Synchronized REPRODUCTION OF SOUND AND SCENE

REPRINTED FROM
BELL LABORATORIES RECORD
NOVEMBER, 1928

Reprinted from
Bell Laboratories Record
A Monthly Magazine of Information for Members of
BELL TELEPHONE LABORATORIES
November 1928

Reproducing Sound and Scene

Theatrical audiences throughout the United States have been enjoying for many months audible motion-pictures, recorded and reproduced by methods and apparatus developed in Bell Telephone Laboratories and made available to producers and exhibitors by Electrical Research Products, Incorporated. A furor has arisen in the theatrical and motion-picture professions, excited by the wondrous possibilities and amplified by public demand and interest.

Relating to this new art of talking motion-pictures much has already been published, for it is based upon years of telephonic research in speech and hearing, the conversion of energy between acoustic and electrical systems, and electrical methods for recording, amplifying and reproducing sound.

Publication of the technical processes involved, however, is following the actual accomplishment in accordance with the tradition of the Laboratories that demonstration shall precede exposition and the hazard of prophesy be avoided.

During September a series of papers was presented to the Society of Motion Picture Engineers by members of the technical staff of the Laboratories. These dealt with the fundamental principles of synchronized recording and reproducing of sound and scene, and with some of the apparatus developments of the Laboratories in that field.

The substance of these original expositions was then presented by the authors to the readers of BELL LABORATORIES RECORD, in a series of articles printed in the November, 1928, issue of that magazine and reprinted therefrom in the present form.

Contents

Fundamentals of Speech, Hearing and Music

By JOHN C. STEINBERG
Research Department

SOUND-PICTURES unite two long lines of development: one of the recording and reproduction of sound, affecting the sense of hearing; the other of the recording and reproduction of visual objects, affecting the sense of sight. The success attained in these lines has been largely due to research continually carried on in the fundamental facts of each and to the perfection of apparatus modified with each advance in knowledge.

The present paper briefly recounts some of the outstanding facts of the science of sound as it affects the development of the sound-picture. It compasses in reality five sub-sciences: one pertaining to hearing, the sense organ on which all perception of sound depends, two covering speech and music, the types of sound most commonly reproduced, and two covering musical instruments and the voice, the generators of the two main classes of sounds.

Speech is produced by streams of air forced out through the vocal passages by the lungs. The trachea or wind pipe is terminated at the upper end by the larynx which contains two muscular ledges, known as the vocal cords, forming a straight slit through which the air stream passes. During speech these cords vibrate so that the slit is alternately opened and closed and by this action a train of sound waves is set up in the lower part of the throat. As these waves pass out into the air, certain resonant and transient characteristics are impressed upon them by the vocal cavities and the movements of the tongue and lips, and it is these variations that are interpreted as speech.

All speech sounds are produced in this manner, except those symbolized by the letters p, k, t, f, s, ch, sh, and th (as in thin). These, called unvoiced sounds because the vocal cords play no part in their production, are produced by frictional vibration set up in the mouth itself. Both voiced and unvoiced sounds may be divided into two classes: those produced by a continuous flow of air, called the con-

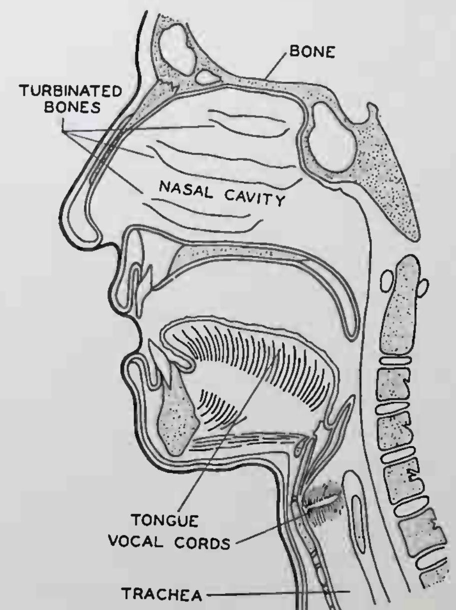


Fig. 1—A cross-section of the human head showing relative positions of the vocal organs.

tinuants, and those produced by the sudden stoppage of the air, called the stops. In the former class are such sounds as a, v, and f, and in the latter class such sounds as p, g, d, and t.

Although the vocal cords lend quality to the voice they do not give, to any great degree, the distinguishing characteristics of the speech

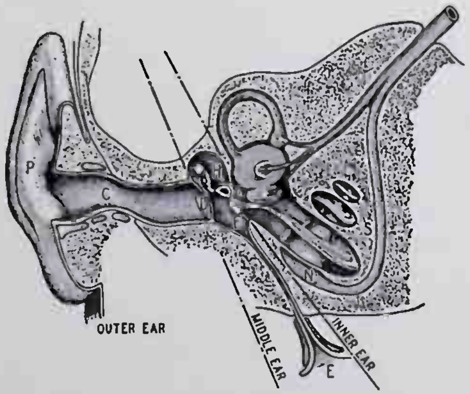


Fig. 2—A cross-section of the ear canal can only suggest its complex structure.

sounds. These are produced mainly by the mouth and nose cavities, as is evidenced by the fact that we can understand whispered speech, in which the vocal cords play no part. As a matter of fact counterparts of the lungs and vocal cords can be located quite outside the body and still produce intelligible speech. This is actually accomplished by the artificial larynx, a piece of apparatus developed by Bell Telephone Laboratories for those who have undergone an operation known as tracheotomy, in which the larynx is removed and the wind pipe is terminated by a small hole in the patient's neck, through which he breathes.

The mechanism of hearing may be divided into three parts: the outer, the middle, and the inner ear. The outer ear, consisting of the external parts and the ear canal, terminates at the drum. The middle ear contains three small bones, the hammer, anvil, and stirrup, which connect the drum with the small window or diaphragm, "O," of the inner ear. The inner ear is a spiral cavity, "S," in the bone, which is filled with liquid, and into it, as may be seen from the illustration, projects a spiral ledge. The liquid above this ledge is separated from that below it by a flexible membrane along which are distributed the nerves of hearing. Two windows, "O" and "E", retain the liquid at the base but at the apex there is a small hole through the membrane which allows liquid to flow from the upper side to the lower.

Sound enters the ear as successions of minute changes in the air pressure which are known as sound waves. These cause the ear drum to vibrate, and it is supposed that the liquid in the inner ear vibrates similarly, affecting the central membrane in different positions depending upon the frequency—the high tones disturbing one end of the membrane and the low tones disturbing the other end. This membrane may be compared to the keyboard of a player piano in operation, the keys at different parts of the scale being disturbed successively with the progress of the music. With the ear the pattern of the disturbance on the membrane is carried to the brain and there interpreted as speech sounds.

The range of pressure and frequency that the ear can perceive is represented on Figure 3 where the scale of abscissas is frequencies in cycles per second, and the ordinate

scale is pressure in dynes. Frequencies above about 20,000 cycles are not perceived as sound nor are those below about 20. Any frequency between these limits, however, is recognized as sound if its pressure is above the lower boundary curve marked "Threshold of Audibility." The upper boundary, marked "Threshold of Feeling" indicates the pressure at which feeling begins. Above this line the sounds are felt, actually causing pain by their excessive pressure.

Frequency and pressure are only the physical characteristics of a sound; our mental responses are called pitch and loudness. Both of these vary logarithmically with their stimulus, difference in pitch between two sounds corresponding to the logarithm of the ratio of their frequencies, and similarly, differences in loudness are proportional to the logarithm of the difference in pressure, but with loudness the proportionality is not quite constant so that constant loudness lines are not truly horizontal. Because of this logarithmic law the illustration is plotted with logarithmic scales, and in addition an arbitrary loudness scale is shown on the right, the units of which, called sensation units, are defined as twenty times the logarithm of the pressure.

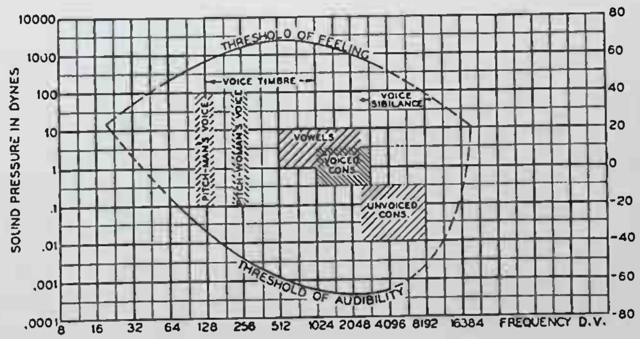


Fig. 3 — Any sound that can be heard lies within the field outlined here. Areas covered by the most prominent speech sounds are indicated.

Studies on the wave forms of speech sounds have shown that the pitch of man's voice is of an order of 128 cycles per second, that of woman's voice of an order of 256 cycles, and with both, overtones of the fundamental cord tone occur. These studies have shown further that frequencies as high as 8000 or 9000 cycles exist in various speech sounds. Studies on the interpretation of speech sounds have indicated the presence of tones covering a large part of the audible frequency range. The location of the various parts of speech on the entire sound area is indicated in the illustration. Although in this figure the speech sounds have been grouped in sharply defined areas, actually the sounds overlap somewhat, and the indicated areas are those which are most important in the interpretation of the sounds. The three voiced consonants, symbolized by the letters v, z, and th (as in them), are exceptions and belong in the unvoiced consonant area.

In general, woman's speech is more difficult to interpret than man's, which may be due in part to the fact that woman's speech has only one-half as many tones as man's, so that the membrane of hearing is not disturbed in as many places. The greatest differences occur in the case of the more difficult consonant sounds which in woman's speech are not only fainter but require a higher frequency range for interpretation. The range from 3000 to 6000 cycles for man's voice corresponds roughly to the range from 5000 to 8000 cycles for woman's voice and since the ear is less sensitive at these higher frequencies and the sounds are initially fainter, their difficulty of interpretation is

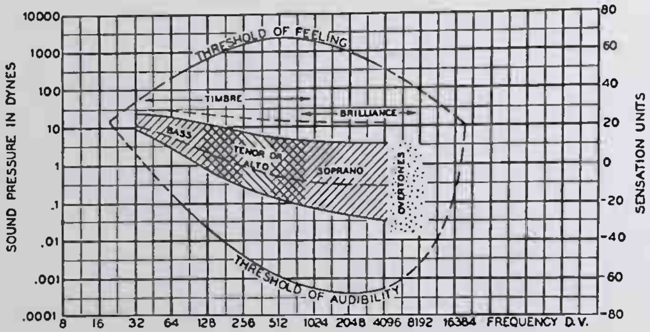


Fig. 4 — Constant loudness lines are not quite horizontal but slope down with increasing frequency as shown.

of course proportionately greater.

Musical sounds, like those of speech, consist of a fundamental frequency and various overtones, but unlike speech are sustained for appreciable lengths of time, and changes in them usually take place in definite steps, known as musical intervals. The frequencies that are present depend upon the type of instrument and the character of the music but the pitch of the tone is determined by the fundamental frequency which, however, need not be present in the musical tone, as the overtones, which are multiples of the fundamental, may cause the correct pitch sensation.

Musical instruments may be divided into two general classes, string instruments and wind instruments. Tones of string instruments, produced by plucking, striking, or bowing, are usually reenforced by resonating air cavities and sounding boards. Tones of wind instruments may be produced by the aid of reeds as with the clarinet or with the flute, or by the lips of the player acting as reeds, as with the horns. Each class may be further subdivided into melody and harmony instruments. In the former class only one note is usually produced at a time; in the latter class several notes are usually produced simultaneously. In general, harmony instruments are capable of producing notes over a much wider frequency range than melody instruments and a given type of instrument of the latter class may, therefore, include several instruments each covering different frequency ranges, such as the bass, tenor, and alto trombone.

Experiments have indicated that notes of different frequency or pitch as produced by a musical instrument appear about equally loud to the ear, which might be expected as the ear has played an important part in their design. In Figure 4 contour lines of equal loudness are shown for the frequency range from 32 to 4000 cycles, which has been divided into three parts, the bass, tenor or alto, and soprano registers, corresponding to the notes produced by various instruments. The contour lines indicate that the notes of the lower registers have greater sound pressures than those of the higher. The range of pressures for various instruments, however, is smaller for low notes, as has been determined by direct measurements of the pressures produced

when played by musicians. Contour lines for loud tones show a smaller change in pressure in going from low to high notes than do the contour lines for faint tones so that it would seem that music played faintly would cover a greater pressure range than loud music.

Percussion instruments such as drums and the various accessory traps produce the greatest pressures that are used in music and although the fundamental frequency of the notes which they emit is fairly low, the complete notes are particularly rich in tones of higher frequency, extending as high as 10,000 cycles. Although these higher tones die out rather rapidly, they are essential to good definition.

The organ, the piano, and the harp have the greatest span, covering a frequency range from about 16 to 4000 cycles. All three of these instruments are characterized by a rather prominent first overtone, so that their effective range extends as high as 8000 cycles.

Melody instruments, owing to their limited range, are among the easiest to reproduce. In any given register, wind instruments produce greater intensities than string instruments, of which the violin produces the faintest sounds. As a class these instruments produce notes covering the frequency range of 32 to 4000 cycles.

From the auditory sensation area, we have seen that the ear is able to perceive a large number of tones of different intensity and frequency. We have also seen that the voice and various musical instruments produce tones which cover a large portion of the auditory sensation area. In order to obtain information as to the relative importance of various parts of this area to the sensory characteristics of speech and music, experiments have been performed in which the tones falling in various parts have been eliminated from the sounds by means of filters.

When frequencies below 100, 200, 300, or on up to 1000 cycles are progressively eliminated from speech, its character changes markedly, the terms "timbre" or "tone color" best describing the characteristic lost. This characteristic appears to be associated with the fundamental and the first few overtones of the voiced sounds and their presence is necessary, therefore, in order to convey this quality, but for the correct interpretation of the speech sounds frequencies below 300 cycles do not appear to be essential.

When frequencies above 8000, 7000, or on down to 3000 cycles are eliminated, the character of the speech again changes markedly; the term "sibilance", appearing to describe best the characteristic lost, refers to the prominence of the hissing or frictional character of speech. If attention is directed to such sounds as s, f, th, and z, the elimination of frequencies above 6000 or 7000 cycles is readily detectable, but it requires rather close attention to detect the elimination of frequencies above 8000 cycles. Elimination of frequencies above 7000 cycles, however, slightly impairs the interpretation of the s and z sounds of woman's voice and elimination of frequencies above 6000 cycles those of the f and th sounds of man's voice, and of the f, th, s, and z sounds of woman's voice. The impairment due to eliminating higher frequencies is usually greater in the case of female voices.

As with speech, the tone color or

timbre of musical tones also appears to be associated with the fundamental and the first few overtones of the note produced. Timbre is probably more important in music than in speech as it is one of the things that distinguish the tones of various instruments. In general, the fundamental and the first three or four overtones are necessary in order to distinguish the tones of various instruments. When overtones higher than these are eliminated the tones lose a characteristic best described by the terms "brilliance" or "definition"; they seem to lose life and become dull. The prominence of these characteristics varies with the type of instrument, the composition of the music, and the personality of the musician.

The notes which are used most in music are contained in the octaves below and above middle C, or from 128 to 512 cycles, and as the fourth overtone of 512 cycles has a vibration frequency of 8192 cycles, tones of this frequency and below occur frequently in music. A trained ear could no doubt detect the elimination of frequencies above this range from the ordinary run of music, but the average individual would have difficulty in detecting the elimination of frequencies above even 6000 or 7000 cycles, unless he gave particularly close attention to the percussion instruments.

Another phenomenon of hearing which enters into the sensation of sound is called masking. Lower pitched tones in a sound deafen the auditor to the higher tones, and this deafening or masking effect becomes very marked when the sound pressures of the lower tones are greater than twenty sensation units. The optimum loudness for the interpretation of speech corresponds to a sound pressure between 0 and 20 sensation units. If the sound pressure is less than this the fainter sounds are inaudible while if it is greater, the masking effect impairs the interpretation. When sounds are increased in loudness the lower registers are accentuated because of this masking effect so that for most faithful reproduction sounds should be reproduced with about the same loudness as the original sounds.

General Principles of Sound Recording

By EDWARD C. WENTE
Research Department

THAT sound is perceived by the ear as the result of a disturbance in the air was known to the ancient Greeks, and that objects are set in vibration by intense sounds must have been observed by primitive man, but it was not until 1857, or less than a century ago, that the first instrument was constructed for making a graphical record of sound waves. In that year Léon Scott patented in France an instrument (Figure 1) which he called the phonautograph. A piece of smoked paper was attached to the cylindrical surface of a drum, so mounted that when rotated by hand it moved forward at the same time. A stylus was attached to the center of a diaphragm through a system of levers in such a manner that it moved laterally along the surface of the cylinder when the diaphragm vibrated. Over the diaphragm was placed a barrel-shaped mouthpiece. When the drum was rotated, words spoken into the mouthpiece caused the stylus to trace a wavy line upon the smoked paper. This wavy line was the first known record of sound vibrations.

It was twenty years later, in 1877, that Edison brought out an epoch-making invention. He constructed a machine very similar to the phonautograph but differing in two important details. The smoked paper was replaced by a sheet of tinfoil, and the stylus was attached directly to the diaphragm so that it traced an impression of variable depth, as the diaphragm vibrated, instead of a wavy line as with the phonautograph. After such a record had been made the drum was returned to the starting point and, with the stylus in place, again rotated as before. The recorded sound was then intelligibly reproduced. Thus Edison gave us the first phonograph.

In subsequent models the tinfoil was replaced by a wax cylinder. For many years the wax record, either in cylinder or disc form, was used almost exclusively for the recording and reproducing of sound. Although many other methods of recording have been suggested, it is only in the last few years that records made photographically have come into the

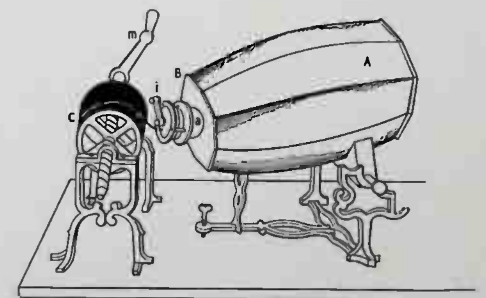


Fig. 1 — Sound was first recorded on the phonautograph of Léon Scott.

commercial field as competitors. At the present time both the wax and the photographic records are used in conjunction with motion pictures.

Photographic records are now being made by many different types of apparatus, but they may be divided into two general classes. In one of these, the record is a trace of constant photographic density but of variable width, while in the other the width is constant but the density varies. In one or two proposed methods the record is a combination of both types.

Almost all systems experimented with today have at least one element

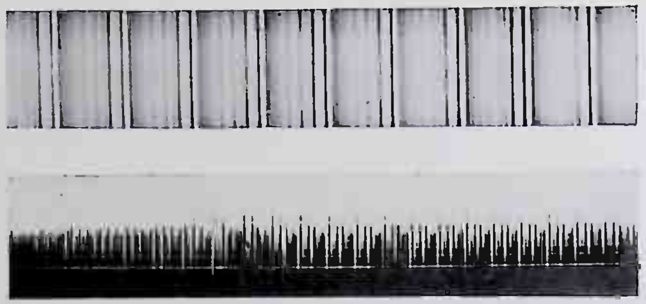


Fig. 2—Photographic records of sound may be either of constant width and varying density, shown above, or of constant density and varying width, show below.

in common with the phonautograph: a diaphragm that is set in vibration by the sound to be recorded. The diaphragm may be mechanically connected to the engraving mechanism or recorder as in the phonautograph, or, it may be electrically connected as in most modern systems. In practically all systems the diaphragm forms an essential element.

Unfortunately a diaphragm does not in general have the same response at all frequencies. A favorite experiment in lectures on elementary physics is to sound a tuning fork and with it, through the air, set in vibration a second tuning fork. In this experiment it is important that the pitch, or the resonant frequency, of the two forks be very nearly the same, or the motion set up in the second fork will be too small to be observable. Diaphragms, and in fact almost any other type of mechanical system, will have at least one resonant frequency, which means that, under the action of sound waves, the response will be much greater at this frequency than at any other.

In the older methods of recording resonance was purposely introduced in order to obtain records of sufficient amplitude. The frequencies lying in the resonance region were then much over-emphasized. The sound reproduced from such records had a blasting and metallic quality, and well deserved the title "canned music."

Because of the complex nature of speech and music and of the great amount of distortion introduced by the early recorders and reproducers, the surprising thing is not that the quality of reproduction was poor, but that the reproduced sounds were at all intelligible. In fact it has been suggested that the invention of the telephone, which preceded the phonograph, might have been delayed for many years had the complex nature of speech sounds been generally known at the time, since its inventor probably would have dismissed his ideas as altogether impracticable.

Although considerable distortion may be introduced by the recording and reproducing systems before the character of the sounds is so changed that they can no longer be recognized, the amount of distortion must be kept extremely small, if all classes of sounds are to be reproduced to such a degree of fidelity that the ear cannot distinguished them from the original. It is necessary, therefore, to diminish the distortion by the diaphragm to a negligible value.

The electrical method of recording, which is today widely used in the production of commercial sound records, has been developed primarily so that a diaphragm giving a uniform response may be used and at the same time a record of sufficient amplitude be obtained. In the modern method the pick-up diaphragm is made a component part of the recording microphone. Here a small amplitude of motion will serve, as the voltage generated may be amplified to an amount sufficient for operating a rugged and distortionless recorder. A comparison of the diaphragm in the Edison recorder with that of the microphone used in the majority of present recording systems is interesting. In the former the maximum amplitude required for the loudest sounds is about 0.001 inch, whereas in the latter under ordinary recording conditions it is only about one-tenth of this amount, and the weight of the microphone diaphragm is only one-twentieth as great as that of the Edison recorder. It can thus be seen how the problem of design of a pick-up diaphragm is greatly simplified in the electrical method.

It is important, of course, that the rest of the recording system shall also be free from distortion. If a microphone of uniform response is available, however, the design of a distortionless recorder is made comparatively easy, for its sensitivity may to a large extent be disregarded, as the required power can be obtained by vacuum-tube amplifiers. In the electrical method, extraordinary improvements have been made over the older systems in elimination of distortion.

The problem of developing recording apparatus is in many respects identical with that of developing high quality radio transmitters. With recording apparatus, however, there is the additional problem of distortion introduced by the record itself. If, for instance, a record is run at a speed of ten inches per second, and a tone having a frequency of 5,000 cycles per second is recorded, the length of one cycle on the record will cover a distance of only 0.002 inch. In the case of wax records the needle must have a very fine point; and in the photographic record the width of the light beam as measured along the direction of motion of the film must be extremely small. At whatever speed the record may be driven, there will always be some frequency beyond which all tones will become more and more attenuated. Although the loss of the higher frequencies does not

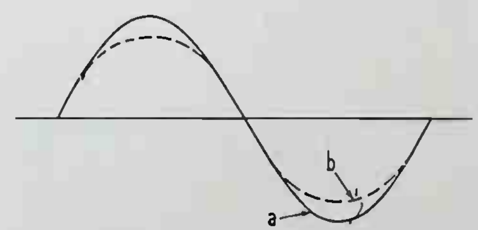


Fig. 3—Non-linear distortion changes the shape of a sine wave form from "a" to the flatter top type of "b".

impair the tone quality as much as does the presence of sharp resonance regions, yet it reduces the intelligibility of speech and the richness and brilliancy of musical sounds.

There is another type of distortion commonly present in reproduced sound, which is frequently designated non-linear distortion. It is introduced when the response of any element of the system is not proportional to the stimulus. A pure tone, for example, of sine wave form, as shown in "a" of Figure 3, may be reproduced so as to have a wave form like that of "b". Distortion of the wave form in this manner is equivalent to the introduction of extraneous frequencies. If the magnitude of these added frequencies is too great, the tone quality will be very disagreeable. A small amount of distortion of this kind, however, is not noticeable, because the primary tone will mask the extraneous one. It is a well known fact that a tone must be much more intense to be heard if another, and particularly a lower, tone is sounded simultaneously.

A type of distortion peculiar to recording is that introduced by a non-uniform speed of the medium on which the record is being engraved. This may not always be serious, but in certain cases of sustained tones, speed variations cause a disagreeable flutter and in some types of music a decided harshness of tone.

One of the most serious problems with which the radio engineer has to contend is static interference. This also has its counterpart in sound reproduction from records. As the ether through which the radio waves are sent is non-homogeneous because of extraneous electrical disturbances, so the sound record is non-homogeneous on account of the non-uniformity of the material on which it is engraved. The noise resulting from these irregularities is often designated as surface noise. With the wax record, most of this noise has its origin in the minute irregularies of the material and with the photographic record, in the finite size of the grains forming the photographic image.

The difficulties of eliminating this noise arise from the fact that the physical intensity of audible sounds covers an exceedingly wide range. The ratio of pressures of the maximum to the minimum is about ten million. If a record of this extreme range of volume were to be recorded the amplitude of the loudest tone would have to be ten million times as great as for the faintest tone. There is a maximum amplitude that a record can accommodate, which in the case of the wax record is about 0.002 inch. If a tone having an intensity near the maximum level is recorded at this amplitude, the amplitude of a tone just audible would be only 0.000,000,000,2 inch. It is difficult to get a material having a degree of homogeneity corresponding to this value.

A similar condition exists with photographic records where the pattern is formed by grains in the emulsion which have a magnitude somewhat less than 0.000,05 inch, depending upon the type of emulsion used. The range of volume considered here is extreme, of course, and in practice it is not necessary to record a range of this extent, but it serves to illustrate the extraordinary requirements placed upon the recording medium. When the range of frequencies that are to be reproduced is increased, the surface noise effect becomes greater. As in the case of the different types of distortion discussed above, the difficulties to be met are increased as the quality of reproduction is improved.

Recent Advances in Wax Recording

By HALSEY A. FREDERICK
Research Department

SUCCESS of recording and reproduction of sound by the so-called "electric" method with a disc record may be considered as depending on several factors. In order, these are: the studio, with its acoustic conditions; the microphone; the amplifier; the electro-mechanical recorder; the "wax" record; the copying apparatus and procedure; the hard record, or "pressing"; the electric pick-up; the amplifier; the loud speaker; the auditorium. The chief problem is that of making the speech or music reproduced in the auditorium a faithful duplicate of the original sounds, using this chain of apparatus. Cost, reliability, and the time required for the process of recording are among a number of other considerations, but these are all subordinate to the problem of fidelity. While it may be convenient or even necessary to introduce distortion into certain of the steps to compensate for such distortion as may be unavoidable in others, experience shows that it is desirable for the sake of simplicity, reliability and flexibility to reduce such corrective warping to a minimum, and to make each step in the process as nearly perfect as possible.

Perfection of a complete recording and reproducing system may be judged by the practical method of listening to the overall result. Each element of the system must be analyzed thoroughly, however, if outstanding excellence is to be attained. One of the most useful of the means of analysis is study of the response-frequency curve. In order that all frequencies be reproduced equally and that the ordinary faults of resonance be avoided, this curve must be flat and, particularly, free from sharp peaks. Good reproduction requires that frequencies from 50 to 5,000 cycles be included without discrimination. If however frequencies down to 25 or 30 cycles be included, a noticeable improvement will be obtained with some classes of music, while if the upper limit be increased to 8,000 or even 10,000 cycles, the naturalness and smoothness of practically all classes of reproduction will be noticeably improved.

A second important criterion of any system is that the ratio of output to input shall not vary over the range of currents or loudnesses from the minimum up to the maximum used. If this requirement be not met, sounds or frequencies not present in the original will appear in the reproduction. This is the type of distortion commonly produced by an overloaded vacuum-tube amplifier; it is often called non-linear distortion.

A third requirement not entirely dissociated from the first two is that any shifts in the phase relations shall be proportional to frequency.

Since our standards of perfection in sound-reproducing systems are growing constantly more exacting, over-all results that seemed excellent a short time ago are only fair today, and before long may seem intolerable. It has, therefore, been necessary for the analysis of each step of the system to be constantly more searching and fundamental.

Of the eleven links in the chain of apparatus for electrical recording and reproduction, only five are pe-

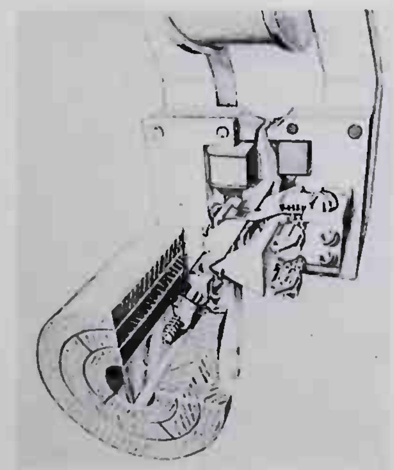


Fig. 1—Section of electrical recorder.

culiar to the wax method. These are the electro-mechanical recorder, the wax record, the copying apparatus, the "pressing," and the pick-up or reproducer. The extent to which the wax method is capable of the highest quality of reproduction will be disclosed by an examination of these five links. Any consideration of the practical advantages or disadvantages of the method can logically follow this examination into the quality possibilities. The considerations which follow refer to the so-called lateral-cut records, in which the grooves are of constant depth and oscillate or undulate in each case about a smooth spiral. This type of record is used in the Western Electric Company method of synchronized motion pictures which uses disc records. Some, but not all, of the considerations might apply to "hill and dale" records, but the characteristics of these records will not be discussed.

The first piece of apparatus in the chain unique to the wax process is the electro-mechanical recorder, whose function is to receive power from the amplifier, and with it drive a mechanical recording stylus. The present-day recorder is a highly developed apparatus based on extensive experimental as well as theoretical studies. Recorders which have been supplied by the Western Electric Company have been designed to operate over a range of frequencies from 30 to 5,500 cycles. The device operates in linear fashion over the range of amplitudes involved in speech and music. As is seen in Figure 2, the response falls off below about 250 cycles. This falling characteristic is necessary in order that the maximum loudness be obtained from a record for a given spacing between grooves without cutting over from groove to groove. A characteristic of the pickup is that the voltage induced in its windings is proportional to the velocity with which the armature moves. In order therefore that a lateral oscillation of the needle point may furnish constant output voltage, it is necessary that the lateral velocity of the needle point be constant. For a sine wave, velocity is proportional to the product of amplitude and frequency, so that as frequency increases, amplitude must decrease proportionately. With the characteristic shown with these recorders, constant velocity is obtained from about 250 cycles to 5,500 cycles. Below 250 cycles an approximately constant amplitude is obtained. If,

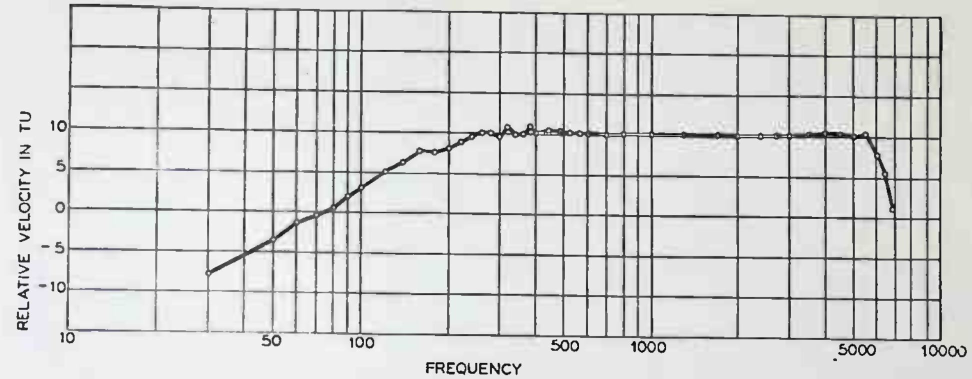


Fig. 2—Typical frequency characteristic of a commercial recorder.

therefore, sounds of constant absolute intensity are to be recorded over this range of 30 to 250 cycles, there is equal tendency for sounds of the different frequencies in this range to over-cut the record groove. Attenuation of the lower frequencies by the recording apparatus may be corrected in reproduction by a suitable electric network. Such a network will increase the subsequent amplification required but, as this additional amplification occurs in the first stages, it is not expensive. Practically it has not been found necessary or desirable to introduce such a corrective network since the correction has been largely cared for by the characteristics of the pickups which are used.

Recent development studies have established the possibility of flattening the response at the low-frequency end and of raising the high frequency cut-off of the recorder. The characteristic obtained with a laboratory model shows uniform performance within ± 1 TU from 250 to 7,500 cycles and within ± 4 TU from 30 to 8,000 cycles. Although its immediate practical value may be limited by other portions of the system, this device is of great interest in that it establishes beyond question the fact that an extremely broad range of frequencies can be successfully recorded in the wax.

The broad, flat characteristic obtained with electric recorders has

Fig. 3—Frequency characteristic of a recent laboratory recorder.

been made possible by so designing their elements that they constitute correctly designed transmission systems. In such a transmission system, whether it be an electrical recorder or a long telephone line, a correct terminating impedance is required. The load imposed by the wax is somewhat variable but fortunately is rather small. It has been found desirable to make the other impedances in the recorder relatively large so as to dominate the system and thus minimize the effects of any changes in the impedance imposed on the stylus by the wax. The mechanical load used as a terminating impedance and to control the device has consisted of a rod of gum rubber ten inches long.

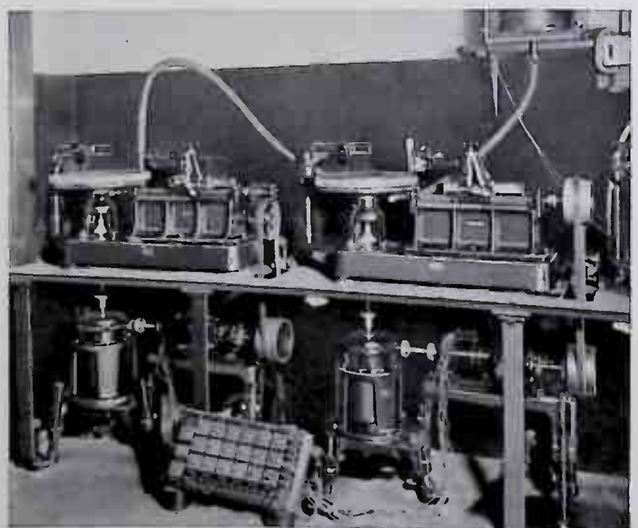


Fig. 4—Two recording machines arranged to be driven synchronously with cameras.

Torsional vibrations are transmitted along this rod at about one hundred feet per second so that its length is equivalent to an ideal electrical line of about 1,500 miles. Loss of energy along this rubber rod is such that a vibration is substantially dissipated by the time it has travelled down the line and back. Thus the rod constitutes a substantially pure mechanical resistance, whose magnitude is approximately 2,500 mechanical ab-ohms, referred to the stylus point as its point of application.

In recording, the usual procedure is to use a disc from one inch to two inches thick and from thirteen to seventeen inches in diameter, composed of a metallic soap to which small amounts of various substances have been added to improve the texture. This disc, commonly called a "wax," is shaved to a highly polished surface on a lathe, and then is placed in a recording machine, essentially a high-grade lathe by which the wax is rotated at a very uniform rate and in definite relation to the film with which it is being synchronized. The recorder with its cutting tool is moved radially across the surface of the disc, common phonograph procedure being to record from the outer edge of the disc toward the center, whereas with records for Western Electric sound pictures the direction of cutting is reversed. After a record has been cut the wax may be handled, and with proper precautions may be shipped from place to place.

The shape of the groove varies somewhat in commercial practice. That used in records for Western Electric apparatus is approximately

0.006 inch wide and 0.0025 inch deep, and the pitch of the spiral is between 0.010 and 0.011 inch, so that the space between the edges of the grooves is about 0.004 inch. Thus the maximum safe amplitude is about 0.002 inch. If this is reached at 250 cycles the corresponding amplitude at 5,000 cycles, assuming that the sound is constant in absolute intensity over the intervening range, will be only 0.0001 inch.

In the records used with Western Electric apparatus the linear speed of the groove past the reproducer point ranges from 140 feet per minute, at the outside of the spiral, to 70 feet per minute at the inside. The rate of rotation is dependent upon the outer diameter of the grooves which is determined primarily by the length of time to be covered by a single disc. When the minimum linear speed and the groove spacing are decided upon, there is an optimum relation between the size of the record, the rate of rotation and the playing time.

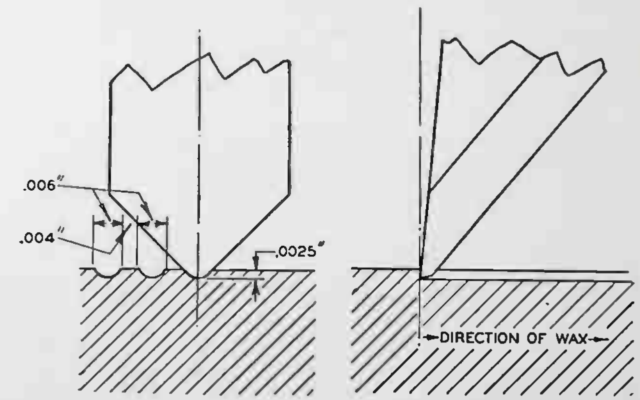


Fig. 5—Details of the stylus point of the recorder. Left, a radial section through the wax; right, a section looking from the stylus point to the center of the disc.

Since any roughness in the walls of the groove introduces extraneous noise in the reproduced sound, it is important that the groove be kept truly smooth. Before starting, the surface of the disc must be shaved to a high polish, and the texture must be fine and homogeneous. Not only must the composition be correct, but the proper temperature must be maintained during recording. Waxes may be obtained commercially whose texture is satisfactory over the ordinary range of room temperatures. The disc must be levelled in the recording machine with reasonable care, and the stylus must be sharp and of a shape to insure a clean cut. The wax shaving is removed as cut by air suction. To aid in maintaining the cut at the correct depth there is a so-called advance ball which rides lightly on the surface, supporting the stylus at the proper height in spite of small inaccuracies in levelling the disc or deviations from planeness. For adjusting the advance ball with respect to the stylus, the groove is observed with a calibrated microscope. Maintenance of the necessary adjustments and satisfactory operation of the recording machine requires an ordinarily skilled mechanic with reasonable experience.

After a record has been cut, the sound may be reproduced directly from the wax by means of a suitable pick-up or reproducer. Ordinary reproducers rest much too heavily on the records to be used on wax; the vertical pressure between needle point and record in an ordinary phonograph is of the order of 50,000 pounds per square inch. Obviously any such pressures would destroy a groove in the wax.

These high pressures have been necessary in order that the groove might drive the needle point of the reproducer properly; their reduction requires reduction of the impedance offered by the needle point to transverse vibration. Such reduction of

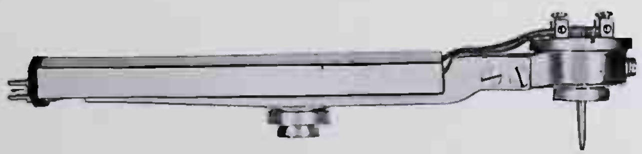


Fig. 6—*Playback for reproduction directly from wax discs.*

the impedance of a "playback" mechanism requires that both its mass and its stiffness be cut down to a minimum. That now available has been designed with those requirements in view, and represents a large advance toward ideal reproduction. Whereas playbacks formerly in use failed to reproduce the higher and lower frequencies with much satisfaction, response of the new piece of apparatus is not widely different from that obtained from finished records with the best electric pick-ups now available. The response is sufficiently good to serve as a most valuable criterion in judging the quality of a record immediately after recording. At the same time, a record may be played a number of times without great injury. At low frequencies there is little change and at the higher frequencies a loss of about 2TU per playing. The opportunity thus made available for an artist to hear and criticize the results of his efforts, immediately at the end of his performance, can hardly be overestimated in its value to studio and recording operation.

After a groove has been cut into the wax record, the usual procedure is to render the surface conducting by brushing into it an extremely fine conducting powder. It is then electroplated. The technique in this step varies somewhat among the various companies producing records, although fundamentally the process is the same with all. The electro-plate thus made, a negative of the original wax, is called a "master." From it two test pressings are usually made, using a molding compound such as shellac containing a finely ground filler. If they are satisfactory the master is then electro-plated after being treated to permit easy removal of the positive plate. From the positive, commonly called an "original," there is plated a second negative—a metal mold, commonly called a "stamper." From it, duplicate "originals" may be plated, and from them, duplicate "stampers." These successive plating processes involve no measurable injury to the quality of the record, and are comparatively simple and extremely safe in practice. By the custom of making a number of duplicates the master is protected from accidents to which it would be subject were it to be used directly for making finished records. The stampers instead are used to mold the final product, or "pressing"; it is not unusual to make a thousand pressings from a single stamper. Test pressings are commonly obtained from the wax in twelve hours, and recent refinements have so reduced the time for the various processes that finished pressings may now, when necessary, be obtained within three hours after

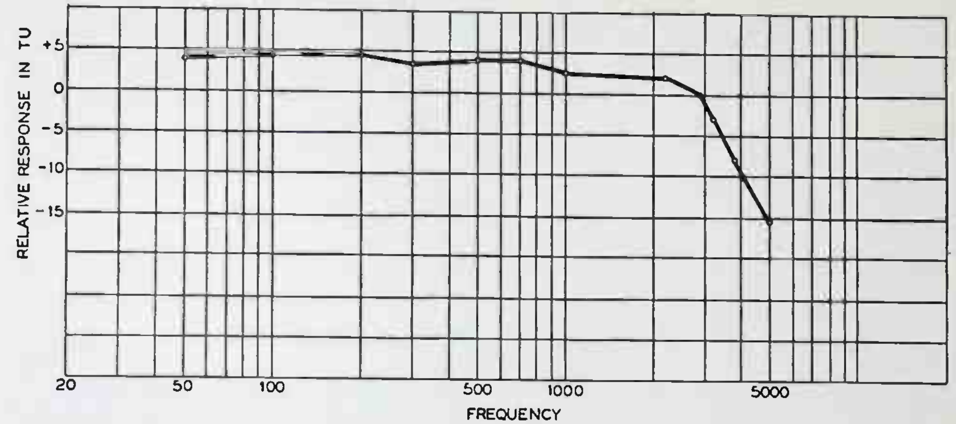


Fig. 7—*Response of the wax playback of Fig.* 6 *driven by a constant-velocity wax record.*

delivery of the wax. The pressing copies the wax record with such a high degree of accuracy that if frequency characteristics alone be considered, it shows almost complete perfection. Moreover it is cheap and durable, and as with an ordinary phonograph allows reproduction of sounds without careful adjustments or complex apparatus.

Various materials have been used in making the pressings. In some cases the material has been homogeneous, and in others the surface and the body of the record have been of different materials. Some records have been of laminated structure. There has not, however, been much latitude allowed the experimenter in his selection of materials. The records must be quite hard and, to have a reasonable life, must contain enough

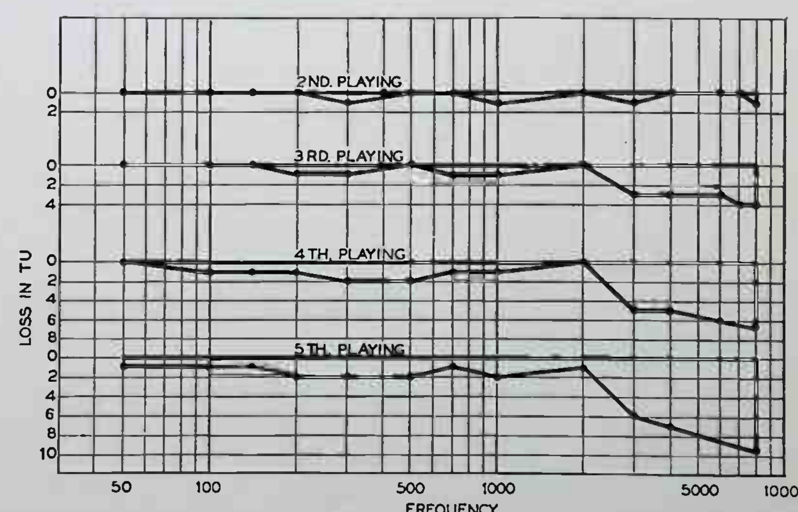


Fig. 8—*Loss in response of a wax playback driven by a constant-velocity wax record, on successive playings.*

abrasive to grind the needle quickly to a good fit. At the beginning of the run of a new needle the pressures are very high, on account of the small bearing surface. They decrease rapidly however, so that with an ordinary loud steel needle in a phonograph of the usual type the bearing surface is increased to such an extent after one minute's wear that the pressure is only about 50,000 pounds per square inch. The high pressures and the necessary abrasive characteristics have introduced irregularities which are responsible for most of the extraneous noise commonly known as "surface" or "needle scratch."

Recent development in the material of the finished records, together with refinement in the plating processes, have reduced the surface noise of records used in Western Electric Company theatre equipment by fifty to seventy-five per cent within the last two years. It is not necessary to reduce the level of surface noise to the zero-point, but merely to the threshold of audibility under the minimum auditorium noise. Moreover, the important point is not the absolute amplitude of the imperfections giving rise to surface noise but their relative magnitude with respect to that of the useful sound amplitudes. Thus an effective reduction in "surface" could be made by using larger records or reducing the playing time of the present records, in either case increasing the spacing of the grooves and the amplitude with which they are cut. Conversely, any large reduction in surface noise made by an improvement in the record material would make it possible to increase the playing time for records of a given size. There is no known absolute or fundamental reason why further improvements in record materials may not be expected, with corresponding reduction in surface noise. Furthermore large advances in pick-up design offer distinctly new possibilities for reduction in "surface."

It has sometimes been thought that in order to reproduce high frequencies properly, the linear speed of the record would have to be increased or the size of the needle point reduced. The factor determining whether a needle will follow the undulation of the groove is not its diameter relative to the length of the undulation, but the relation between the radii of curvature of the needle and the bend of the groove. At present the bearing portion of a representative needle

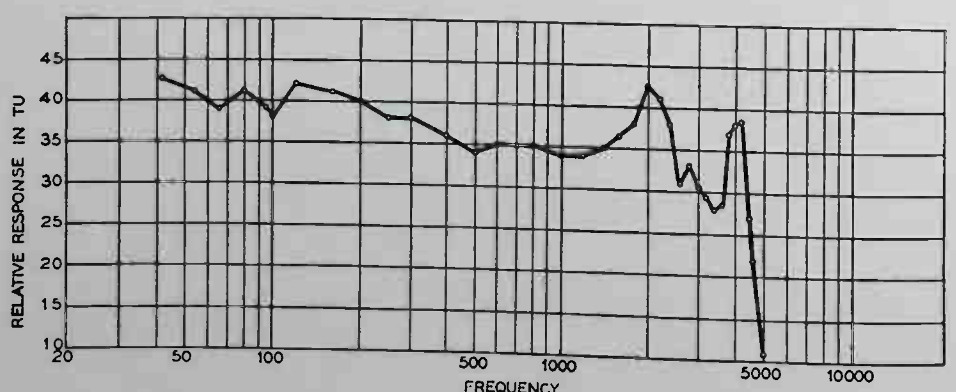


Fig. 9—Response of a 2-A pick-up driven by a constant-velocity pressing.

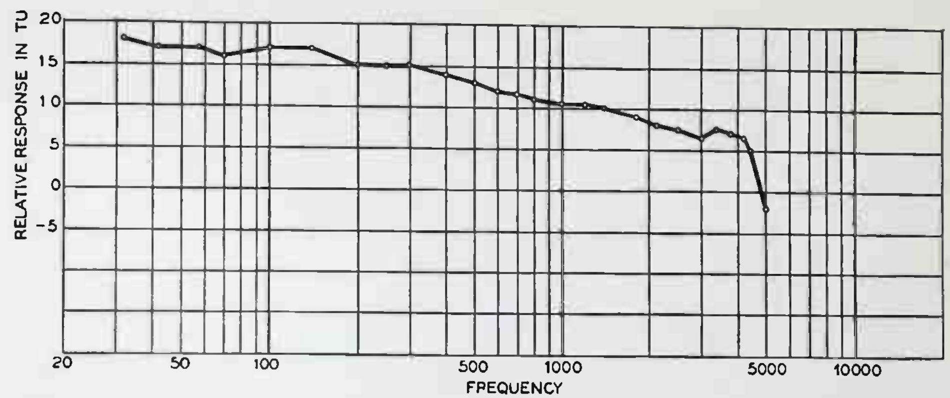


Fig. 10—Response of a 4-A pick-up driven by a constant-velocity pressing.

is about 0.003 inch in diameter whereas the half wave-length for a 5,000 cycle wave is 0.0014 inch. As mentioned before, the amplitude at 5,000 cycles would be only about 0.0001 inch if sounds of that frequency were as intense as those of lower frequencies. With the assumption of an amplitude of 0.0001 inch at 5,000 cycles and a linear record speed of seventy feet per minute, the minimum radius of curvature of the undulation is 0.00193 inch. On a corresponding basis, the radius of curvature of the undulation becomes equal to that of the needle point at about 7,000 cycles. As a matter of fact, sounds of 5,000 cycles or more in speech and music are characterized by lower intensity than those of lower frequency, and the amplitude of their undulations is correspondingly less. Present commercial needle points are therefore quite capable of following the undulations up to frequencies of at least 10,000 cycles. The limitations to high-frequency reproduction commonly found in the past have come from imperfections in the design of the pick-up or reproducer. They were caused mainly by inability of the record groove to drive the needle point, with resultant chatter, and by failure of the pick-up structure to transmit high-frequency motions from the needle point to the armature.

Large advances have been made within the last two or three years in designing electric reproducing structures. The mechanical impedance at the needle point has been reduced so that the point follows the undulations of the groove truthfully without the necessity of somewhat destructive bearing pressures. At the same time the mechanical transmitting structure has been so designed that a very broad range of frequencies is conveyed properly from the needle point to the armature. Moreover, proper mechanical loads have been provided so that the vibrations are absorbed after they operate the armature; resonance as ordinarily considered has been eliminated.

The curves shown in Figures 9 to 11 show the improvement which has been made. The pick-up shown in Figure 10 is free from the resonances shown in Figure 9. In the earlier pick-up there were high needle-point impedances in the region of these resonances, involving large driving forces destructive to both needle and record. Certain records were injured by only a few playings with this reproducer.

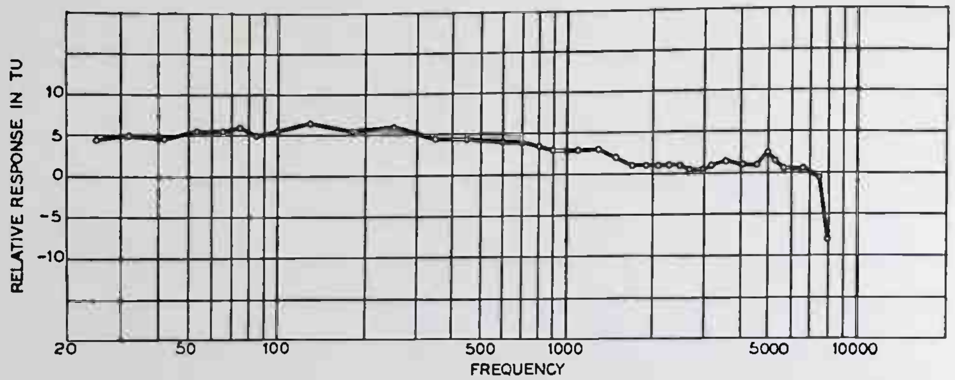


Fig. 11—Response of experimental pick-up driven by a constant-velocity pressing.

The reproducer of Figure 10 is characterized not only by considerably reduced average needle-point impedance but, as shown on the curve, by a practical elimination of resonance. Hence there is an even greater reduction from the maximum impedances which occurred at resonance in reproducers of the earlier type. Both needles and records have a longer life with the later type pick-up, which has been in commercial use for several months. In addition, as is seen, the higher frequencies are reproduced in considerably better fashion. A third curve, Figure 11, was obtained with a more recent experimental model in which a lighter, though very much more rigid, structure is used to connect the needle point with the armature. In this model a further large reduction in needle-point impedance has been effected, with corresponding reduction in wear on the records; in addition greatly improved reproduction has been secured at the high-frequency end of the scale.

The developments and inquiries heretofore discussed, although in many cases undertaken for their application to sound pictures, are of direct application to phonograph reproduction for any purpose, irrespective of whether the sound is to be accompanied by a film. Application of the processes to records which are to be synchronized with motion pictures has in addition involved meeting a number of conditions not previously encountered in the phonograph field. One of the most important of these relates to editing, cutting and rearranging of a picture, with the attendant necessity of rearranging the speech or music. Various methods have been used to copy or "dub" a disc record by playing it and recording on a new wax the parts that are wanted. The prime requirement is that the sacrifice in quality be kept at a minimum. To attain this end records have sometimes been copied at very low speed. This method appears unnecessarily laborious and slow and the results obtained are not altogether satisfactory in the light of possibilities presented by pick-ups and recorders of the characteristics shown above. Rearrangment of material on records is entirely practicable, and portions may be deleted, new portions added, or new sounds added to those already recorded; in fact any changes of this type may be made which can be made in the picture. Although the detailed technique of rearranging, or "dubbing," is not highly developed at present, its advancement appears to offer no serious technical difficulty. Hence the refinement reached will probably depend on the rate at which improvement takes place, and the extent to which the method is used in the field of sound pictures.

Sound Recording with the Light Valve

By DONALD MacKENZIE
Apparatus Development Department

OF the several ways by which sound can be recorded on motion-picture film, one has seemed to engineers of Bell Telephone Laboratories to offer most immediate promise. This employs a light beam of constant intensity and varying width to produce a trace of varying density. Modulation of the light beam is effected by an electromechanical light valve actuated by speech currents which have been amplified to a suitable volume.

The light valve consists of a loop of duralumin tape suspended in a plane at right angles to a magnetic field. When the assembly of magnet and armature is complete, the two sides of the loop constitute a slit 0.002 by 0.256 inches, its sides lying in a plane at right angles to the lines of force and approximately centered in the air gap. The ends of the loop are connected to the output terminals of the recording amplifier. If the magnet is energized and the amplifier supplies an alternating current, the loop opens and closes in accordance with the current alternations.

When one side of the wave opens the valve to 0.004 inches and the other side closes it completely, full modulation of the aperture is accomplished. The natural frequency of the valve is set by adjusting the tension of the tape; for reasons which involve many considerations, the valve is tuned to seven thousand cycles per second. Under these circumstances about ten milliwatts are required for full modulation at a frequency remote from resonance; about one one-hundredth of this power at the resonant frequency. The impedance of the valve with protecting fuse is about twelve ohms.

If this appliance is interposed between a light source and a photographic film we have a camera shutter of unconventional design. Figure 2 shows a diagram of the optical system for studio recording. At the left is a light source, a ribbon-filament projection lamp, which is focussed on the plane of the valve. The light passed by the valve is then focussed with a two-to-one reduction on the photographic film at the right. A simple achromat is used to form the image of the filament at the valve plane, but a more complicated lens, designed to exacting specifications by Bausch and Lomb, is required for focussing the valve on the film. The undisturbed valve opening appears on the film as a line 0.001 by 0.128

inch, its length at right angles to the direction of film travel. The width of this line varies with the sound currents supplied to the valve, so that the film receives exposure to light of fixed intensity during the varying time required for a given point to traverse the varying aperture of the slit.

Recording in the studio is carried out on a film separate from that which receives the picture. This practice permits the use of two machines to make duplicate sound records, an insurance which is well worth its cost. The practice of separate negatives for sound and picture also permits the picture negative to be developed and printed according to well-established technique, and allows the necessary latitude in developing the sound record. The recording machine is driven by a motor synchronously with the camera. Lest there be any variation in the velocity of the film past the line of exposure, the sprocket which carries the film at that point is driven through a mechanical filter which holds the instantaneous velocity constant to one part in one thousand.

In the recording machine a photoelectric cell is mounted inside the left-hand sprocket which carries the film past the line of exposure. Fresh film transmits some four per cent of the light falling on it, and modulation of this light during the record is appreciated by the cell inside the sprocket. This cell is connected to a preliminary amplifier mounted below the exposure chamber, and with suitable further amplification the operator may hear from the loud speaker the record as it is actually being made on the film. Full modulation of the valve implies complete closing of the slit by one side of the wave of current; this modulation may not be exceeded or photographic overload will abound.

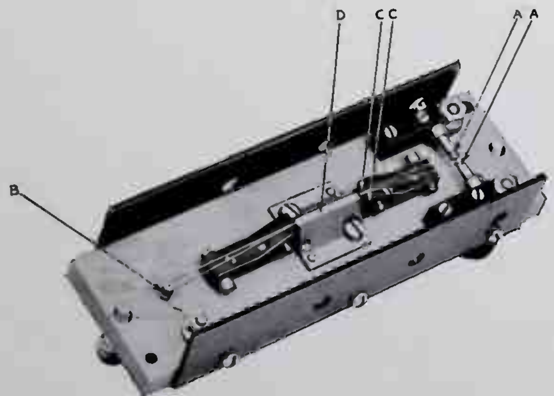


Fig. 1—*The light valve. A duralumin tape,* 0.006 *inch wide and* 0.003 *inch thick, is secured to windlasses A and A′ and stretched tight by the spring-held pulley B. At points C and C′ insulated pincers confine the central portions of the tape between windlasses and pulley to form a slit* 0.002 *inch wide. Supporting this loop and adjusting devices is a slab of metal with central elevation D, which constitutes the armature of an electromagnet. The central portions of the loop are supported on insulating bridges just above the face of D; here the sides of the loop are centered over a tapered slot. Viewed against the light, the valve appears as a slit* 2 *mils by* 256 *mils.*

Adding sound to the picture introduces no complication of studio technique other than to require sufficient rehearsing to make sure of satisfactory pick-up of the sound; microphone placement must be established and amplifiers adjusted to feed to the light valve currents which just drive to the edge of overload in the fortissimo passages of music or the loudest utterances of speakers. Provision is made for combining if desired the

contributions of several microphones on the set. This combination is under the control of the mixer operator in the monitoring room, who views the set through a double window in the studio wall. The mixer controls also the gain of the amplifiers for the recording machines. Relays permit the mixer to connect the horn circuit either directly to the recording amplifier or to one or the other of the monitoring photoelectric cells in the film recorders. The direct connection is used in preparing the sound pick-up in the studio: the program is rehearsed until satisfactory arrangement of microphones and of amplifier gain is effected. The electrical characteristic of this direct monitoring circuit is so designed that the sound quality heard in the horns shall be the same as the quality to be expected in the reproduction of the positive print in the theater. Acoustic treatment of the walls of the monitoring room secures the reverberation characteristic of the theater, and the monitoring level is so adjusted that the mixer operator hears the same loudness that he would wish to hear from the theater horns. It is capitally important that the operator judge his pick-up on the basis of sound closely identical in loudness and quality with that to be heard later in theater reproduction.

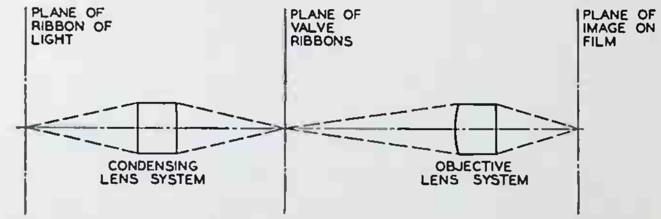


Fig. 2—*Optical system for studio recording.*

After the pick-up has been established on the direct monitoring circuit, the output of the recording amplifier is applied to the light valves and the monitoring horns are connected to the photoelectric cell amplifiers on the recording machines. With no film in the machine and at a convenient lamp current a complete rehearsal is made to verify the operation of the valves at the proper level. Film is then loaded, cameras and sound recorders are interlocked and starting marks made on all films by punches or light flashes.

A light signal from the recording room warns the studio, which after lighting up signals back its readiness to start. The machine operator starts the cameras and sound recorders, brings up the lamp current to the proper value, and when the machines are up to speed signals the studio to start. During the recording, the mixer operator monitors the record through the light valves, thereby assuring himself that no record is lost.

In printing these sound negatives in combination with pictures for projection in the theater, it is customary at the present time to print one negative, masking the space needed for the other, then to run the positive again through the printer with the other negative, masking the space already printed. In printing the sound negative, the light is regulated to result in thirty-five per cent transmission of the unmodulated track after positive development. Provision of suitable masks in the camera has been made to show in the finder and expose on the film only the portion which will be available for picture projection. In the theater projector, the sound gate is located fourteen and one-half inches below the picture

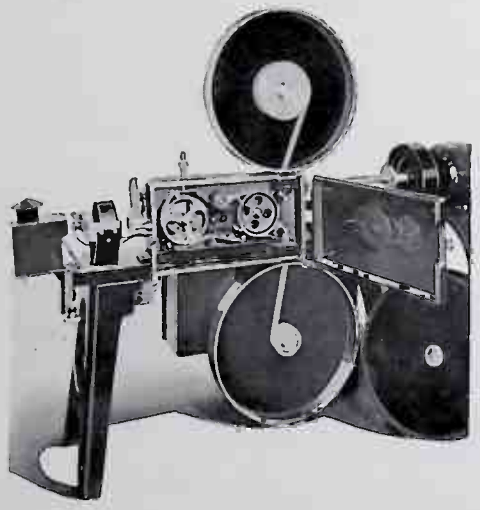


Fig. 3—A studio recording machine with the door of the exposure chamber open. The left-hand sprocket engages twenty perforations and is driven through a mechanical filter.

gate, in order to project the sound record at a point where the film is in continuous motion.

As with other systems for sound-transmission, that which includes film recording and reproduction has certain inherent faults which may be minimized by careful design. These are background noise, irregular response at various frequencies, and distortion due to non-linear characteristics.

Background noise results principally from casual variations in the light-transmission of both positive and negative films. Raw stock that is entirely satisfactory for pictures may be too irregular for sound records, since the photoelectric cell recognizes variations of 0.1 per cent while the eye ignores contrasts under two per cent. The remedy is to use "positive" stock for the sound negative as well as for the print and to use developer as little granular as possible in its effect. Fortunately, it is necessary to reduce the background noise only to a point below the threshold of audibility which will exist in the theater during the softer parts of the program. This point determines the level of the faintest sound-record which can be reproduced unmarred by noise.

Due to the facts that the element of illumination is 0.001 inch wide, instead of infinitely narrow, and that the film is travelling at ninety feet per minute, at a frequency of 18,000 cycles per second it will require the time of one cycle of a sound wave for a given point on the film to cross the slit. Then as each successive element of film crosses the slit, it will receive an exposure proportional to the integral of a complete cycle of the valve. Since the integral of one cycle is the same regardless of the phase at which the integral starts, each successive element will receive the same exposure, and the film will develop to a uniform density. Consequently no record will be made of the sound. Fortunately this frequency is far outside the range of interest to us, and the effect decreases as frequencies become lower. The drooping characteristic resulting, called the film transfer loss, may be largely offset by judicious choice of electrical characteristics and by tuning the light-valve mechanically to a frequency near 7000 cycles. How successful these measures are, in the range of importance in program reproduction, is evident from the curves of Figure 4.

When the curve connecting power input to the film system with its power output is not a straight line, distortion results, as in purely electrical systems. This takes the form of an introduction of harmonics of the frequencies normally present. The curve which connects exposure of photographic emulsions with resulting opacity is a straight line only when development is so controlled as to produce a contrast-ratio* of unity. Picture-recording practice is to develop the positive print to a contrast-ratio greater than unity. Development of the sound-negative is therefore so controlled as to give a contrast-ratio the reciprocal of that to be expected in the positive, so that the overall ratio is unity. Distortion from this cause is then so completely annulled that the resulting harmonics are undetectible.

The correction just outlined is available over only that part of the photographic range where exposure is correct. For exposures outside this range, the characteristic curve of film becomes curved in a way which cannot be compensated in the printing, and a liberal crop of harmonics is inevitable. However, the range of positive film is about twenty to one, and with the combination of light-source and optical system developed in Bell Telephone Laboratories, it is not difficult to set the unmodulated light at a value which will give an exposure of ten times that corresponding to the beginning of under-exposure. Then ninety per cent modulation of the light can be permitted without running into exposure on the "faint" side of the wave. For sound currents reaching one hundred per cent modulation of the light, ninety per cent of the wave is free from distortion; if the average light were halved, still eighty per cent would be free from distortion. There is therefore considerable latitude in the average exposure, and the negative is satisfactory if the transmission of the unmodulated track lies between fairly wide limits.

The volume of reproduced sound for a given reproducing light source varies directly with the average track density and the per cent modulation of this average. In printing the sound

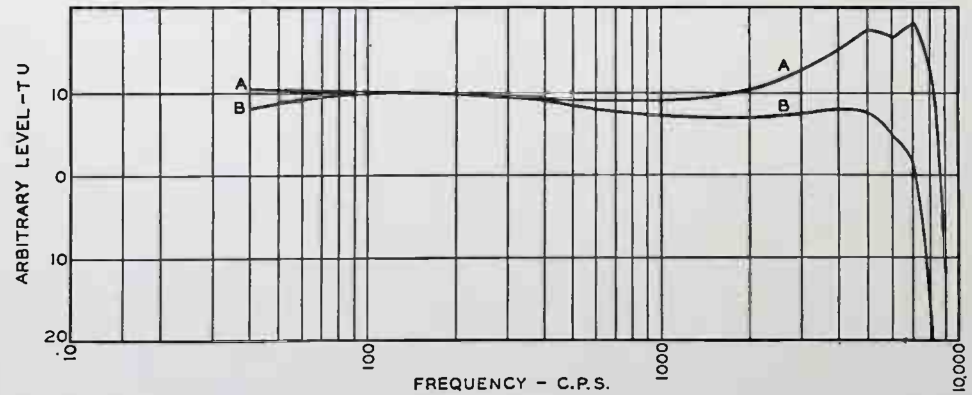


Fig. 4—Recording and reproduction characteristics: from acoustic pressure at microphone diaphragm to (A) light modulation by valve and to (B) current delivered to loud speaker in theater.

* *Technically known as "gamma," it is the slope of the Hurter and Driffield characteristic curve.*

negative, a uniform density for the print of the unmodulated track is desired, lest there be changes in the sound output during the showing. For Eastman positive film a suitable transmission of the unmodulated portion of the sound print is thirty-five per cent. At this average transmission only the peaks of the recorded sound will encroach on the region of under-exposure. For the reciprocally-developed negative track the region of under-exposure will have been reached by occasional peaks on the other side of the wave, and whatever photographic distortion exists will be balanced between positive and negative.

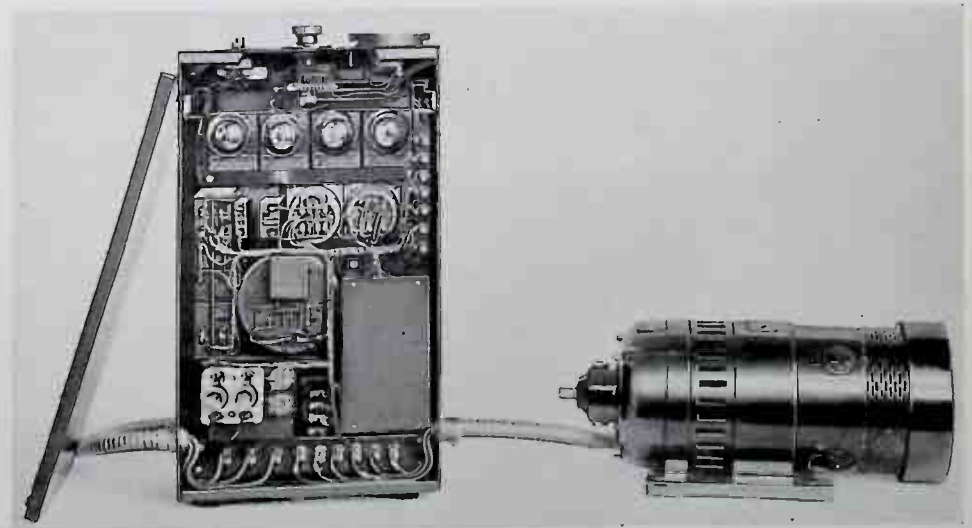
If the entire negative exposure has been confined to the under-exposure region of the emulsion chosen, a huskiness in the reproduction will result which can not be corrected by any known technique. But with correct exposure, which is readily possible with the light-valve method, ninety per cent of the wave will be clear of under-exposure, and experience shows that the ear detects no distortion. In telephonic terms, everything at a level one TU below full modulation will be free from distortion, and the peaks will be substantially perfect.

Speed Control for the Sound-Picture System

By HUGH M. STOLLER
Apparatus Development Department

SYNCHRONIZATION between sound record and film is required for all types of sound-pictures. In the Western Electric System it is secured by electrical methods for the recording apparatus and, for the reproducer, by mechanically coupling the picture projection machine and the sound record. With satisfactory synchronization thus secured, there still remains, however, the problem of speed control. Musical pitch varies directly with frequency or rate of vibration. The faster the record is rotated, the higher the pitch of the sound given out. In order, therefore, that the reproduced music may be of the same pitch as that recorded, the record must run at an assigned speed, and to keep the pitch from varying during the playing of the record, this speed must be prevented from changing. To attain these ends, the speed of the driving motor must be accurately controlled.

In determining how nearly constant this should be held, the criterion is the smallest pitch change that is noticeable, and it has been found that abrupt variations are more readily perceived than slow ones. A good musical ear will detect sudden changes in pitch produced by a change in speed of only one-half of one per cent. To make sure, therefore, that a discernible change in pitch never arises, speed regulation better than one-half of one per cent is required at all times. As further allowances seemed desirable to provide a suitable factor



An alternating-current drive motor for the sound picture connected to the governor.

of safety, a regulation of two-tenths of one per cent was agreed upon.

A survey showed that no commercial governing mechanism was available which would meet the requirements. The most suitable was probably the governor used with ordinary phonographs. This governor applies friction as the speed increases,

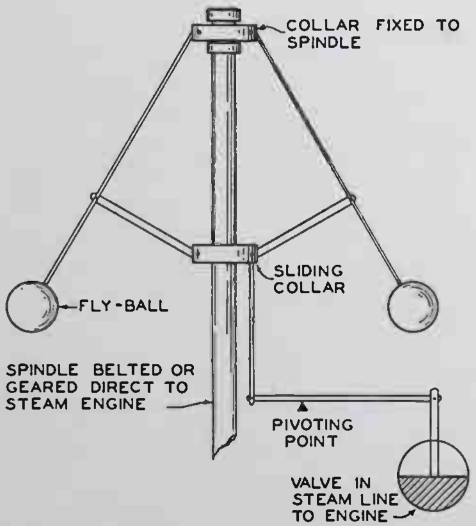


Fig. 1—Diagramatic representation of a fly-ball governor showing how sensitivity may be changed by moving the pivoting point to the left or right.

and it becomes increasingly difficult, because of maintenance difficulties, to design a satisfactory governor of this type for the larger motor required to drive both the projecting machine and turntable. A completely new design seemed the only satisfactory course.

The nature of the problem and the difficulties to be overcome are perhaps most readily brought out by considering a simple fly-ball governor controlling the speed of a steam engine, shown in Figure 1. Rotation tends to make the balls move outward due to centrifugal force. As they move out, however, they pull up a sliding collar and this action, through a system of levers, closes the steam valve supplying the engine, or in some other manner decreases the supply of steam, and the engine slows down. The tendency of the balls to fly outward is opposed by their weight so that there is a definite equilibrium position for each speed.

As a result of this the engine will run at a given speed for only one load. A load greater than this will require a wider valve opening so as to admit more steam, and this in turn will require a lower position of the fly-balls to allow it. This necessary drop in speed to allow a wider valve opening may be reduced, however, by moving the position of the pivoting point of the lever to the left. This changes the lever ratio so that smaller and smaller movements of the sliding collar will produce greater and greater valve openings. Inherently, however, some speed change must be permitted in order that the valve may be moved to accommodate the new load. To enable the engine to run at one speed regardless of the load requires an additional mechanism that will admit enough more steam to the engine to carry the increased load at the desired speed.

When, however, the sensitivity of the governor is made too great by moving the lever pivot too far to the left, or when the additional mechanism is added to make the engine always run at the same speed, an unstable condition is brought about. Under these conditions, the engine, at each action of the governor, tends to over-shoot its mark, either not attaining equilibrium speed at all or reaching it only after several oscillations. This instability may be overcome by adding a dashpot or some similar contrivance to the governor which, allowing an initial rapid adjustment, will later introduce a slower compensating force that prevents oscillations of speed.

The same governing principles apply whatever may be the type of governing apparatus. A change in speed must be able to cause some force to act on the prime mover tending to counteract this change and, if the governing is to be made very close, some dashpot equivalent must be provided to prevent the tendency to hunt or oscillate back and forth before the equilibrium speed is reached.

The prime mover for the sound-picture equipment is an electric motor. Either direct or alternating current may be used but as the principles involved are the same with each, only the latter need be discussed. The speed of the motor used is controlled by changing the impedance in its armature circuit. With high line voltage or light load the impedance is made a maximum, while with low line voltage or heavy load the impedance is made a minimum. A suitable governor, therefore, must be arranged to change this armature impedance at a very small change in motor speed (below 0.2 of one per cent) and must incorporate some arrangement to prevent oscillations.

In the sound-picture system the somewhat cumbersome fly-balls with their sliding collar and connecting levers to the steam valve are replaced by a few comparatively simple electrical circuits which act silently and instantly to correct the slightest change in speed of the driving motor. The electrical governing apparatus may be split up into three parts corresponding to those of the fly-ball governor. One part will substitute for the driving link (not shown in Figure 1) between the engine and the governor spindle, another for the fly-balls themselves and their connecting levers, and a third for the steam valve that changes the torque of the engine.

The driving link for the electrical system is a 720 cycle generator coupled to the main driving motor of the projector unit.

The governing circuit proper is a special bridge circuit shown as Figure 2. One arm of the bridge has a fixed inductance and condenser in series, which are adjusted to tune the circuit to 720 cycles. At this frequency the impedance of this arm is a resistance only and the impedance of the arm D is made a resistance of the same value. At 720 cycles the ratio of these two arms, therefore, is unity, as is that of the other two arms, made up of the primary of a transformer divided at its half tap. The small

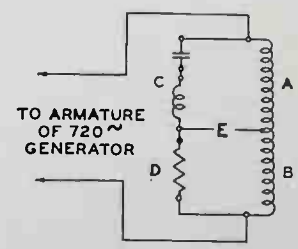


Fig. 2—The heart of the sound-picture governor. Potential E shifts 180° in phase as the speed changes from any value below 1200 r. p. m. to any value above it.

alternator, connected across the transformer, thus becomes the source of power for the bridge.

This arrangement makes an extremely sensitive analogy to the fly-ball governor and lever system. The potential, E, from the mid-point of the coils A and B to the junction of arms C and D, which is zero at 720 cycles, shifts its phase 180° as the speed changes from any frequency below 720 cycles to any above it. Below 720 cycles the current in C is leading due to the predominance of the condenser and above 720 it is lagging due to the predominance of the inductance. Instead, therefore, of a gradual change as the speed changes from its desired value there is an abrupt one which furnishes a basis for accurate speed control.

Acting as the steam valve to control the speed of the motor is an impedance coil with three legs shown as Figure 3. The two outer legs, L_1 and L_2, are connected to the motor armature and serve as the impedance controlling the speed. The middle leg, G, carries a direct current winding. As the current in this winding increases the magnetic flux in the two outer branches increases to saturation and their impedance decreases. The torque of the motor varies inversely with the reactance of the coils, L_1 and L_2, and as a result the greater the direct current flowing into the coil G, the higher will be the torque of the motor.

The link between the bridge and the three-legged inductance is a vacuum-tube circuit which causes more direct current to flow as the motor speed tends to fall. This complete circuit, shown in Figure 4, includes the detector tube V_4, and two tubes V_1 and V_2 which supply current for the middle winding of the impedance coil. Tube V_3 is a rectifier to supply excitation for the 720 cycle alternator and grid biasing voltage for V_1, V_2, and V_4.

Plate potential for V_4 comes from the alternator through the transformer T_4, and its phase is fixed. Grid potential for this tube is from the bridge output circuit, E, the phase of which shifts 180° as the alternator speed changes from below to above 1200 r.p.m. This 180° shift of phase changes the potential of the grid relative to the plate from negative to positive or vice versa, and thereby causes a relatively large change in the plate current, which, flowing through R_1 causes a correspondingly large change in the grid potential of tubes V_1 and V_2, and thus in the direct current to the impedance coil.

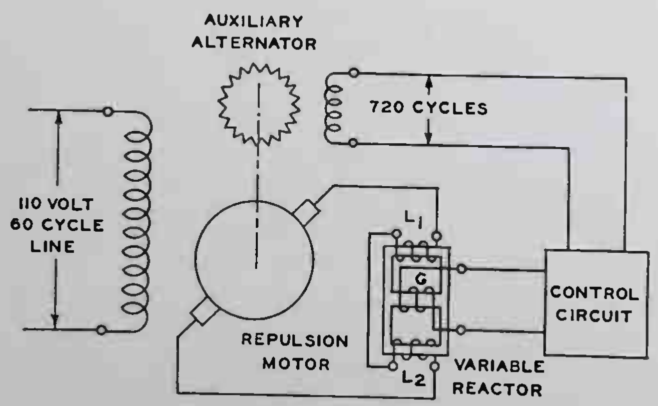


Fig. 3—Torque control for the driving motor is obtained by a three-legged inductance coil, the impedance of which is changed by direct current flowing around the middle leg.

The bridge circuit makes a very sensitive governing device but with it alone some permanent change in speed would be required to compensate for each change in load or line voltage. Something else must be added if the speed is to be constant

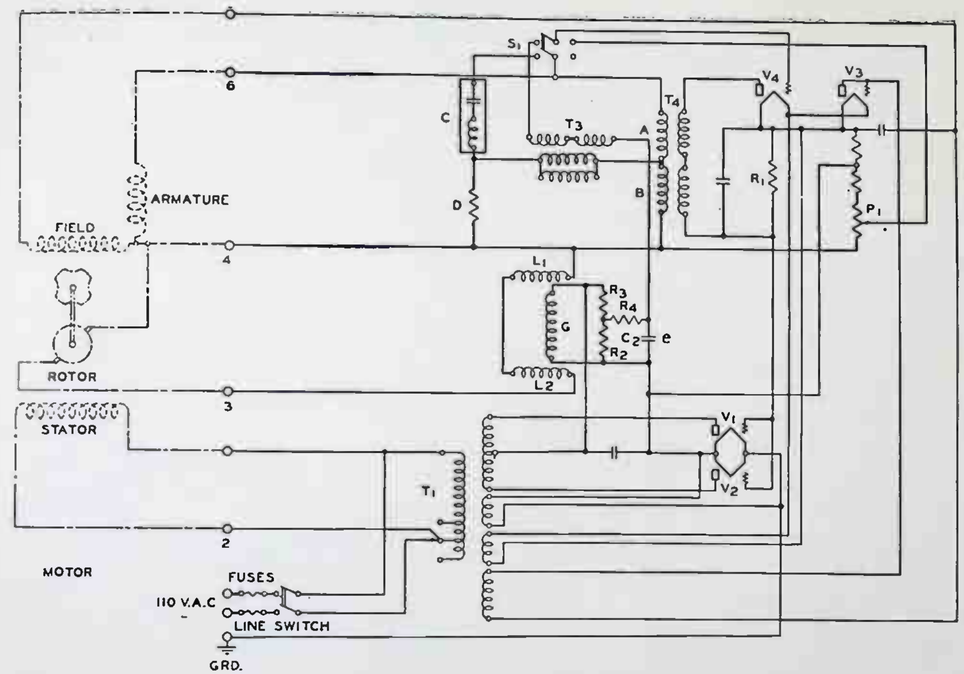


Fig. 4—Complete diagram of the governing circuit showing the three-legged reactance coil in the center. Switch S_1 is normally thrown to the left.

for all conditions. Also, some dashpot equivalent must be used to prevent the speed oscillations mentioned.

These two functions are accomplished by a network composed of R_2, R_3, R_4, and C_2, properly connected to the other circuits. Current flowing in the plate circuit of V_1 and V_2 to the coil of G flows also through R_2 and R_3. The drop across C_2 which after a short preliminary period is the same as that through R_2, feeds back a potential to the grid circuit of V_4, and thereby causes the additional regulation required. To slow down the action of this feed back, however, R_4 in series with C_2 is connected in parallel with R_2 as shown. The feed back potential is that across C_2 and this rises or falls slowly as the condenser must be charged or discharged through the high resistance R_4.

By these means, in the comparatively simple and compact electrical circuit shown in the accompanying cut, is provided a governor, similar in its functioning to the fly-ball governor of Figure 1, that will control the speed of the driving motor under all ordinary conditions to within the required two-tenths per cent. The switch S_1 was added to the circuit so that, when the picture-projection machine was used for ordinary silent motion-pictures, when such close speed regulation is not required, it could be thrown to the right and hand regulation obtained by the potentiometer P_1. This is needed to vary operating speed of the projector to meet a definite time schedule for showing the picture. When the schedule permits, however, accurate speed regulation is preferred since it enables a leader to keep his orchestra in better step with the picture.

Sound Projector Systems for Motion-Picture Theaters

By EDWARD O. SCRIVEN
Apparatus Development Department

IN theaters at which motion pictures accompanied by synchronized speech or music are presented, the records come in two forms. Some are composition discs similar to ordinary phonograph records, while others are standard motion-picture film bearing at one side a track of alternate light and dark bands, of varying density. In either case there must be apparatus synchronized with each projector to derive from the records an electric current in which all the variations in pitch and loudness are accurately represented. There must in addition be apparatus to amplify the current, to effect its conversion into sound and so to direct the sound into the theater auditorium as to create the illusion that it emanates from, rather than merely accompanies, the picture. When a theater is being prepared for presenting sound pictures, the film projectors in use are ordinarily retained but each is fitted with a new motor and driving mechanism; it is provided with a turntable and electric pick-up for disc records, and with analogous equipment for film records, or both. The pick-up used for disc records is in some ways similar to the reproducer of an ordinary acoustic phonograph, with a needle holder connected to a clamped diaphragm of highly tempered spring steel. To the diaphragm there is fastened an armature made of a special high-permeability alloy, so arranged that as the diaphragm and the armature vibrate, the flux in the air-gap of a permanent magnet varies correspondingly; in appropriately placed coils currents are induced which are the electric representation of the wave groove which moves past the needle. Although this instrument delivers energy at a comparatively low level, it has a very uniform response over a wide range of frequencies. That result has been secured largely by preventing distortion which would arise from resonance in any part of the system; the members have been designed with natural periods

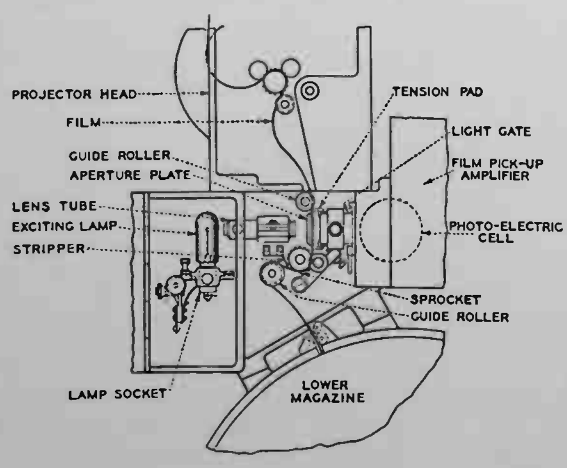


Fig. 1—Arrangement of projector for reproducing from sound films.

beyond the range of frequencies to be transmitted, and the magnet chamber back of the diaphragm is filled with a heavy oil to damp free vibration. The films used with the disc records, called synchronized films, differ from ordinary films only in that one frame at the beginning of each is marked to give the starting point.

Fig. 2—Photoelectric cell of the type used for reproducing from sound film.

With the optical or film records, the sounds are represented by parallel bands, alternately light and dark. Intensity or loudness is represented by differences in density of the record, and pitch by the closeness of the bands. For reproduction from these another apparatus group is required, and it too is connected to the projector. A narrow light beam of high intensity passes through the film record and falls upon a photoelectric cell to produce a current corresponding to that from the original recording transmitter. There is fastened to the projector an "exciting" lamp and a system of lenses for focusing its light upon an aperture 0.0015 inch by 3/16 inch; by other lenses the image of the aperture is then brought to focus upon the film record as a line 0.001 inch by 0.080 inch. Since the track on the film is 0.100 inch wide, there is an allowance of 0.010 inch on each side for variation in its position. The position and focus of the lens tube are fixed, but the exciting lamp is mounted on a movable carriage so that new lamps as installed may be brought properly into focus.

A photoelectric cell of the type used is shown in Figure 2, and the circuit in Figure 3. When polarized by a proper voltage, the cell passes a current proportional, within limiting values, to the intensity of the light falling upon it. The polarizing voltage is supplied to the cell through such a high resistance that in operation there is obtained from the cell a voltage across the resistance proportional to the incident light. The

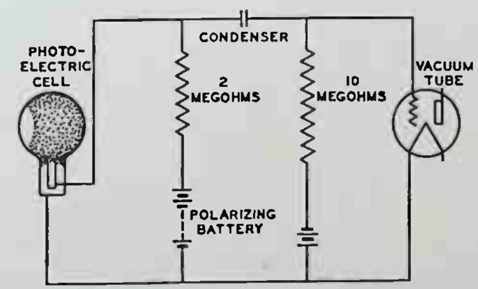


Fig. 3—Circuit from the photoelectric cell to the adjacent amplifier.

voltage bears therefore at any time an inverse relation to the density of that part of the sound track then between the exciting lamp and the cell.

The photoelectric cell circuit is inherently one of high impedance. In such a circuit local interference—"static," to use the radio expression—is readily picked up, and since the energy level is low, the current so acquired may be appreciable in comparison with the sound currents themselves. In addition the shunting effect of the capacity between the conductors

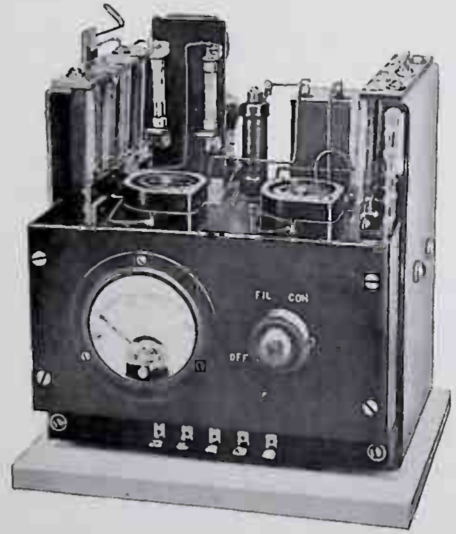


Fig. 4—Amplifier to which the photoelectric cell is connected, showing the suspension for absorbing vibrations. The tubes are not in place.

is noticeable, particularly at the higher frequencies. Hence a vacuum tube amplifier, which serves both to increase the energy and to make that energy available across a low impedance circuit, is closely associated with the cell upon the projector itself. Cell and amplifier are enclosed in a heavy metal box made fast to the frame of the projector, and the frame is carefully grounded. As a further precaution, the amplifier is supported within the enclosing box by a rather elaborate flexible suspension, lest vibration of the vacuum tubes introduce noise components into the current. The amplifier brings up the energy level to about that obtained from the magnet coils of the reproducer for disc records.

It is evident from the relative location of apparatus, shown in Figure 1, that it is not feasible to print the film with the pictures directly opposite corresponding parts of the sound record. Furthermore the pictures move intermittently before the projection lens, while the sound record must of course move uniformly in front of the photoelectric cell. Picture and sound record are therefore separated longitudinally by 14½ inches, and a certain amount of slack is allowed between the sprocket carrying the film in front of the projection lens and that carrying it before the photoelectric cell. To prevent vibration of the projector or variations in the supply voltage or load from varying the speed of the latter sprocket, it is connected to the other moving parts of the system by a mechanical filter which absorbs any abrupt changes in speed. The driving motor is held electrically at the correct speed, but at will the automatic control can be disconnected and the speed regulated manually by the operator.

As with ordinary motion-pictures, two projectors must be used alternately to present a continuous program. At the end of a record, the music or speech coming from one machine must be blended imperceptibly into that from the other just as the picture from one reel is faded into that from the next. At the end of each sound film or disc the music overlaps that at the beginning of the next; to make the transition there is a device called a fader, a double potentiometer. As the starting projector goes into operation, the fader knob is turned and the current delivered to the amplifiers is changed quickly until it comes entirely from the

new record. Ordinarily the fader is installed with auxiliary dials and handles, so that the operator can control it from any position around the projectors. In its lower range, used in changing between projectors, the steps are rather large, whereas in the upper range the volume changes in scarcely perceptible steps. The fader thereby fills another use; it makes possible any volume of sound desired, within reasonable limits, by choice of the proper step in the upper range, and thereby permits equalizing the level of sound obtained from different records. There is provided as well a switch for changing from film to disc records, and the reverse, and a key for connecting a spare projector in place of either of the regular machines.

After passing through the fader, the sound currents go to the main amplifier, where their energy is raised to a level adequate for the loud speakers of the particular theater. This combination of apparatus is capable of multiplying the energy 100,000,000 times, and is so designed that all frequencies in the range from 40 to 10,000 cycles are amplified about equally. A potentiometer is provided on the amplifier but after it has once been adjusted at the time of installation it is ordinarily not changed; necessary adjustments in energy level are made on the fader instead. The amplifier* is built in three units, of which the first consists of three low-power tubes connected in tandem, resistance coupled, with the filaments heated by a twelve-volt battery. In the second unit there are two medium-power tubes with a push-pull connection, whose filaments are heated by low-voltage alternating current. Two similar tubes in this unit act as a full wave rectifier, and supply rectified alternating current for the plate circuits of the amplifier tubes in the first

* *Described by H. A. Dahl in the* RECORD *for May, 1928.*

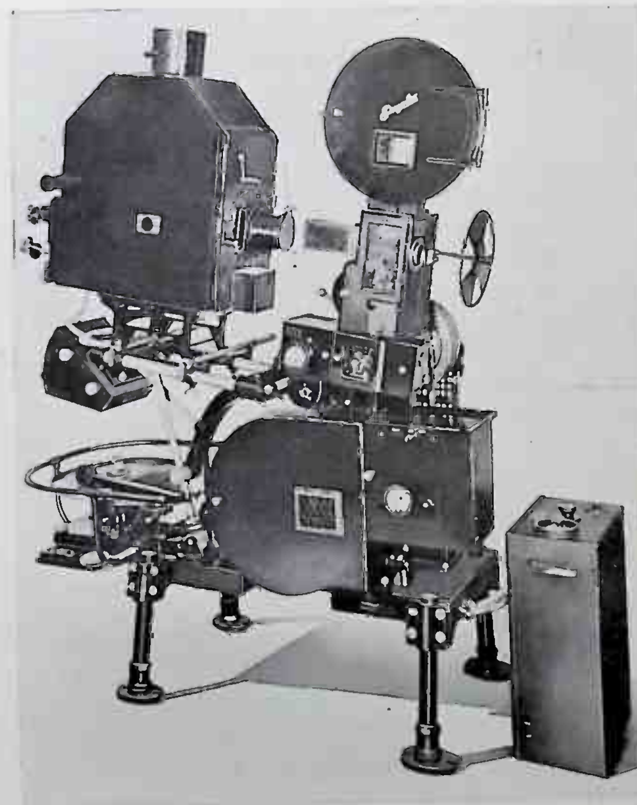


Fig. 5—Western Electric projector for sound pictures, using a Simplex head. A Powers or Motiograph head may also be used.

and second units. The third unit has a single stage of high-power push-pull amplifier tubes and push-pull rectifier tubes; like the second, it operates entirely on alternating current. The three units can be arranged to meet any conditions. In small the-

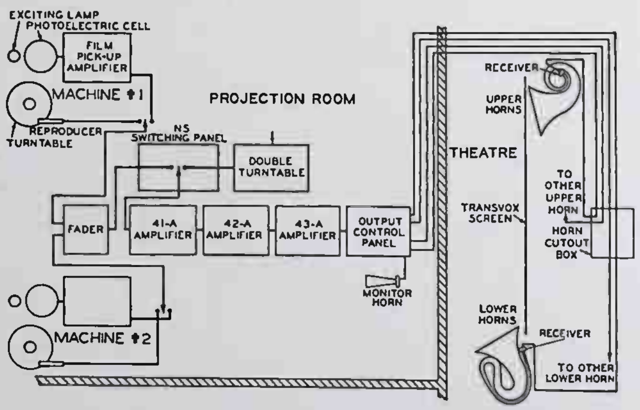


Fig. 6—Layout of a typical theater for presenting sound pictures.

aters only the first two are required, and in larger houses the high-power unit, the third, is used as well. For unusual conditions two or more of the high-power units may be operated in parallel from the output of the second unit to give a greater volume of sound.

Following the amplifier there is an output control panel consisting of an autotransformer having a large number of taps which are multipled to a number of dial switches. To the switches are connected the loud-speaking receivers, so that the impedance of the amplifier output can be matched to the desired number of horns. Thereby there is secured the most efficient use of the power available, and adjustment of the relative volumes of the individual horns is made possible at any time.

A theater installation ordinarily contains four horns. They are mounted behind a transvocent screen, on which the pictures are shown, so that the sound may seem to come directly from the picture. Two horns are mounted at the line of the stage and pointed upward toward the balconies and two are mounted at the upper edge of the screen, or above it, and directed downward.

One or more Western Electric No. 555 receivers* are used with each of the horns. Since these show extremely high efficiency, converting into sound energy thirty per cent of the electrical energy supplied them, they reduce to a minimum the output required from the amplifier. A horn is ordinarily fitted with one receiver, but for outdoor use or other special requirements it may be fitted with two, four or nine by a throat such as that shown in Figure 8. The maximum electrical input to a horn for continuous safe operation is approximately five watts per receiver. To disperse the sound-waves over a large angle, more horns are needed than for a comparatively small angle. This directive characteristic of the horns is important, since it is responsible for the illusion that the sound comes directly from the mouth of the horn, that is, from the screen. When the horn is replaced by a loud speaker of otherwise identical characteristics which radiates its sound over a very wide angle, the sound seems to come from a point some distance behind the screen, so that the illusion of coming from the picture is destroyed.

*Described by A. L. Thuras in the Record for March, 1928.

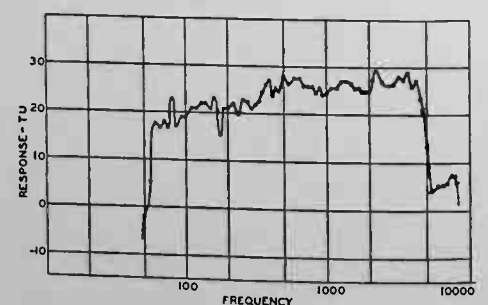


Fig. 7—Response-frequency characteristic of exponential horn and 555-W receiver.

Fig. 8—Throat by which a horn may be fitted with nine receivers, for outdoor use.

In addition to its function as a part of the talking motion-picture equipment the sound projector system may also be used as a public address system for voice reenforcement.

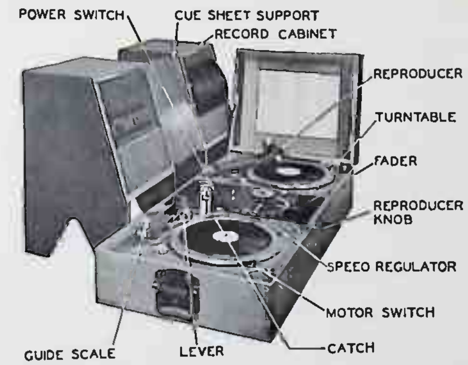


Fig. 9—Cabinet for reproducing from disc records not synchronized with pictures.

Microphones may be concealed in the footlights, and the horns so placed as not to affect them; for announcements ordinarily there will be a transmitter in the manager's office.

By means of auxiliary equipment provided, the system can also be used to provide non-synchronized music as an accompaniment to pictures. There is a cabinet containing two turntables, each with a pick-up and means for locating it accurately upon a record, and a fader to make possible continuous playing. The same amplifier and loud speakers are used as for the synchronized speech and music.

Installation and Adjustment of Western Electric Sound-Projector Systems

By HOWARD B. SANTEE
Electrical Research Products, Incorporated

THAT there may be realized the full advantages which a Western Electric Sound-Projector System can bring to a theater, each installation is planned and directed individually by engineers of Electrical Research Products, Incorporated, the organization which provides the apparatus. Preceding the installation, engineers of that company conduct a detailed preliminary investigation to decide upon the exact apparatus, and to determine any changes that may be needed in the theater building. The equipment is installed under supervision of an engineer of the organization who in addition sees that whatever supplementary work was needed in the building has been carried out properly. As the work proceeds he sees that the equipment is so placed and so regulated as to give the best results, and as the various pieces of apparatus are made ready for use, he trains the projection-room employees in their operation. Finally the completed installation is kept at the highest standard of operation and maintenance through supervisory visits made at intervals by a field engineer, and routine maintenance work is directed by the organization's service group.

Although its need may not at first be fully apparent, the preliminary survey receives the most careful attention to insure that everything will be in readiness for the work of installation to progress smoothly, and that the apparatus best suited for the particular theater will be used. Acoustic characteristics of the stage and auditorium receive thorough study, upon which in many cases the success of the installation rests. Size and location of the stage are noted, for their bearing on placing of the horns and of the special screen needed. Where blue-prints of the theater building are not available, the engineer must measure the dimensions himself and must prepare sketches showing the size and shape of the auditorium and the number and arrangement of the seats. He must also appraise the acoustic characteristics, discovering any sections where there is an echo, interference, noise from an extraneous source, or any other hindrance to proper hearing. A final study of acoustic conditions is best made on completion of the installation, but data obtained during the preliminary survey give a useful basis for the installation work.

In the projection room the engineer notes the type and condition of the projection machines, the power supply and its regulation, the angle of projection. He sees whether there is room between the machines to allow convenient operation after installation of the sound apparatus, and whether there is a suitable space for the amplifier and the associated controls. In cases where the room is inadequate he decides upon the rearrangement or enlargement that is needed.

At the end of the study, the engineer's recommendations, along with the sketches and detailed information on which they are based, are submitted to the home office. There all the circumstances are examined again, and a report embodying a final set of recommendations is prepared.

When details of the installation are fully determined, dates are assigned for shipping the apparatus and for the start of the installers' work, and the theater management is informed. The management is in addition notified at this time if architectural changes are needed in the projection booth or elsewhere, so that they can be finished before the work of installation is commenced.

The entire work of installation is directed by a field engineer whose main function is to see that all parts of the work, and all supplementary work in preparing the theater, are so carried out that the completed system will operate at its full possibilities. He first establishes satisfactory contacts with the theater manager and staff, and with information obtained from them coordinates the work of installation, testing and rehearsal with circumstances necessary for operation of the theater during the preliminary period. Likewise he chooses an electrical contractor, if possible one already familiar with the wiring in the theater. Thus he insures that the work will proceed smoothly, and that everything will be ready for the scheduled opening.

Most commonly the work starts at the projectors themselves. While these remain practically intact in other respects, on each the driving motor is replaced with one whose speed is regulated electrically. As these motors require from four to five seconds to come up to running speed, it is desirable to have them in place early, that the projectionists may become accustomed to them and so be able to give full attention to handling the records by the time sound programs are to be presented. Next the batteries are installed, with their switching panel and charging apparatus, and a motor-generator if that is to be used. The rack for the main amplifier is placed in position, and the units mounted upon it. Next the fader is installed, and its auxiliary controls put in place and connected to it. Likewise the monitoring horn is mounted at a carefully chosen location in the operating booth, so that the projectionist will be able to hear the program clearly at all times and thus be prepared for the fader changes between records. Usually the turntable and pick-up for disc records are then installed on each machine and, for film records, the exciting lamp, lenses, photoelectric cell and associated amplifier, and subordinate equipment; in some cases however these members are assembled on the projectors before the main amplifier and fader are installed. Although the reproducing apparatus adds materially to the mechanism of a projector, the parts have been so designed for each of the commonly-used makes of machines that their installation does not involve loss or extensive change of parts not immediately involved. Since the conduit and wiring work have been keeping pace with the installation of equipment, it should now be possible to connect the pieces of apparatus and check their correct installation and the wiring work.

Installation procedure is constantly being improved, and within a few months a noteworthy advance is to be made. In the present arrangement the driving motor and disc turntable are mounted on individual pedestals, the film reproducer and associated amplifier are supplied as separate units to be mounted on the projector, and the film-discs transfer panel is mounted on the wall of the room. Instead there will be a single pedestal bearing all of these members. The change will facilitate the work of installation and reduce the wiring needed, and furthermore will make the work of operation more convenient as well.

With the work well advanced in the projection room, the engineer next turns his attention to the stage. The horns are placed on their mountings, which should be ready by this time, at locations presumably correct for the size and shape of the auditorium. After the loud-speaking receivers are attached and connected, there is the first opportunity to produce sound through the complete system. A cursory test follows to assure that no major errors are present, and any immediately obvious imperfections are corrected. Comparative tests are then made between the reproducers on the projectors, so that any discrepancies in volume can be overcome. Each receiver and horn is heard separately, and then they are operated in unison so that the engineer can check their proper poling. Before the work proceeds further, the semi-porous screen on which the pictures are to be shown is installed immediately in front of the horns. Through this the sound will pass, seeming to come from the pictures themselves.

Although for hearing of speech reverberation is objectionable as preventing good articulation, when not excessive it is desirable for music since it tends to give the effect of fullness and roundness.* A varying factor complicating the situation further is the absorbent property of the audience itself. There are theaters which empty, are entirely too reverberant, but which are thoroughly satisfactory when filled. Attendance varies at different performances of course, so that at best a compromise must be tolerated in considering acoustic adjustments.

While the use of ordinary drapes and similar material causes absorption of the higher frequencies to a somewhat greater degree than the lower, it is about the only practical method which can be employed to overcome acoustic difficulties from excessive reverberation and echo. Excessive damping must be avoided, to prevent a deadness in the sounds which may almost be depressing. Heavily damped houses require greater amplification than would otherwise be necessary but the presence of the audience has less effect than in a somewhat more reverberant house.

Correction of specific acoustic faults may be made with almost complete success, however. Echoes may sometimes be corrected by moving one or more of the horns, but the reflecting surface must be covered with heavy drapes when it is so located that no suitable echo-free position

* *For a consideration of auditorium acoustics see Speech Interpretation in Auditoriums by E. C. Wente,* Bell Laboratories Record, *October, 1928.*

is available for the horns. Interference usually results from reflection of sounds, and likewise can be prevented by adequate covering of the surfaces responsible. Resonance occurs when a surface made up of a thin, hard material, free to act as a diaphragm at its own natural frequency, is in the path of the sound waves. Distortion necessarily results, since only that part of the sound near the natural period of the resonator is reenforced. Here again covering is ordinarily the treatment used.

Bearing in mind the three prime requisites to good hearing, that sound be sufficiently loud for all auditors without being unnaturally loud for any, that successive sounds be clear and distinct, and that the components of complex sounds retain their relative intensities, the engineer makes the acoustic adjustments. He usually tests first for sound distribution, arranging the horns so that as far as possible all parts of the house receive the same volume. Proper flaring and tilting of the horns is usually sufficient to give adequate distribution. Any acoustic peculiarities are next investigated, and where possible corrected. Many hours are often spent in moving the horns and adjusting their placing, since slight changes in position sometimes produce marked differences in the acoustic effect. When the best results demand further acoustic treatment of the auditorium, involving alterations to the building or extensive change of the wall surfaces, the manager is so informed.

One standard practice followed in all cases is to cover the horns, and the back of the screen except the areas occupied with the horns, by absorbent drapes. It is usual also, where the screen is set back from the proscenium arch, to hang drapes around it, in shadow box effect. By these precautions back-stage echoes and reflections are avoided.

After the horn locations have been definitely fixed and the equipment given a final check, the engineer calibrates the amplifying apparatus for the particular theater. He has at hand several test records which make available a wide variety of the types of entertainment to be shown, each marked with the fader setting determined in a theater which has been chosen as "standard." By playing these records with the fader set at the marked value and adjusting the main potentiometer until the proper effects are obtained, he sets the potentiometer at a presumably permanent adjustment, in keeping with the size of the theater, so that in the future good results can be obtained from any records by adjustment of the fader only. In addition, various types of records require that varying volumes be obtained from the different horns. From the tests the engineer can choose the relative volumes for different types of selections, and establish the horn settings which will normally be required.

During progress of the installation, the projectionists and their assistants are instructed in the operation and maintenance of the new equipment. At the outset each man is given a complete operating instruction bulletin, which gives detailed information on every necessary point. By the time the theater is calibrated the men should be familiar with their new duties, and should be able to present a complete program without assistance. An opportunity soon comes to show their facility, in the rehearsal of the opening program.

At the early performances the closest attention is given, to be sure that the system is being operated to give the desired effects. In addition, however, watchful attention should be continued as a permanent function of the theater management, if the best results are to be secured. Whenever a sound program is in progress either the manager or somebody specially appointed should be present in the audience. By his own reactions, and by noting the effect on the audience, he observes the need for any changes in volume or otherwise. Likewise he watches for any extensive change in the degree to which the house is filled, or any other circumstance having a bearing on the picture as presented. A buzzer signal and a telephone are provided for communicating instructions to the operating room.

The installation engineer remains at the theater until it is evident that the local staff are fully competent to continue independently. At that time the installation is transferred to the service organization of Electrical Research Products, Inc. The service engineer in whose district the theater is located is available for emergency calls at all times. In addition, he makes scheduled visits to check the equipment and recommend any minor adjustments which may seem advisable. In this way each installation is given constant engineering supervision, to keep the sound equipment operating at the peak of performance which is possible.

Printed in U. S. A.

Dr E C Wente
with author's compliments

History of Sound Motion Pictures

EDWARD W. KELLOGG

SOCIETY OF MOTION PICTURE AND TELEVISION ENGINEERS
JUNE, JULY, and AUGUST 1955

History of Sound Motion Pictures

By EDWARD W. KELLOGG

Excellent accounts of the history of the development of sound motion pictures have been published in this Journal by Theisen[5] in 1941 and by Sponable[6] in 1947. The present paper restates some of the information given in those papers, supplementing it with some hitherto unpublished material, and discusses some of the important advances after 1930.

One of the numerous omissions of topics which undeniably deserve discussion at length, is that, except for some early work, no attempt is made to cover developments abroad. The subject of 16mm developments is discussed with a brevity altogether out-of-keeping with its importance. This has been on the theory that basically the problems are similar to those of 35mm sound, and that whatever has brought improvement to one has been applied to both.

Edison invented the motion pictures as a supplement to his phonograph, in the belief that sound plus a moving picture would provide better entertainment than sound alone. But in a short time the movies proved to be good enough entertainment without sound. It has been said that although the motion picture and the phonograph were intended to be partners, they grew up separately. And it might be added that the motion picture held the phonograph in such low esteem that for years it would not speak. Throughout the long history of efforts to add sound, the success of the silent movie was the great obstacle to commercialization of talking pictures.

Early Sound Pictures Using the Phonograph

The idea of combining recorded sound with the motion pictures is as old as the motion picture itself[33] (if we exclude the early "zoetrope" invented in 1833 by W. G. Horner).[39] In a paper, "What Happened in the Beginning," F. H. Richardson[7] reproduced a letter in which Thomas A. Edison quoted from his early notes: "In the year 1887, the idea occurred to me that it would be possible to devise an instrument which should do for the eye what the phonograph does for the ear, and that by a combination of the two all motion and sound could be recorded and reproduced simultaneously." The letter proceeds to tell of the development of the motion picture (and is followed by letters from Thomas Armat, George Eastman, C. Francis Jenkins and others, related to motion-picture inventions). Edison in 1895 tried on the public the combination of a phonograph with his "peep show" moving picture.[5,11] He built at least 50 (and probably more) of the combination machines.

Gaumont. Leon Gaumont, in France,[5] began as early as 1901 to work on combining the phonograph and motion picture. He worked on the project during several widely separated intervals. Theisen[5] refers to a series of shows of the "Film Parlant" at the Gaumont Palace in Paris in 1913 and to demonstrations in the United States. After 1926 the "Etablissements Gaumont" used the system developed by Peterson and Poulsen.

Presented on May 5, 1954, at the Society's Convention at Washington, D.C., by Edward W. Kellogg, Consulting Engineer, 276 Merion Ave., Haddonfield, N.J.
(This paper was received on October 25, 1954.)

Laemmle. An attempt by Carl Laemmle of Paramount in 1907 to exploit a combination of phonograph and motion picture is mentioned in Sponable's paper.[6] This was a German development called "Synchroscope." It was handicapped by the short time which the record would play, and after some apparently successful demonstrations, was dropped for want of a supply of pictures with sound to maintain programs in the theaters where it was tried.[13]

Pomerede, Amet, Bristol. Theisen's paper[5] mentions combinations of phonograph and motion pictures using flexible shafts or other mechanical connections, by Georges Pomerede[2] (1907 patent), and E. H. Amet[14] (1912 to 1918) who used electrical methods for the sound. Wm. H. Bristol[15] began his work on synchronous sound about 1917.

Siren Type of Amplifier. An ingenious attempt to obtain amplification in reproduction used the movements of the phonograph needle to vary the opening of an air-valve, connected to a source of air pressure. This device was employed for sound pictures by Oskar Messter[5,16] (Germany 1903–4). In England, where it was known as the "Auxetophone," it had some use for phonographs. Its invention is credited by the Encyclopedia Britannica to Short (1898), with improvements by the Hon. C. A. Parsons.

Edison. In 1913 Edison made a serious effort to provide synchronized phonograph sound. The equipment is on exhibit at the Edison Museum in West Orange, N.J. The phonograph is of special construction, to provide maximum volume and long playing, the cylinder record was oversize, and the horn and diaphragm considerably larger than those of home phonographs. Between the reproducing stylus and the diaphragm was a mechanical power amplifier, apparently using the principle of capstans used on shipboard. There was a continuously rotating amber cylinder and a hard rubber brake-shoe subtending about 130° of arc. One end of the shoe was connected to the reproducing stylus in such a manner that an upward displacement of the stylus would increase the pressure between shoe and cylinder; and the other end of the shoe was connected through a slender rod to the diaphragm, in such a way that the shoe movement resulting from increased friction would give an upward push on the diaphragm.[17] One may well imagine that the adjustment of this device to give substantial gain without producing chattering must have tested the skill of the best of operators. Nevertheless, it must have worked, for the record indicates that the Edison talking-picture show ran for several months in Keith's Colonial Theatre in New York, with much acclaim, and was shown in other large cities of America and in other countries.

The arrangement for synchronizing was not in accordance with present practices. The phonograph behind the screen determined the speed, being connected through a string belt to a synchronizing device at the projector. The belt pulleys were about 3 in. in diameter. The belt passed from the phonograph up over idler pulleys and overhead, back to the booth. The synchronizing device applied a brake to the projector, and the brake-shoe pressure depended on the relative phase of phonograph and projector, increasing rapidly as the projector got ahead in phase. With an even force

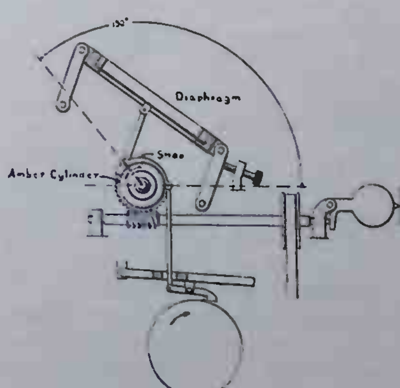


Fig. 1. Mechanical power amplifier of Thomas A. Edison and Daniel Higham.

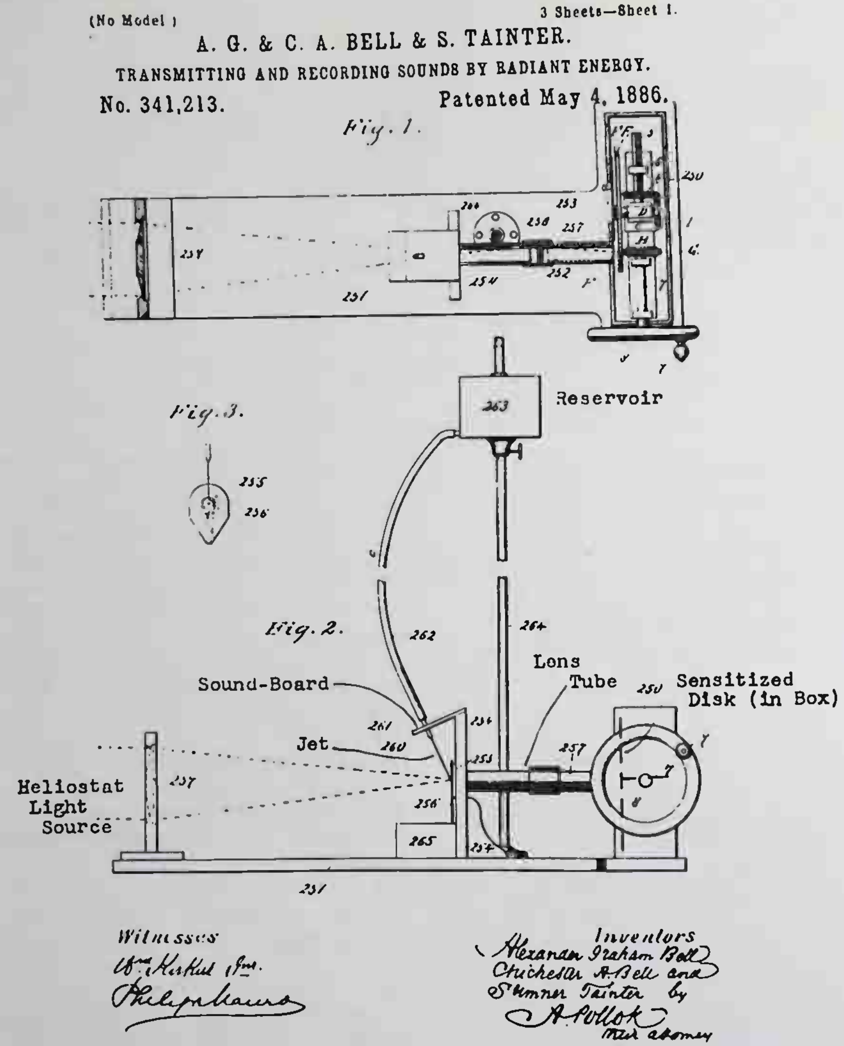


Fig. 2. Variable density recording system of A. G. Bell, C. A. Bell and Sumner Tainter, 1886.

on the projector crank, normal phase relation was maintained. The projectionist watched for synchronism and had a slight degree of control by turning the crank harder if the picture was behind or easing it off if it was ahead.

So far as I have learned, there were few further efforts (at least in the U.S.) to provide sound for pictures by means of phonograph (mechanical) recording until the Warner Brothers' Vitaphone system of 1926.

Photographic Sound Recording

A history of sound pictures necessarily includes the many efforts to record sound photographically, whether or not the experimenters made any attempt to combine the sound with pictures, or were even interested in that application. Despite the obvious advantages, from the synchronized-sound standpoint, of a photographic record of the sound on the same film with the picture, it does not appear that this consideration was necessarily an important factor in directing experimentation toward photographic recording, nor even that ultimate application to synchronous sound for motion pictures was (in many cases) a main objective. It was rather that photographic recording represented a new medium, which seemed to offer promise of much superior results. A mechanical system seems inherently crude where such delicacy is needed as in reproducing sound; in contrast to which recording by a beam of light would seem ideal. The experimenters have all been conscious of the handicap imposed by the necessity of making ponderable mechanical parts vibrate at high frequency.

So we find that efforts to record sound photographically began before there were such things as motion pictures on strips of film. Before the invention of the telephone, Alexander Graham Bell, interested in aiding the deaf, had made photographic records of "manometric flames," showing voice waves. His patent, No. 235,199, filed in 1880, shows a system for transmitting speech over a beam of modulated light, and uses a light-sensitive device (selenium cells) to detect the received fluctuations, thus anticipating the essential principle of the reproducing system which was used in many later experiments.

Blake. Prof. E. W. Blake of Brown University in 1878 made photographic records of speech sounds on a moving photographic plate, using a vibrating mirror.[6,18]

Fritts. U.S. Patent No. 1,203,190, filed in 1880 by Charles E. Fritts,[5,6] discloses photographic soundtracks and a great variety of devices for recording and reproducing, but there does not appear to be evidence of much significant experimental work.

Bell and Tainter. In the Smithsonian Museum in Washington, D.C., are a number of large glass disks carrying spiral sound tracks. These were made by a method described in U.S. Patent No. 341,213 (filed 1885) to Alexander Graham Bell, Chichester A. Bell and Sumner Tainter. Light from a steady source was transmitted in a relatively narrow beam through a piece of stationary glass, and then further restricted by a slit where it reached the circular photographic plate. Just above the place where the light entered the stationary glass, a tiny jet of ink (or other light-absorbing liquid) was directed against the surface. The nozzle was attached to a "sounding board" (small plate) which picked up the sound vibrations. The jiggles of the nozzle caused waves in the stream of ink which flowed down over the surface, and these modulated the transmitted light.

Some years ago it became desirable, in connection with a patent suit, to demonstrate that the spiral track was really a soundtrack. Contact prints (on celluloid films) were made of several of the most promising looking of the glass plates, and a reproducing system arranged, giving the record the benefit of modern equipment in this respect. The approximate best speed was found by trial. (The original recording machine was hand-cranked). The photographic image had suffered from age and was very noisy, and the total recording lasted only a few seconds. But it was with something of the thrill of an antiquarian that we listened to the voice from the past. "This is . . . . I am . . . in the . . . laboratory." The date was given too " . . . , eighteen eight- . . .?"

Others. Sponable's historical paper mentions numerous other workers and their patents. Several of these modulated the light by means of a small mirror connected to a diaphragm so that vibration caused rotation, thus anticipating features of equipment used by C. A. Hoxie in the work at General Electric Co. Of the developments which, although not leading to any commercial system, deserve special mention, I shall speak of several inventions or discoveries which laid foundations for later developments, and of the direct contributions to photographic recording of Rühmer, Lauste, de Forest, Reis and Tykociner.

Basic Inventions and Discoveries

Selenium Cells. For many years, reproduction from photographic-sound records was made possible by the selenium cell. The photoconductive properties of selenium were discovered by Willoughby Smith in 1873, and a practical selenium cell was made by Werner Siemens in 1876.[19] The response of a selenium cell to changes in illumination is sluggish, making it a very imperfect tool for sound reproduction, whereas the photoemissive effect on which photocells depend is practically instantaneous, but the electrical output from a selenium cell is very much greater.

The Photocell. The first indication of photoemission was discovered by Hertz in 1887 and later studied by Hallwachs (1888), Stoletow (1890) and Elster and Geitel (1889 to 1913).[19,20] Although by 1900 much had been learned, practical photocells did not become generally available till some years later, nor were they of help toward sound reproduction without electronic amplifiers.[21,22]

Thermal Emission — The "Edison Effect." Edison discovered in 1883 that a small current could flow through evacuated space in a lamp bulb, between a hot filament and a separate electrode. The Fleming "Valve," invented in 1905, made use of this principle, played an important part in early wireless telegraphy and was the forerunner of thermionic amplifiers.[26]

The Audion. The invention of the "Audion" by Lee de Forest in 1907 marked the beginning of the electronic era. As has been emphasized by many writers, it was the electronic amplifier which unlocked the door to progress and improvement in almost every phase of sound transmission, recording and reproduction. However, amplifying tubes did not become generally available to experimenters for over a decade. The de Forest patent[23] (acquired by the Telephone Company) was basic and unchallenged, but the vacuum techniques of some of the foremost laboratories of the country[24] were needed to make of the audion a dependable and reasonably rugged tool.*

The Oscillograph. The oscillograph, consisting of a small mirror mounted on a pair of conductors, close together, in a strong magnetic field, was invented by Blondel in 1891 and improved in 1893 by Duddell, who put it into practically the form still used. It has played a vital part in photographic sound recording.

Magnetic Recording. The invention by Poulsen of Copenhagen in 1900 of recording magnetically on a steel wire laid the foundation for modern tape recording, which has almost revolutionized methods of making original recordings.[27]

Auditorium Acoustics. The modern science of room acoustics and acoustic treatment dates from the work of Prof. Wallace C. Sabine of Harvard in the years 1895 to 1900.[28] With little other equipment than a whistle, a stop watch and brains, he worked out the acoustic principles on which successful sound recording and reproduction so largely depend.

Gas-Filled Incandescent Lamps. Beyond a certain point, optical-recording systems cannot give increased exposure by increasing the size of the source, but only by increasing the intensity (candles per square centimeter), which means higher temperature. Early incandescent lamps were well exhausted because all gas results in loss of heat by convection and hence lowered efficiency. In 1911–13 Irving Langmuir of General Electric Co. studied the effects of inert gas not only on heat loss, but also on the rate of evaporation of tungsten from the filament surface, which is the factor which determines permissible operating temperature. He showed that such gases as nitrogen, or better yet argon (the heavier the better), at pressures well up toward atmospheric or even higher, could with suitably formed filaments so retard the evaporation of tungsten that the higher permissible temperature much more than compensated for the added heat convection, thus giving several-fold increase in efficiency as well as whiter light. With the gas, the evaporated tungsten is carried to the top of the bulb instead of blackening the sides, in the optical path.[29]

Magnetic Materials. The development of several alloys of iron, nickel and cobalt having extraordinary magnetic properties is reported by H. D. Arnold and G. W. Elmen in the *Bell System Technical Journal* of July 1923, and by Elmen in the January 1929 and July 1929 issues. The extremely high permeability and low hysteresis of Permalloy have made it possible to greatly reduce distortion in transformers and in many electromechanical devices, and to provide more successful magnetic shielding than would otherwise be possible. In another alloy which has been called Perminvar, constancy of permeability and low hysteresis (making for low distortion) have been carried still farther. Another alloy named Permendur can carry very high flux densities before saturation, making it possible to produce intense fields which make for sensitivity and damping in devices of the moving conductor type.

Important for the reduction of cost and weight of magnetic devices was the discovery by the Japanese physicist T. Mishima of the properties of certain aluminum-nickel-cobalt alloys for permanent magnets,[30] and subsequent improvements.

Improvements in Vacuum Tubes and Phototubes. In any list of the advances which contributed in an important way to the technical attainments in modern sound reproduction, several improvements in amplifier tubes deserve an important place. Among these are:

(1) The Wehneldt (oxide coated) cathode and other low-temperature emitters, which in turn made indirectly heated unipotential cathodes possible.

(2) The screen-grid tube.

(3) The pentode.

(4) Remote cutoff or exponential tubes, and other variable gain tubes.

(5) The caesium phototube with its high sensitivity to infrared light.

(6) The gas-filled phototube with its increased output.

Early Work on Sound on Motion-Picture Film

Rühmer. Ernst Rühmer in Berlin[5,6,31] in 1901 began publication of the results of his work on photographic sound recording, which extended over a period of about twelve years. As sources of modulated light he superimposed voice currents on the continuous currents in electric arcs. He used considerably higher film speeds than those used for pictures. Sponable reported (ref. 6, p. 278) that some of Rühmer's Photographophon films were brought to this country by the Fox Film Corp., and that the articulation was clear; also, this reference shows a sample of Rühmer's soundtrack. A variable-area track by Rühmer is shown in the Theisen history (ref. 5, p. 421), the *Scientific American* of 1901[31] being cited as reference. Presumably Rühmer experimented with both systems.

Lauste. This Society has taken special note of the work of Eugene Augustine Lauste, in a 1931 report of the Historical Committee,[32] in a paper by Merritt Crawford[33] and in placing his name on the Society's Honor Roll. The young Frenchman joined the staff of Thomas A. Edison in 1887, where he did construction and experimental work till 1892. For two years he worked on another project and then, in association with Maj. Latham, developed a projector which was the first to incorporate the extra sprocket and free loops with the intermittent. Lauste's interest in photographic sound recording was first aroused when in 1888 he found in an old copy of the *Scientific American* (May 21,

* Much higher vacuum than de Forest had been able to obtain was necessary. This was independently accomplished by I. Langmuir of General Electric Co. and H. D. Arnold of Western Electric Co.[24]

1881) an account of Dr. Bell's experiment in transmitting sound over a modulated light-beam, and converting to electrical modulation by means of a selenium cell. This suggested the thought of recording the sound photographically on the same strip with the picture. It was not till about 1900 that he began to find opportunity to work on this project. He worked for several years in the United States and then went to England where he pursued his experiments. A British patent (No. 18,057, filed in 1906) shows a well thought-out system. Lauste received some financial backing in 1908 from the manager of the London Cinematograph Co.

To modulate the recording light, Lauste used rocking mirrors and what have been described as "grate-type light-valves." The mirror system was too sensitive to camera vibrations, and the grate-type valves which he was able to build had too much inertia. In 1910 he began working with modulators of the string galvanometer type, with excellent results. The historical account by Theisen,[6] shows photographs of some of Lauste's apparatus. He spent some time with Ernst Rühmer in Berlin, a stimulating and profitable association. He visited America in 1911 and as part of his demonstration made what was probably the first actual sound-on-film motion picture made in the U.S. A necessary return to England, shortage of capital, and the war, halted Lauste's sound-picture researches. In his paper on Lauste, Crawford expresses the thought that had it not been for this unfortunate interruption, plus very limited resources, and had electronic amplifiers been available to Lauste, commercialization of sound pictures might well have gotten started a decade before it actually did.

E. E. Ries filed application in 1913 for a patent (No. 1,473,976, issued in 1923) in which broad claims were allowed on the essentials of a single-film system. The patent became the basis of later litigation.[6]

Tykociner. In 1918 and following, Prof. J. T. Tykociner of the University of Illinois carried on experiments and developed a system. This work was described before the American Institute of Electrical Engineers and in the *SMPE Transactions*.[34] After pointing out that three new tools had in comparatively recent times become available for the solution of the sound-picture problem, (namely, high-frequency currents, photoelectricity, and thermionic amplifiers), Prof. Tykociner gives a broad discussion of requirements and possible arrangements. As a source of modulated light he used for the most part a mercury arc with either modulated continuous current or modulated high-frequency current, and for reproduction a Kunz (cathode of potassium on silver) photocell. The light from the mercury arc is particularly potent photographically, but is sluggish in following the input modulation, which results in some loss of the higher audio frequencies.

Foreign Developments Which Led to Commercial Systems

Tri-Ergon (meaning "the work of three"). Josef Engl, Joseph Massole and Hans Vogt, in Germany, began in 1918 the development of a system of sound pictures which later was commercialized under the name Tonbild Syndicat AG (abbreviated to Tobis).[6,35] They used a modulated glow discharge for recording, and a photocell for reproducing. Of chief concern in this country were the Tri-Ergon patents,[35] in which numerous claims allowed by the U.S. Patent Office were so broad that had their validity been sustained they would have almost swamped the industry. In particular, one patent (1,713,726) which claimed the use of a flywheel on the shaft of a roller or sprocket which carries the film past the translation point, to take out speed variations, was the basis of prolonged litigation, being finally declared invalid by the U.S. Supreme Court (1935).[114] But in the meantime the efforts to avoid what were thought to be dangerous infringements of the Tri-Ergon flywheel claims, had for seven years steered the course of mechanical designs on the part of the major equipment manufacturers into inferior or more complicated constructions. (See section on Mechanical Systems.)

In Germany the Tri-Ergon patents controlled the situation. The large picture producing companies, U.F.A. and Klangfilm (a subsidiary of Siemens & Halske and A.E.G.), took licenses under the Tri-Ergon patents. A brief account of the patent negotiations and agreements in this company and in Germany will be found in the Sponable paper.[6]

Peterson and Poulsen in Denmark developed a system (1923) which was commercialized in Germany under the name Tonfilm.[6] They used an oscillograph as the recording light modulator (giving a variable-area soundtrack), and a selenium cell for reproduction. (One of the Tri-Ergon U.S. patents[35] claimed the use of a photocell for this purpose, and it is likely that a German patent accounts for the use of a selenium cell by Poulsen and Peterson.) This system was used by Gaumont in France and by British Acoustic Films, Ltd.

The de Forest Phonofilm

Dr. de Forest tells the story of this work in the 1923 *Transactions*.[36] The account is particularly interesting because he tells much of his viewpoint as he started, and then, after describing the system which he had evolved, gives his reflections on the applications and future of sound motion pictures.

The man whose invention gave us amplifiers in which the heaviest object that had to be moved was an electron, surely had a right to wish to do away with moving mechanical parts in microphones, light-modulators and loudspeakers. For microphones he experimented with the conductivity of gas flames and of open arcs as affected by sound waves, and with fine platinum wires heated to a dull red by a direct current and subjected to the cooling effect of the air vibrations superimposed on a slight continuous air movement. The changes in resistance of the wires with variations of temperature gave rise to telephonic currents.

For light modulators he tried "the speaking flame" (probably the "manometric" flame of König) and a tiny incandescent lamp, carrying voice currents superimposed on direct current. The lamp was designed to have very rapid filament cooling (partly by using a short filament, so that heat conduction to the lead-in wires would be high). On listening to these sources by means of a photocell and amplifier, de Forest was convinced that they gave exceptional quality (even compared with the condenser microphone), but they proved entirely inadequate for making a useful soundtrack giving very small percentage of modulation and probably also underexposure. Finally a successful source of modulated light for recording was found in a gas-filled tube excited by modulated high-frequency currents from a 5- to 10-w radio telephone transmitter. This was named the "Photion." A slit, $1\frac{1}{2}$ to 2 mils wide and 3/32 in. long, adjacent to the film, was used to restrict the size of the exposing beam.

A similar slit was used in reproduction. Both potassium photocells and Case Thalofide[37,40] cells were used in reproducing equipment, the greater sensitivity obtainable with the Thalofide cell being a consideration offsetting the faster response of the photocell. The design and construction of amplifiers using his Audion were of course very familiar to de Forest.

Lament is expressed that loudspeakers depending on some principles other than diaphragms and horns were not to be had, but after some discouragements with "talking arcs" and sound radiators on the thermophone principle, the commercially available horn and diaphragm speakers were accepted as the only solution at the time.*

Practical models of recording and reproducing equipment were built, and recordings made, using principally a combined camera and recorder, and many demonstrations given.

* It is of interest that in the early part of our investigation which led to the direct radiator dynamic speaker (*Trans. AIEE*, 1925, p. 461) Chester W. Rice and I tried talking arcs and thermophones, and also a corona discharge device — all of which avoid mechanical moving parts — but none of these appeared promising.[38]

The de Forest paper[36] reviewed earlier history of efforts to record sound photographically, and gave appreciative acknowledgment of the help that had been given by Theodore W. Case.[37]

To have guessed wrong on some subject is no reflection on the insight of an experimenter, but several instances are striking, in the light of later developments. Speaking of the efforts to provide sound by means of the phonograph, the author said: "The fundamental difficulties involved in this method were so basic that it should have been evident from their inception, that commercial success could hardly be achieved in that direction." (Consider the Warners' Vitaphone.) Speaking of loudspeakers, after saying that the loudspeaker has been developed "to a high state of perfection" but left much to be desired, he said: "I am convinced that final perfection will come not through any refinements of the telephone and diaphragm, but by application of entirely different principles." (Yet phenomenal improvements were made with the identical elements, through refinements.)

In speaking of the future of sound pictures, Dr. de Forest gave a definite "No" to the question whether the existing type of silent drama could be improved by the addition of voice. But he foresaw the evolution of an entirely new type o dramatic scheme and presentation, taking advantage of the freedom which had been such an asset to the silent moving picture (as contrasted with the stage) but using sound and voice where these could be effective. He also had visions of great utility for travel films, newsreels, records of notable persons, and educational films.

The work just described was done from 1918 to 1922. About a year and a half later[36] Dr. de Forest gave a brief account of progress, reporting improvements in many details, better articulation, thirty theaters equipped, much interest on the part of operators, films made of a number of celebrities and contracts with leading chain exhibitors. Again the opinion was expressed that the talking picture would not ever take the place of the silent drama.

The Phonofilm system was used in numerous theaters, with sound films made under Dr. de Forest's direction; but he did not succeed in interesting the established American picture producers. Perhaps the industry was prospering too well at the time, but judging from the initial coolness of film executives to the technically greatly improved systems a few years later, it is easy to imagine that numerous imperfections which undoubtedly existed (as, for example, defective film-motion, limited frequency range, and loudspeakers that gave unnatural voices, and perhaps too, demonstration films that were uninteresting) contributed to loss of the impressiveness needed for doing business.

Several years later the "de Forest Phonofilm Co." was bought by Schlesinger of London and South Africa.

Work at the Theodore W. Case Laboratory (Movietone)[6]

Theodore W. Case[37] became interested in modulating light and deriving telephonic currents from it in 1911, while a student at Yale. In 1914 he organized his laboratory at Auburn, N.Y., devoting special attention to the study of materials whose resistance is altered by light, of which selenium was the best known example. These studies resulted (1917) in the development of the Thalofide cell, in which the photosensitive material is thallium oxysulfide.[40] These cells, which are especially sensitive in the near infrared range, were widely used in Navy communication systems during World War I. Case was joined in 1916 by E. I. Sponable. Experiments were continued with the help of an Audion amplifier obtained from de Forest. One of Case's postwar developments was the barium photoelectric cell.

In 1922 attention was turned seriously to sound pictures. Manometric* flames (oxyacetylene) were tried as a possible source of modulated light. Soon afterward Case found that the light from an argon arc in one of the tubes that had been used for infrared signalling could be readily modulated and was photographically potent. These tubes had oxide-coated hot cathodes. A tube for recording, based on this principle, was developed and named the Aeo-light.[6,41] It operated at between 200 and 400 v. Helium was substituted for argon in 1922, with benefit to the actinic power and also to the speed with which the light followed the current variation. The commercial Aeo-lights were rated at 350 v.

From 1922 to 1925 Case cooperated with de Forest, furnishing numerous items of experimental equipment.

Several sound cameras were built under the direction of Sponable, in 1922, 1923 and 1924. The 1924 model was a modified Bell & Howell camera rebuilt to Sponable's specifications by the Bell & Howell Co. The film motion in this and other cameras was unacceptable until they had been reworked for greater mechanical precision. In the final designs of sound camera the sprocket was driven through a mechanical filter, consisting of damped springs and a flywheel on the sprocket shaft. The sound was recorded on the sprocket. The Aeo-light was mounted in a tube which entered the camera at the back. Directly against the film was a light-restricting slit made by silvering a thin quartz plate, ruling a slit 0.0006 in. wide in the silver, and cementing over it a thin piece of glass which was then lapped to a thickness of about 0.001 in. The slit was thus protected from collecting dirt from the film. The end of the Aeo-light, where the glow was concentrated, was close behind the slit. A Bell & Howell contact printer was modified to make possible the independent printing of picture and sound.

Up to the fall of 1925, when the working arrangement with de Forest was terminated, the Case laboratory efforts were directed largely to recording principles and apparatus. It was decided then to work on a system independently of de Forest, and one of the next projects was to build reproducing equipment in the form of an attachment which could be used with existing picture projectors. It was in this design that the decision was reached to place the soundhead under the projector, and the offset of 20 frames or $14\frac{1}{2}$ in. between picture and sound was established. The speed of 90 ft/min was adopted for the Case system. In the first projector attachment a light-restricting slit was used similar to the one used in the camera, but later a straight tungsten filament was imaged on the film, and in a still later model, a concentrated straight-axis helical filament was imaged on a slit which was in turn imaged on the film.

With the essential elements of a sound-on-film system developed, Case and Sponable began study of the patent situation, with a view to obtaining licenses, if necessary, for the commercial use of their system. There appeared to be no very strong patents to interfere, except those on the use of thermionic amplifiers. A contract between General Electric, Westinghouse and Radio-Corporation on the one hand and Western Electric Co. on the other, was in effect, specifying the fields of activity in which each might use amplifiers, but, if I have not misinterpreted the account in Sponable's historical paper, sound-pictures had not been specifically mentioned, and there was some question as to the right to license use in the Case system, the eventual decision being that both groups had rights. The Bell Telephone Laboratories were interested themselves in developing sound pictures, and so were not immediately ready to license what would be a competing system. However their engineers were much interested in the performance attained, and there was some thought of combining efforts. There were demonstrations of both systems, but no plan to merge them was reached. The experience of Case and Sponable at

* A gas jet so arranged that sound vibrations produce changes in the gas supplied to the jet.

General Electric Co. was rather similar.

In 1926 demonstrations were made to representatives of the Fox Film Corp., who became greatly interested, and finally to William Fox. After thorough testing on their own premises, the Fox Film Corp. purchased rights to the Case developments (July 23, 1926), leaving the question of amplifier rights to be worked out later. The Fox-Case Corp. was organized to exploit the system, which was given the name Movietone. Courtland Smith, who had been with the Fox Film Corp. and had been instrumental in bringing about the purchase, was made president of the Fox-Case Corp. The Movietone News service was established.

Sponable left the Case organization to give his services to the new company, one of the first of his activities being the design of recording studios in New York and later in Hollywood. In 1927 he developed a screen which transmitted sound freely, permitting loudspeakers to be located directly behind the picture. The first public showing of Movietone recordings was in January 1927.

The Fox-Case Corp. obtained license to use amplifiers, first in 1926 through the Western Electric Co. and the Vitaphone Corp., and the next year revised contracts were made with Electrical Research Products, Inc. (ERPI), which was formed in January 1927 to handle the sound-picture business for the Western Electric and Telephone companies.

In the Movietone reproducing system, Western Electric amplifiers and loudspeakers were used. The years 1928 and 1929 were marked by rapid expansion in facilities and personnel, successful showings and stepped-up schedules of newsreel releases. In March 1929 the making of silent pictures by Fox was discontinued. Six months later the Fox and Hearst newsreel services were united.

The British Movietone News was organized in 1929. In 1930 William Fox sold his interests in Fox Film and Fox Theatres.

As the Fox Film Corp. was already an ERPI licensee, and therefore had rights to use other Western Electric developments, the Western Electric light valve was adopted for the Movietone service (as well as for Fox studio recording), displacing the Aeo-light.

Work at Western Electric Co. and Bell Telephone Laboratories

The Western Electric Co. brought to a commercial stage almost simultaneously a sound motion-picture system based on disk records, and one based on sound on film. Various developments which laid the foundations for these systems had been taking place through a number of years. The citation of the life and work of Edward B. Craft in this *Journal*[42] indicates that his interest and enthusiasm were in large measure responsible for the undertaking of a full-scale project for developing systems of sound for motion pictures. Craft was assistant chief engineer of the Western Electric Co. from 1918 to 1922, when he became chief engineer. With the transfer in 1924 of research activities to the newly organized Bell Telephone Laboratories, Craft was made executive vice-president, and continued to guide activities.[42]

Whether or not there was a definite policy of not putting all of the eggs in one basket, work on both systems was stepped up at about the same time (1922) and pushed with equal vigor.

The two systems had identical requirements with respect to many elements, but, in particular, microphones, amplifiers and loudspeakers. The Western Electric Co. had acquired rights to de Forest's Audion in 1913 and made great improvements in it during the next few years, building up wide experience in its applications and circuitry.

Second only to electronic amplifiers in importance for the development of high-quality recording and reproducing systems was a microphone of uniform response and with low distortion. With amplifiers available Dr. E. C. Wente[43] was able largely to ignore the question of output level, and to develop by 1916 a microphone of the condenser type, having extraordinarily high fidelity and freedom from distortion and noise.[44–47]

In the loudspeaker field, the company had had considerable experience and had developed units for public address work. The public address installations had afforded experience with auditoriums and requirements for intelligibility, while experience in acoustics for sound pickup had been gained in radio broadcasting.

With respect to the recording itself and reproduction, I shall separate the two stories of the disk and photographic systems.

The Disk System

In 1946 there was published a history of sound recording in the laboratories of the Western Electric Co.[48] Since the transmission of speech was the main business of the Telephone Co., a program of studying every aspect of speech waves was initiated about 1912, and as part of this project, efforts were directed to recording the sound. The interest soon spread to include music. In connection with work with disk records, Crandall and Kranz built an electromagnetic reproducer in 1913. In 1915 H. D. Arnold suggested that the improvement of disk recording be undertaken, using the then available electrical equipment (which included amplifiers). By this time the electrical reproducer had been improved.

The war interrupted these projects, but they were resumed soon after its close. A group under J. P. Maxfield undertook the improvement of wax recording and the phonograph. The story of this development was told in 1926 to the American Institute of Electrical Engineers.[49] The recording system made use of a magnetically driven cutter so designed that with constant current input, the vibratory velocity of the cutting stylus was substantially constant from about 200 to 5000 cycles, while from 50 to 200 cycles the amplitude was constant, a characteristic practically necessary to avoid overcutting by the low notes. Two features of the design were of special interest: (1) the separation of the total mass that must be driven into three parts (armature, stylus-bar and coupling disk), connected together through portions of shaft whose torsional flexibility was carefully calculated to make of the structure a mechanical low-pass filter of calculable mechanical impedance; and (2) a mechanical resistance consisting of a thick-walled rubber tube (which may be thought of as practically a rod of soft rubber) subjected at one end through the coupling disk to torsional vibrations. The propagation of torsional waves in such a soft rubber rod is so slow that in a length of about 6 in. there would be many wavelengths for all but the lowest frequencies.

Vibrations imparted to the rubber reach the far end very much attenuated, are reflected, and propagated back toward the start, but are of neligible magnitude when they reach it. Under such conditions the rubber line acts as a nearly pure mechanical resistance to load the filter, and, if properly matched to the filter impedance, results in practically complete (and therefore uniform) transmission through the filter structure, throughout the frequency band below the filter cutoff. The features just described are, I believe, the inventions of H. C. Harrison. The great improvement in records which electrical recording brought, is well known to all of us.

Without a better reproducing system than the phonographs of the types in use about 1920, the improvements in the records would have been largely lost, so there was developed a greatly improved (nonelectrical) phonograph called the Orthophonic (also largely the outcome of H. C. Harrison's approach to the problem). However this part of the program had no direct bearing on the talking-picture project. In early 1925 the Columbia and Victor Companies took licenses from Western Electric Co. to use the recording methods and apparatus, and to build phonographs of the Orthophonic type.

Sound-on-Disk Synchronized With Pictures. Little time was lost in trying and demonstrating synchronized sound and pictures using the new electrically recorded disks. Craft arranged for a demonstration at Yale University in 1922 and another in February 1924, the equipment and many details of the system having been developed and improved in the interval.

To provide sound for pictures, using the disk-record system,[50] it was necessary to have records which would play continuously for at least the projection time of a 1000-ft reel (about 11 min), to plan a synchronous drive, and to use electrical reproduction in order that, with the help of amplifiers, adequate sound output could be had.

It was not desirable (in view of background noise) with record materials then available, materially to reduce amplitudes of cuts, and so groove pitch had to be kept nearly the same as then in current use (about 100 grooves per inch). To maintain quality the minimum linear groove velocity must not be reduced. With a given groove pitch and minimum velocity, the maximum playing time for a given record diameter is obtained by recording to half the maximum diameter, and the required playing time determines the needed size and corresponding rotation speed. While the engineers could take some leeway, the choice of 16 in. outside diameter and $33\frac{1}{3}$ rpm, approximately met the conditions indicated.

For synchronous recording, the camera and the recording turntable can be driven by selsyn motors, which driving system gives the equivalent of both being geared together and driven from one shaft. Starting marks on both film and disk are of course essential.

For reproducing, the turntable and projector were mechanically geared together. A simple magnetic pickup, if not damped, has a high-frequency resonance in which the armature whips, giving excessive output and high mechanical impedance at the needle tip.[51] The magnetic pickup used in the sound-picture system was designed for use with replaceable steel needles and damped by enclosing the moving elements (except the needle-holder and needle) in oil.[52]

The turntable driving systems[52–54] evolved for the sound pictures are discussed in the section on "Mechanical Systems" — the great problem being (as had been the case throughout the history of sound recording) to obtain sufficiently nearly constant speed.

The loudspeakers which had been developed for public address applications[55] were of the "balanced armature" type, had good power-handling capacity, and were regarded as fairly satisfactory from the standpoint of articulation. Designs of horns had been evolved which fairly successfully controlled the directivity for auditorium purposes. In 1923 Dr. Wente built a speaker of the moving-coil type which gave greatly improved quality[56] (especially the better bass response which is possible with the moving-coil drive), but in terms of efficiency and power-handling capacity it was not satisfactory. It was not until 1926 that a speaker of the moving-coil type was developed by Wente and Thuras[57] which met the requirements for quality, efficiency and power-handling capabilities. Speakers of this design rapidly superseded those of earlier design, and continued in use for years.

According to the account of Lovette and Watkins[48] the sound-on-film system, on which another group of engineers had been engaged, was capable in 1924 of matching the quality of the disk system, but the latter represented an older art in which there were fewer uncertainties. The greater confidence with which the company could offer the disk system, and with which a potential customer would consider it, were responsible for choosing it as the first to be pushed. However, interest on the part of most of the picture producers was cool, nor did Craft, conscious of the numerous failures of previous efforts by others, think it desirable to hasten the commercialization of either system until its weaknesses were worked out.

Samuel Warner and Vitaphone.[58] With many details omitted, the foregoing is the description of the sound-on-disk system which became known as Vitaphone. Col. Nathan Levinson,[48] then serving the Western Electric Co. in the Pacific district where he had had close association with Samuel L. Warner, made a business trip to New York early in 1925 and saw a demonstration of the sound pictures. He felt sure that Mr. Warner would be interested, and arranged for a demonstration at the first opportunity. Samuel Warner was more than convinced, and his enthusiasm quickly spread to his brothers. More thorough tests were arranged, using cameramen, technicians and artists of the Warner staff, in cooperation with Western Electric engineers. The adoption of sound by a large picture-producing company would mean a huge outlay, and its success was a question not only of technical performance, but of the artistic, dramatic and psychological results which could be achieved through the addition of sound. The tests were convincing to the Warner Brothers, if not to the executives of some other picture companies who witnessed them. To develop and market sound motion pictures and equipment, the Vitaphone Corporation was organized in April 1926, with Samuel L. Warner as its president.

The first major Vitaphone sound picture to be released was *Don Juan*,[1,48] (August 1926) in which music by the New York Philharmonic Orchestra was featured. The new loudspeaker developed by Wente and Thuras was ready in time for this. Preparations were made for producing sound pictures in Hollywood, where sound stages were erected embodying the recommendations of the foremost experts in acoustics. The production of *The Jazz Singer* with Al Jolson, was begun in the spring of 1927 and it was shown in New York on October 6. Its success was such that the industry was convinced "overnight" that the day of sound pictures had arrived.

Improvements in the Disk System. Under the title "Recent Advances in Wax Recording"[59] H. A. Frederick tells of a number of advances subsequent to the 1926 account by Maxfield and Harrison. By improvements in record material and wax processing techniques, it had been possible to reduce surface noise by 3 to 6 db. A new pickup (4A) is described with smoother response and good to about 4500 cycles, as compared with 4000 cycles for the previous model. A response curve for the commercial recorder shows practically uniform response to 5500 cycles. Laboratory models of recorder and reproducer are mentioned as carrying the response to 7500 cycles. The new recorder used a longer rubber damping line. Frederick gives the groove pitch as 10 mils and the minimum groove velocity as 70 ft/min. He also reported very satisfactory results with re-recording.

Western Electric Sound on Film

Mention has been made of fundamental studies of speech waves, begun in 1912 and carried on through several years until interrupted by the war. Amplifier tubes became available as laboratory tools in 1913. Photographic records of speech waveshapes were made, using at first a carbon transmitter, an amplifier and a Duddell oscillograph. The weakest link in this chain of equipment was the transmitter, whose response varied greatly with frequency and which had a high level of background noise, making it difficult to get reliable traces of consonants and other relatively weak speech sounds. The development of a better transmitter was one of the first undertakings of Edward C. Wente,[43] who came to the company in 1914.[44–47]

The Condenser Transmitter. If the charge on a pair of condenser plates is maintained through a sufficiently high resistance, the voltage is directly proportional to the separation of the plates, so that a transmitter based on this principle is an amplitude-sensitive device. If the diaphragm, which is one of the condenser plates, is so stiff in relation to its mass that resonance occurs above the required frequency range, the diaphragm deflection is proportional to the instantaneous air pressure. Wente met this mechanical requirement by using a stretched steel diaphragm 0.002 in. thick and spaced 0.001 in. from a relatively massive backplate. The very thin layer of air contributes greatly to the stiffness of the diaphragm, but the flow of air through the narrow space toward and from a relief space around the edges causes damping, so

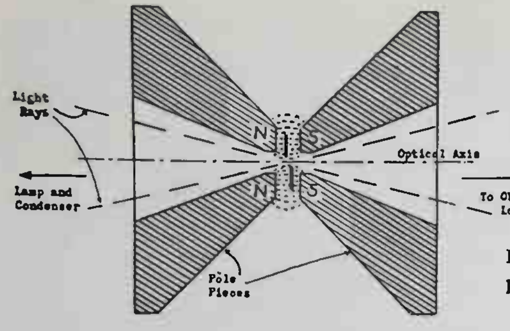


Fig. 3. Light-valve ribbon and pole piece arrangement; section at right angles to ribbons.

that a nearly flat (uniform) response was obtained up to about 15,000 cycles.

Wente left the company in 1916 for graduate study and returned in 1918. In the meantime Dr. I. B. Crandall had made a theoretical analysis of the air-film damping, and improved the instrument by means of grooves of appropriate size and shape in the backplate.[45] For measurement purposes it was essential to calibrate the condenser transmitter, and Wente accomplished this by working out the theory of the thermophone, which enabled him to make a reliable pressure calibration.[46] Free field calibrations were made later, using a Rayleigh disk as reference. In a later design,[47] which was used commercially for sound recording, the sensitivity was greatly increased, in part by use of aluminum alloy 0.001 in. thick instead of 0.002 in. steel for the diaphragm, and in part by not carrying the response as far into the high-frequency range. (In 1931 W. C. Jones published a pressure calibration curve for a #394 transmitter which showed a rapid drop above about 7000 cycles.[47]) The condenser transmitter is rated as a very insensitive device, but it is of interest that a diaphragm deflection of a millionth of an inch will give a fifth of a volt, the gradient in the space between electrodes being 200 v per mil. It is the extreme stiffness of the diaphragm which makes the sensitivity low.

Photographic Recordings. The condenser transmitter with amplifier gave better waveshape traces, but the narrow mirror of the bifilar (or Duddell) oscillograph causes diffraction effects which make the light-spot at the film blurred or fuzzy. Prof. A. C. Hardy showed[59] that this trouble could be largely eliminated by radical changes in the optical system in which the oscillograph vibrator is used, but his analysis was not published until 1927 (in time to be of much help in the General Electric recording developments, but the Western Electric experiments with the oscillograph were before 1920).

An article in a British Journal (1920) came to Wente's attention, describing experiments of Prof. A. O. Rankine in transmission of sound over a beam of light. The light modulator, in which a rocking mirror caused an image of one grating formed on another grating to move transversely to the bars, appeared well adapted to making photographic records of the variable-density type. While a variable-density record would not give as much information to the eye as a variable-area record, it could be analyzed by instruments of the microdensitometer type. The faithfulness of the recording could be checked by playing it back. (The previous oscillographic recordings had not been designed for playing back.)

Some of the recordings were played in May 1922 for Craft and others. A few months later apparatus-development engineers were requested to construct an electrically interlocking driving system for camera and recorder. Further demonstrations were given in December 1922. In these recordings the principle was recognized, that for linear relations between exposing light and print transmission, the product of positive and negative "gammas" should be unity.[61, 62]

Light Valve. The grating type of modulator had several drawbacks, one of which was diffraction by the grating. Because of these difficulties, Wente in January 1923 proposed using a two-string light valve.[63–65, 4] Such a valve was ready for test a month later. The tension on the ribbons was adjusted to bring their resonance to 6500 cycles. Condensing lenses imaged the light source on the slit between the ribbons, and an objective lense imaged the valve slit on the film.

Results with the light valve were definitely better than with the previous modulators, and arrangements were made for tests on a larger scale. A recording studio was set up in 1923 and sound pictures made for demonstration purposes.

In the latter part of 1922 and subsequently, much of the study of film emulsions, exposures and developments was carried on by Dr. Donald MacKenzie. He showed that by running the lamp at slightly over-voltage, it was possible adequately to expose positive film, which thereafter was the standard sound-recording stock. The relatively fine grain of the positive stock was of great benefit from the standpoint of resolution and low background noise.

In 1928 MacKenzie described the light-valve model in use at the time, and recording and processing practice (exposure ranges and developments) as worked out at the Bell Telephone Laboratories.[64] The valve is mounted with the slit between ribbons horizontal — so that its image on the film is transverse to the film. The ribbons are in a strong magnetic field and currents in the two are in opposite directions, so that they are deflected (edgewise) to increase or decrease their separation depending on the direction of the current. The width of the slit with no current in the ribbon was 0.002 in., and it was masked to a length of about 0.2 in. It was imaged on the film with a 2 : 1 reduction. With the slit width 0.002 in., the light could be modulated 100% by a vibration of each ribbon of 0.001 in. amplitude. Since the ribbon need be only slightly wider than its double amplitude, thick enough to be opaque, reasonably easy to handle and long enough between supports to make the deflection substantially uniform throughout the length of the slit, it can be extremely light and readily put under enough tension to place its mechanical resonance above the required audio range. Rather than attempting to control the resonance by damping beyond that obtainable electromagnetically, an electrical low-pass filter was used in the input, to prevent the passage of any impulses of high enough frequency to excite the resonance. However the cutoff was not too far below the frequency of resonance to permit a considerable rise in amplitude just before cutoff, the maximum being at about 7000 cycles. This rise was regarded as advantageous in that it compensated for loss of high-frequency response due to image spread in the film. For monitoring, a photocell behind the film picked up some of the light which went through the film.

The subject of sensitometry for soundtracks of the variable-density type also received attention from many other writers for a number of years after the advent of photographic sound.

In the matter of the frequency range attained in the early light-valve recordings, MacKenzie shows an overall (light-valve input to photocell output) curve which was substantially flat to 5000 cycles, a figure not far from what could be obtained at the time with disks.

Recorder. The Western Electric recording machine employed a sound sprocket, having a filtered drive and protected by a feed sprocket from jerks from the magazines.[4] The film was exposed while on the sound sprocket. For synchronism the camera and recorder were driven by selsyn motors.

Soundhead. For reproduction from photographic soundtracks the Western Electric Co. built a "soundhead," to be mounted under the picture projector,[52–54] similar in many respects to that previously mentioned as used in the Fox-Case development. I shall come back to the subject of the mechanical features of the film-motion system, so shall mention here only some optical and electrical features. The scanning light on the film was an image of a mechanical slit, illuminated by a low-voltage incandescent lamp, with condensing lenses. The filament was a close-wound helix with straight horizontal axis. The photocell and preamplifier were cushion-mounted to prevent microphonic noises. Owing to the very high impedance of the photocell and its small output, a very short (low-capacity) connection to the first amplifier tube is important. The preamplifier brought the level up to about equal to that of the disk pickups.

Standard Speed. In the early theater installations most projectors were equipped for both disk and film reproduction. It was obvious that for sound pictures the recording and reproducing speeds must be closely held to a standard. The practice had become widespread of projecting silent pictures at considerably higher speeds than that of the camera, which had for years been nominally 16 pictures/sec or 60 ft/min. The higher projection speeds shortened the show so that more shows could be run in a day, and the public had become inured to the fast action. But there was a better justification in that flicker was much reduced.

For pictures with sound on film there was further benefit from increased speed in that it resulted in better high-frequency response and, in some degree, reduced percentage of speed fluctuation. A speed of 85 ft/min for silent pictures had been recommended for a standard, but practice varied widely. A speed of 90 ft/min or 24 frames/sec was chosen for both of the Western Electric sound-picture systems (sound on disk and sound on film) and this became the standard. On the theory that exhibitors would demand the option of running silent films at other speeds, the Western Electric engineers adopted a driving system with an accurate control which could be made inactive at the option of the projectionist.[54] Either a repulsion motor or a d-c motor might be used. For 90 ft/min a 720-cycle generator fed a bridge with one arm tuned to 720 cycles. At the correct speed the bridge was balanced, but if the speed was not correct the unbalance gave rise to a correcting current which increased or decreased the motor speed as required.

Commercialization. In January 1927 Electrical Research Products Inc. was formed as a subsidiary of Western Electric and the Telephone Co. to handle commercial relations with motion-picture producers and exhibitors.

The adoption of sound systems by the motion-picture industry (except for the case of Fox Movietone and Warner Vitaphone) is discussed in another section of this paper.

Developments at General Electric Co.

Interest in photographic sound recording at the General Electric Co. in Schenectady stems from the development prior to 1920 of a photographic telegraph recorder for radio reception,[66] by Charles A. Hoxie. Transoceanic radio service was by long waves, and static interference caused the loss of many letters. It was thought that a visual record of the incoming signals, even though mutilated by static, might be deciphered at leisure in many cases in which the signals were forever lost if the operator, depending on ear alone, failed to recognize a letter.

For the usual reception, by ear, the incoming continuous-wave code signals were heterodyned to give interrupted tones of audio frequency, short for dot and longer for dash. Hoxie's recorder made an oscillographic record of these code signal tones, on a moving strip of sensitized paper. Instead of actuating a receiver diaphragm the electrical signals vibrated a reed armature, which, through a delicate knife-edge arrangement, imparted rotary motion to a mirror, which caused a small spot of light to dance back and forth across the sensitive strip.

Since the code recorder vibrated at audio frequency, it was a short step to try it and modifications of it for recording voice, and this was one of the many experiments which Hoxie tried which started him on more systematic experimentation in the field of photographic sound recording. Negative film was used at first, in order to get adequate exposure, but Hoxie was among the first to appreciate the advantage of the finer-grain positive film.

As in the case of the telegraph recorder, the track ran down the middle of the film, and was nearly an inch in width. In Hoxie's recording and reproducing machine the film was drawn over a physical slit on which intense light was concentrated. The width of the slit was about 0.001 in. Since an open slit would quickly fill with dirt, a wedge of fused quartz was ground to a thin edge and cemented in place between the metal edges which formed the slit. The face against which the film was to run was then lapped and polished. A photocell close behind the film picked up the transmitted light, and an amplifier and loudspeaker completed the reproducing system. The results were highly gratifying. Theisen[6] says that Hoxie's first sound recorder was completed in 1921, and with it he recorded speeches by President Coolidge, the Secretary of War and others, and the recorded speeches were broadcast over Station WGY (Schenectady) in 1922.

Hoxie called his optical phonograph the Pallophotophone, meaning "shaking light sound." We do not know the identity of the Greek scholar. In another experimental development, Hoxie caused the vibration of a sound-pickup diaphragm to rock the mirror. This device, called the Pallotrope, was used with a photocell as a photoelectric microphone.

Narrow Sound Track Found Sufficient. Hoxie continued his experimenting for several years before any decision was reached to embark on an all-out program of developing a system of sound for motion pictures. One of Hoxie's experiments which undoubtedly played a part in interesting executives in such a program was that of reproducing with part of his track width masked. The development of the General Electric model of the Duddell oscillograph had centered in the General Engineering Laboratory (where Hoxie worked) and it was extensively used as a laboratory tool throughout the company. With such a background it would be natural to think of a photographic sound track as showing the outlines of the sound waves.

In any case the wide soundtracks made in the Hoxie equipment were of the variable-area type. A spot of light moved parallel with the slit, illuminating a larger or smaller fraction of its length. However, the active edge of the light spot was by no means sharp. While experimenting with reproduction from this sound track, Hoxie observed that masking off part of the track had little effect on the sound except some reduction in volume. He repeated the experiment with still more of the track masked off, until he was using only a sample, about $\frac{1}{16}$ in. wide. This experience was sufficient to demonstrate that a track wide enough to show the wave outlines was by no means necessary for sound reproduction. The narrow strip being scanned was obviously a variable-density record of the sound.

At that early stage of the experimenting we had not seen it demonstrated by actual accomplishment that a satisfactory variable-area recording could be confined within so limited a band, but at any rate this test proved that a photographic sound record could be placed along the side of the picture without stealing more picture width than could be tolerated.

Loudspeaker and Phonograph Developments. Another factor which undoubtedly influenced General Electric executives toward increased interest in sound was the success of the loudspeakers developed by C. W. Rice and myself for broadcast radio reception.[38] The coil-driven (or "dynamic") paper cone, freely suspended, surrounded by a baffle and driven by an amplifier with adequate undistorted power, so far surpassed its predecessors in quality of reproduction that within a few years its use for radio receivers and phonographs became practically universal.*

Following the loudspeaker development, the success of the electric phonograph helped to make the sound motion picture seem like a logical next project.

Chester W. Rice. I trust that I will be excused if I take this opportunity to pay a brief tribute to my colleague, whose vision and initiative were largely responsible for our undertaking the loudspeaker project. His thoroughness and tireless energy insured that no hopeful lead was left unexplored. He brought to bear on his work an extraordinary measure of ingenuity and mastery of engineering and physical principles, which he was constantly supplementing by study, and his standards of excellence would permit no compromise with an inferior result.

No one could have been more scrupulously fair and generous in giving credit to other workers. His death in 1951 was a great loss to his associates and to science.

C. W. Stone's Leadership. In addition to L. T. Robinson, head of the General Engineering Laboratory, the man who played the major role in initiating and promoting a large-scale project for developing talking pictures, was C. W. Stone, manager of the Central Station Dept., who had taken great interest in all of the sound developments. His enthusiasm, confidence and influence encouraged those who were engaged in development, helped to secure the financial backing and established fruitful contacts outside the company.

Practical designs; Assistance of Prof. A. C. Hardy and L. E. Clark. When, about 1925, a program of developing commercial sound-on-film equipment was undertaken, Robinson was made responsible for the general program, and, together with others in the Research Laboratory, I was asked to assist in problems where there seemed to be call for research. Engineers in the General Engineering and Research Laboratories had had experience in sound, first with loudspeakers[38] and then in cooperation with the Brunswick Balke Callender Co., electrical recording and reproduction for phonographs[51] (the work represented in the Brunswick Panatrope[51] and the Brunswick electrically recorded disks). Our part in the phonograph project was tapering off, freeing some of the personnel to devote time to the newer development. Our group, however, had inadequate background in optics and photography. Professor A. C. Hardy was engaged as consultant and soon did us two invaluable services: he straightened us out on a number of optical and photographic questions, and he recommended that we engage the services of L. E. Clark, then completing some advanced work at Massachusetts Institute of Technology. "Pete's" presence was a guarantee that we would not again get off the beam on optical questions, but his associates at General Electric, then at Photophone headquarters in New York, and later in Hollywood, carry a memory of something far more cherished than his valuable technical help.

Variable-Area System Chosen. A fundamental question on which we took Prof. Hardy's advice was in regard to the advantages of the variable-area type of soundtrack.[61] At the time of Hoxie's tests with a masked track, the only tracks that had been made, sufficiently narrow and still fairly satisfactory, were of variable density. A better understanding and application of optical design was needed to make clear, sharp-edged variable-area tracks within permissible limits of width.[59,60]

With the right kind of lenses and optical design, an imaged slit soon displaced the contacting physical slit with which the first tracks had been made. Hoxie's special galvanometer was not adequately damped, but General Electric had long since been building oil-damped oscillographs of the Duddell type, whose response was good up to 5000 cycles. The optics of the recording system are similar in principle to those of the oscillograph, as explained in one of Hardy's papers.[59] Prof. Hardy had shown how important design improvements could be made, greatly increasing the light intensity at the film. An optical system was designed[60] using a regular oscillograph galvanometer, and following suggestions of Prof. Hardy and of L. E. Clark.

The general mechanical features of the first recording machines were due principally to Hoxie, while H. B. Marvin (of the General Engineering Laboratory) designed amplifiers, optical systems and other necessary equipment. High-quality microphones were available in the Western Electric Condenser Transmitter (developed by E. C. Wente of the Bell Laboratories)[44-47] which was used in broadcast studios and had been an essential tool in the loudspeaker[38] and phontograph developments.[51]

General Electric had a well established motion-picture laboratory under the direction of C. E. Bathcholtz, for general company and publicity service, so that with the cooperation of that department, pictures with sound could be made. A number of demonstrations were given in 1926 and 1927, using this equipment. Motion-picture producers showed interest, but no contracts were made at that time.

An incident of much interest to those who were connected with the photographic recording project was a visit to Schenectady in December 1925 by E. I. Sponable from the Case Laboratories.[f] He showed and demonstrated the combined camera and sound-recording system which he and his associates had developed, giving us the benefit of his experience and participating in some demonstrations. However, no arrangements for combining the efforts resulted.

The Road-Show Wings. The first public entertainment picture to be shown, with the General Electric developed sound system, which by this time had been named the Kinegraphone, was a story of the Air Force activities in World War I, entitled *Wings* and produced by Paramount. The sound effects were added after the picture had been shot. The system and equipment were demonstrated and briefly described by H. B. Marvin.[67]

Wings was exhibited in 1927 as a "road show" (about a dozen sets of equipment having been supplied), for few motion-picture theaters at the time *Wings* was shown were equipped for optical sound reproduction. Multiple-unit cone-and-baffle type loudspeakers[38] were used, with a bank each side of the screen. The sound-reproducing device or "head" was mounted on the top of the projector, no standard sound offset having been established at the time the apparatus was designed. The picture width was reduced from 1 in. to $\frac{7}{8}$ in. to make room for a soundtrack. Ninety ft/min had by this time been agreed upon for film speed.

There were many, even of the most enthusiastic advocates of sound-picture development at General Electric, who did not think of the chief function of the synchronized sound as giving speech to actors in plays, but there was high confidence that there was a large potential market for sound systems for furnishing sound effects and background music and providing voice for lectures and speeches.

G.E.–Westinghouse–RCA Working Arrangements

At the time that the synchronized sound development was taking shape, the three-cornered arrangement between General Electric, Westinghouse and RCA was in effect. RCA was the sales outlet for all radio and kindred equipment. Manufacturing was divided between General Electric and Westinghouse. Research and development continued to be carried on at both manufacturing companies, and before production was started, designs were coordinated between them and had also to be acceptable to RCA, which maintained a Technical and Test Dept. in New York, to pass on performance.

At Schenectady, in view of the prospects of manufacturing on a much larger scale than could be handled in the General Engineering Laboratory, the film project had been transferred (1927) to the Radio Dept. where it was under the direction of E. W. Engstrom. The change brought new personnel into the activity. The names of E. D. Cook and G. L. Dimmick deserve mention.

Developments at Westinghouse

Engineers at the Westinghouse Electric and Manufacturing Co. in East Pittsburgh did not turn their attention to photographic sound recording until about 1926 when the project at Schenectady had gained some momentum.

One of the first research projects undertaken was to adapt the Kerr cell to photographic recording. The development was described to this Society in 1928 by V. K. Zworykin, L. B. Lynn and C. R. Hanna.[68] Nitrobenzene has the property of rotating the plane of polarization of a light beam, when the liquid is subjected to an electrical field at right angles to the direction of the light. The amount of rotation depends on the square of the field gradient. Practically, several hundred volts per millimeter are required. Nicol polarizing prisms are used on each side of the cell and rotated to extinguish the light at minimum applied voltage. With increase of voltage, the transmitted light then varies as the sine of the increase in angle of rotation.

One of the design problems is to keep within satisfactory limits the distortion resulting from the nonlinear relation between voltage and transmitted light. Another difficulty is that commercial nitrobenzene is yellow, absorbing the photographically valuable blue light. The investigators were able by double distillation to reduce very largely the absorption of blue light. A third problem was avoidance of electrical arcs through the liquid, which quickly contaminate it. Proper choice of electrode material and surfaces, and purification of the liquid made it possible to produce cells which were regarded as practical.

The unique property of the Kerr cell light modulator which makes it of special interest is its extreme speed. The only limitation is in the ability of the modulation-voltage supply system to charge the extremely small capacity of the cell. As contrasted with this, other light-modulation systems either involve moving mechanical elements, or electrical discharges through gases, which have definite frequency limitations.

Zworykin, Lynn and Hanna were in the Westinghouse Research Laboratory, which was under the direction of Mr. Kintner. A group under Max C. Batsel was responsible for development and design of commercial equipment. One of this group was J. D. Seabert, whose contribution to the theater loudspeaker problem will be described in the paragraph with that heading. Hanna's analysis of the damped flywheel problem[69] laid the foundation for the highly successful rotary stabilizer discussed under that heading in the section dealing with Mechanical Systems.

Organization of RCA Photophone, Inc.

RCA Photophone, Inc. was organized in 1928 as an RCA subsidiary to carry on commercial exploitation of the sound-on-film system. Carl Dreher (later with RKO) was its first chief engineer, followed in 1929 by Max C. Batsel from the Westinghouse Co. A laboratory was established in New York to which a number of engineers were transferred from the Technical and Test Dept. of RCA.

New Designs of Commercial Units. Between the launching of the *Wings* show and the offering by RCA Photophone, Inc., of a commercial sound system,* a number of design changes and advances had been made. C. L. Heisler had designed a new recording machine (R-3) and a combined picture and sound projector (P-2),[70] both of which embodied new principles in film motion. A sound attachment or "soundhead" was developed, by which existing silent projectors could be adapted for sound. The offset between picture and sound had meantime been standardized at $14\frac{1}{2}$ in., with the soundhead mounted under the projector. Because of the much more stringent requirement for accurate and constant speed for sound than for picture, the driving motor was made part of the soundhead, and the projector mechanism driven from the soundhead through gears. The first commercial soundhead to be offered by the RCA group (designated as PS-1) was of Westinghouse design, but the manufacturing was carried on by both companies.

Theater Loudspeakers. The flat baffle type of loudspeaker[38] used in the *Wings* equipment and in almost universal use for home receivers, while excellent for music and sound effects, had not proved satisfactory for *speech* reproduction in reverberant theaters. While a certain kind of directivity can be had by using arrays of direct-radiator loudspeakers, vibrating in phase, this did not confine the radiation in the direction of the audience as successfully as the use of short horns. The first successful units of this type were developed by J. D. Seabert in 1929 (then of Westinghouse). The horns used at first expanded from about the cone area to an opening abut 3 ft by 4 ft. The name "directional baffle" was used to distinguish these horns, whose primary function was to confine the radiation within a limited angle, from the small-throat horns whose basic function was to load the diaphragm, in addition to confining the radiation. The directional baffle type of unit was the subject of later developments by Dr. H. F. Olson and his associates.[71-73]

In spite of the benefits of directive baffles, in many motion-picture theaters satisfactory speech reproduction was not achieved until absorption had been applied to reduce reverberation.

Location Equipment. The RCA equipment also included a truck for location and newsreel service.[75] With batteries for power supply, the truck carried a motor generator for driving apparatus designed for 60-cycle operation, and a studio-type film recorder, to be driven in synchronism with a cable-connected camera. For more remote or inaccessible locations, a single-film system was provided, with portable batteries and amplifier, governed direct-current camera motor, and a sound attachment, mounted on the top of the camera.[74] The first commercial uses of RCA Photophone recording equipment were for newsreel service. Two types of light modulator were employed in the earliest Photophone single-film location equipments, one of which used a galvanometer designed by W. O. Osbon and K. A. Oplinger, under the direction of C. R. Hanna, with optics generally similar to those of the studio system, and the other the Kerr cell (or Carolus cell) system developed by L. B. Lynn and V. K. Zworykin.[68]

Location equipment (sound trucks) of improved design followed within a short time. Of special interest was a new optical system requiring only 3 w for the lamp.[76]

Disk Equipment. Although the RCA group was convinced of the inherent advantages of sound on film for motion-picture sound, disk equipment was wanted in all of the earlier theater installations, and accordingly combined sound-on-film and synchronous disk equipment was designed and built by the G.E. and Westinghouse companies and supplied by RCA Photophone, Inc.

A number of developments and inventions took place at both of the manufacturing companies which did not come into commercial use for several years, and these will be described presently.

Commercialization. The establishment of commercial relations with picture producers is described in the latter part of the following section.

* Many of the elements of this type of loudspeaker, such as coil drive, cone diaphragms and the baffle had been proposed individually by early inventors, but not in the full combination. Nor, I believe, was the principle of placing the mechanical resonance of the diaphragm (with its suspension) at or below the lowest important frequency proposed, except that Adrian Sykes (U.S. Pats. 1,711,551 and 1,852,068) advocated it for a microphone. The Farrand loudspeaker (U. S. Pat. 1,847,935, filed 1921. See Radio Club of America, Oct. 1926) had a large cone, coil-drive and low resonance-frequency, but no baffle or associated power amplifier. It had considerable commencial success during the 1920's.

* H. B. Franklin in *Sound Motion Pictures*[2] gives May 14, 1928, as the date of an announcing advertisement in New York and Los Angeles papers; however the Progress Report, *Trans. SMPE*, No. 31, 438, May 1927, states that Photophone equipment is to be sold direct to theaters, and that recording efforts would be concentrated on music scores.

The Motion Picture Industry Adopts Sound

Many Commercially Unsuccessful Efforts. The historical outline with which our story began contains a very incomplete account of the many efforts to combine sound and picture, some of which attained a fair degree of technical success, elicited praise and held public interest for short periods. We mentioned the work of Edison, Lauste, Rümer, and de Forest, and might add Pathé Frères and Léon Gaumont* in France.[5] Many of these were ahead of their time, for without amplifiers, the production of adequate and natural sound was practically impossible. Even after amplifiers became available the experimenters had little better success in getting picture producers seriously interested. The article by Lovette and Watkins[48] states that by the end of 1924 practically every major producer in Hollywood had rejected Western Electric's sound-picture system.

Economic Hurdles. The same authors give such a convincing statement of the financial obstacles from the producer's standpoint that I cannot do better than quote them:

"The motion picture producers had large inventories of silent films, which had cost millions to produce. They had great numbers of actors and actresses under long term contracts, most of whom knew no dramatic technique except that of pantomime. The industry was universally equipped with stages and studios suited only to the silent film technique.

"Moreover, world-wide foreign markets had been established for silent films. To serve these markets, it was merely necessary to translate the words printed upon the film from English to any language desired. Finding stars and supporting casts who spoke the various languages of the world, or finding ways to give the illusion of their speaking them, appeared to be an insuperable task.

"The art of the silent film had attained superb quality and the public was satisfied. Why then, producers asked, should Hollywood scrap the bulk of its assets, undertake staggering conversion costs, and force upon the public a new and doubtful experimental art?

"Nor were the exhibitors equipped for sound. Many, it was argued, would not be able to meet the cost of sound picture equipment."

These obstacles would not have prevented the producers from introducing synchronized sound, had they been convinced that it would give their pictures greater appeal. A factor which many developers of sound equipment probably did not fully recognize, was that to contribute to the illusion, the sound must have a degree of naturalness far surpassing that which had sufficed for simply transmitting information, or making words understood.

How It Looked in 1926–7. To many, the silent motion picture, with its freedom of action, its settings for much of its action in natural backgrounds, was better entertainment than stage drama, and when one tried to imagine what a talking motion picture would be like, one's thoughts immediately turned to examples of theater drama. I have already quoted some of Dr. de Forest's reflections. The prevailing thought at the General Electric Co. as our system began to take shape is probably typical. Many, even of the most enthusiastic advocates of the sound-picture development were not convinced that the chief function of the synchronized sound would be to give speech to the actors in plays. The art of telling stories with pantomime only (with the help of occasional titles) had been so highly developed, that giving the actors voices seemed hardly necessary, although readily possible. Such a view was actually a very high tribute to the movie makers of the silent era. However, a very large business in synchronized sound seemed assured (even without any use of the system for dialogue) in furnishing sound effects, background music, and providing voice for lectures, speeches and travelogue commentary.

As one who shared in this misjudgment, I would like to suggest to readers that it is difficult today to divest oneself of the benefit of hindsight. At that time, the principal examples of sound pictures we had seen were demonstration films, very interesting to us sound engineers working on the project, but scarcely having entertainment value. None of us had seen a talking motion picture with a good story, and picture and script well designed for the purpose. When in 1927 such a picture was shown (*The Jazz Singer*) the story, the music and the dialogue were splendidly adapted to produce a fascinating picture with great emotional appeal, in which no element could have been spared without serious loss. In short, the excellence of showmanship played no small part in making it clear to everyone who saw it that the day of "Talkies" was here.

The Jazz Singer and its predecessor *Don Juan*, it might be noted, had the benefit of a newly designed loudspeaker,[67] very much superior to those used in the Western Electric 1924 demonstrations.

Warners and Fox Take the Step. Warner Brothers committed themselves to the adoption of sound pictures in 1926, license contract being concluded in April, followed by large investments in sound stages and equipment. In July of the same year the Fox Film Corp. became committed, forming the Fox Case Corp. which took license for the Case Laboratory developments in April, and in December from Western Electric Co. for rights to use amplifiers. Both Warners and Fox operated theater chains. With two major picture producing and exhibiting organizations definitely launched on a program of making and showing pictures, could the other great picture companies remain on the sidelines?

Large Producers Agree to Choose Same System.[6] Early in 1927 the first Fox Movietone Newsreel subjects were shown. The other picture companies must by this time have become convinced that sound pictures were inevitable, for a part, if not the whole of motion-picture entertainment. In February 1927, the "big five" — M-G-M, First National, Paramount, Universal and Producers' Distributing Corp. (or PDC), jointly asked the Hays organization to study and make recommendation as to what system should be adopted. The Movietone and Vitaphone (disk) had already become commercial systems, Western Electric was offering a sound-on-film (light-valve) system, and General Electric had made a number of demonstrations of a variable-area system (later offered to the industry with some modifications through RCA Photophone). There had as yet been no formal standardization, and those participating in the conference probably felt some uncertainty about interchangeability of recordings. It is not strange that the picture companies thought it would be advantageous for all to adopt the same system.

By far the most ambitious demonstration of sound motion pictures that had as yet (February 1927) been witnessed was the Warner Vitaphone *Don Juan* (shown August 1926),[1, 43] with performances by noted artists and score and background music for the play by the New York Philharmonic Orchestra. And the sound quality was good. But it was a demonstration of synchronized sound, and not of sound motion-picture drama. The producers, still "on the fence," continued their "watchful waiting."

The presentation of *The Jazz Singer* in October 1927 dispelled all doubts. But whether the future lay with the disk or the film system was a question not completely settled for several years.

"Big Five" Sign Contracts with ERPI.[6] With such large producers as Warners making pictures with sound on disk and Fox with Movietone releases on film, it appeared that exhibitors might be saddled with a dual system. Perhaps it was the hope that one or the other would very soon forge ahead in the race that caused further hesitancy, but in April and May of 1928 (about six months after the showing of *The Jazz Singer*) Paramount, United Artists, M-G-M, First National, Universal and several others signed agreements with Electrical Research Products Inc. (the commercial outlet for the Western Electric systems) for licenses and recording equipment.

Getting Started.[1] There followed a period of feverish activity in erection of sound stages, and procurement and installation of recording channels and equipment. Deliveries of apparatus were far behind the desires of the customers, and there was great shortage of engineers and technicians with sound-picture background. The manufacturers and associated organizations lent or lost many of their personnel. Intensive training courses and much instructive literature alleviated the situation. The Transactions of the SMPE for the fall of 1928 are little short of an encyclopedia of sound recording and reproduction by both disk and film. To this body of literature, the engineers and processing laboratory experts from the producing companies soon began making their contributions.

Scarcely a step behind the building and equipping of recording studios was the installation of sound reproducing systems in theaters. Theater chains controlled by the picture-producing companies which had already signed contracts, used sound systems of the corresponding make, but the business of furnishing sound equipment to the great number of independent theaters was competitive between ERPI, RCA Photophone and many other suppliers. An idea of the rate of growth of the sound pictures, may be had from the following figures given in Sponable's paper.[6] At the end of 1927 there were some 157 theaters in the U.S. equipped for sound, of which 55 were for both disk and film and 102 for disk only. At the end of 1928, of the 1046 ERPI theater installations, 1032 were for disk *and* film. By the end of 1929 ERPI had equipped about 4000 theaters in the U.S. and 1200 abroad, and RCA Photophone had equipped some 1200 in the U.S. and 600 abroad, most of these being for both disk and film. The SMPE Progress Report of February 1930 states that at the time, Hollywood studios were producing only 5% silent pictures. Installations by other manufacturers brought the total number of theaters equipped for sound in the U.S. to over 8700. There were at the time 234 *different types* of theater sound equipment including the large number which were designed for disk only. At the end of 1930 there were about 13,500 theaters equipped for sound, and about 8200 not equipped, according to the SMPE Progress Report of August 1931.

Contracts for Photophone Variable-Area Recording. In 1928 RCA bought the theater chain interests of B. F. Keith and of Orpheum, and the film producing company Film Booking Office or F.B.O., and organized Radio Keith Orpheum or RKO. The new company (RKO), with Photophone equipment, and drawing heavily on the RCA group for much of its initial sound personnel, made many feature and shorter pictures, using the name Radio Pictures for its product. RCA Photophone made arrangements for license and equipment with Pathé Exchange Inc., Mack Sennett, Tiffany Stahl and with Educational Pictures Corp.

One of the first feature pictures made by Pathé was *King of Kings* directed by Cecil de Mille. The Pathé Newsreels were an important item, using a number of RCA mobile recording equipments or "sound trucks."

Disney switched to the RCA Photophone system in January 1933. Republic Pictures Inc. used the RCA system beginning October 1935 and Warner Brothers in June 1936. Columbia Pictures Inc. began May 1936 to use the RCA variable-area system for part of its operations, but continued for several years to release on variable-density.

Cinephone. The Powers Cinephone system was developed by R. R. Halpenny and William Garity for Patrick A. Powers, who financed the project. It was basically similar to the system of de Forest, with whom Powers had permissive contracts. Cinephone was put on the market in September 1929 and used for several years by Walt Disney and others.

Type of Contract. Most of the initial contracts between the equipment-manufacturing companies and the picture producers were on a lease (rather than outright sale) basis, for a stipulated term of years, with equipment servicing and engineering assistance as part of the suppliers' obligation, and royalties depending on the film footage recorded.

Evolution of a New Art, Under Difficulties.* The idea that the silent motion picture would continue to have its place in theater entertainment died hard. What *The Jazz Singer* had proved was that with a suitable story and presentation, a sound picture could have an appeal far beyond what was possible without sound. It had not proved that sound would help in *all* types of presentation. In March 1929, Fox discontinued making silent pictures. In speaking of this in his historical paper[5] of 1941, W. E. Theisen calls it a daring decision, "since a large number of the leaders of the industry still felt that sound films were only a passing fad." In "The Entertainment Value of the Sound Movie" (*Trans. SMPE*, No. 35, 1928), H. B. Franklin, President of West Coast Theatres, says: "The silent motion picture is too well established. . .to vanish because of this new development."

It took time, much work and some mistakes for the industry to learn to use sound to full advantage, and the great pressure under which writers and producers worked during the years of transition was not conducive to best results. Two quotations from 1928 papers are illuminating. In "The Public and Sound Pictures" (*Trans. SMPE*, No. 35) Wm. A. Johnson, Editor of *Motion Picture News*, speaks of the great demand for sound pictures, and says: "The present hastily turned out crop of talkies are for the most part crude and disappointing." In "Reaction of the Public to Motion Pictures with Sound" (*Trans. SMPE*, No. 35), Mordaunt Hall, motion-picture editor of the *New York Times*, describes the shortcomings of many efforts as due to stories not adapted to talkies, actors who didn't articulate, or had poor voices, and misjudgments in production.

We tend, fortunately, to forget the troubles that are past. Still more do we forget the troubles other people had. We who took part in the development of sound equipment may be tempted to think that we made the talking picture possible. But if we give the credit they deserve to the writers, directors, actors and their bosses, and to the patient guinea pigs who bought tickets, perhaps the only bouquet left to hand ourselves is to say that our stuff was not so bad as to make the talkies impossible.

Mechanical Systems

Of all the tell-tales that remind the listener that the sound he hears is from a record and not "live pickup," the most unmistakable is that due to speed variations — known as "wow" or "flutter," and it is probably the most painful and devastating to realism. The importance of correct and constant speed was recognized by Edison and all his successors in sound recording, but standards were not very high. Phonographs sold despite their shortcomings. But sound for pictures could succeed only by providing better entertainment than silent pictures. In those systems which gained eventual acceptance by the motion picture industry, the engineers spent much effort on providing constant speed. In his story of the development of the Fox-Case system, for example, Sponable[6] tells of having to rebuild cameras, and of mounting a flywheel on the sprocket shaft and driving

* Gaumont, in addition to many inventions and other activities, was a pioneer and successful leader in the motion-picture business, and probably came nearer to success with phonograph sound than others. See account, and references given in the Theisen history[5] from which ref. 77 is taken.

* Many excellent discussions of the requirements for the new form of entertainment have been published. One such is Chapter IX "Comments on Production," of H. B. Franklin's *Sound Motion Pictures*.[2]

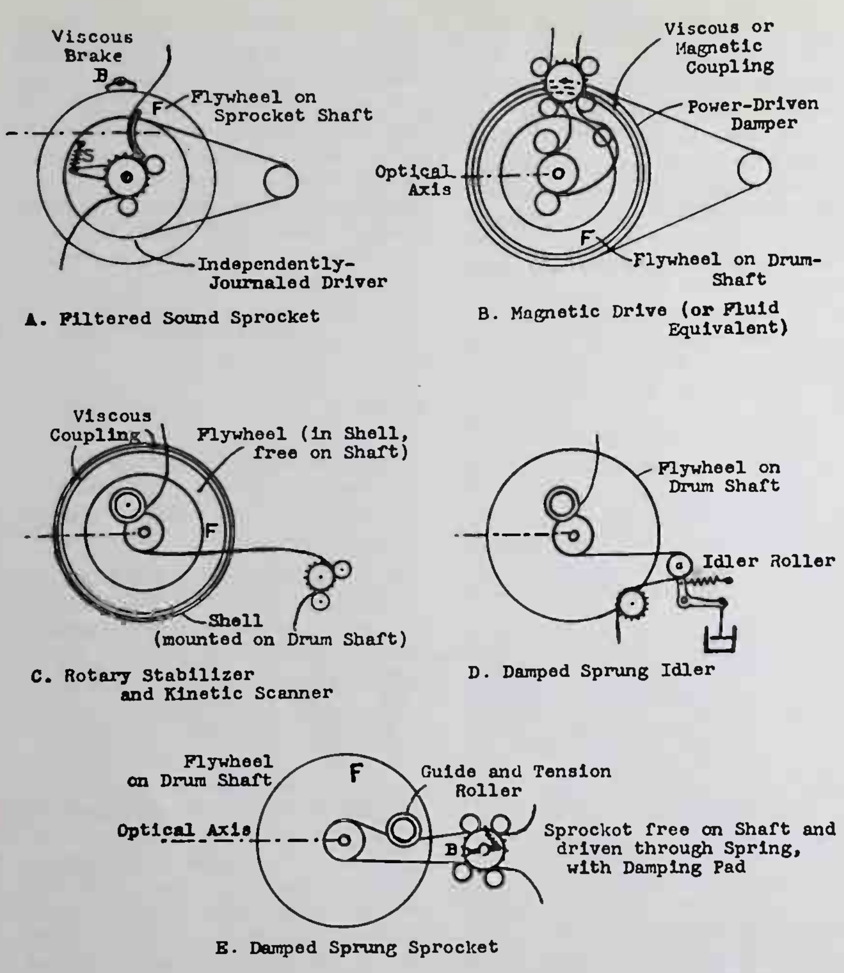


Fig. 4. Mechanical filter systems for reducing irregularities in film motion.

the combination through damped springs.

The literature dealing with speed fluctuations has been devoted largely to discussions of measures for improving the performance of recorders and reproducers in this respect.[70] Until the recent important contribution by Frank A. Comerci,[84a] such information as has been published regarding subjective thresholds or tolerances has been limited largely to continuous tones. Further systematic quantitative studies with typical program material are very desirable. There is no question however that all the present and future improvements in equipment performance are well justified in terms of more satisfying sound reproduction. Some of the more general discussions of the subject will be found in the literature.[70,79–84a]

Wow Meters. Of prime importance toward improving recording and reproducing machines is ability to measure the departures from uniform speed. One of the first such meters was built about 1928 by M. S. Mead[85] of the General Engineering Laboratory at Schenectady. It was improved by H. E. Roys and used extensively at Camden, N.J., being the basis of the flutter-measuring equipment described by Morgan and Kellogg.[86] This meter made an oscillographic recording of the fluctuations. An extremely simple and light-weight indicating flutter bridge used in RCA servicing is described in the *Journal*.[87] Flutter-measuring instruments are described by Scoville.[88] These are of the indicating type with band filters, by which flutter at different rates can be separated. Another design is described by Herrnfeld.[89] A widely used wow meter designed by U. R. Furst of Furst Electronics, Chicago, has been commercially available since 1947 or earlier.[90]

Disk System. In the disk system the change from 78 to 33⅓ rpm increased the difficulties, for at the low speed even a very heavy turntable (although very helpful toward eliminating rapid flutter) was not a practical answer. A flywheel driven through springs, or what we call a "mechanical filter," was a well-known expedient, but such a system is oscillatory and will multiply rather than reduce the speed fluctuations if the disturbances are of a frequency anywhere near that of the resonance, unless the system is damped by adequate mechanical resistance.[69,70,81,91,94,100,102,104,105] The requirement that the transient disturbance of starting shall disappear in not more than one revolution is more difficult to meet with extremely large inertia. The acceptable 33⅓-rpm reproducing turntables had much more inertia than had been customary for 78-rpm machines, and were driven through springs, with enough damping to reach equilibrium reasonably quickly, and dependence was not placed on making the natural frequency low in comparison with that of the slowest disturbance (once per revolution). Damping in some designs was provided by applying friction to the springs,[52,53] and in others by a viscous drag on the turntable. In either case it was essential to have high indexing accuracy in the low speed gear or worm-wheel.

For 33⅓-rpm recording turntables, the Western Electric engineers went to extraordinary refinement.[70,91] On the theory that it would not be practically possible to produce gears with no eccentricity or indexing errors, they made their 33⅓-rpm worm-wheel in four laminae, all cut together in one operation. Then they separated and reassembled them, each rotated 90° with respect to its neighbor. Each had its own spring connections to the turntable. Damping was by means of vanes in oil. Four vanes were rigidly connected to the turntable, while the pot and four other vanes were driven from the gears through a system of equalizing levers (which might be compared to whiffletrees) which imparted to the pot and its vanes a rotation which was the average of that of the four gear laminae. The effect of this was to divide by four the magnitude of each disturbance due to imperfection in the cutting of the gear, but to make it occur four times per revolution instead of once, and both of these effects are helpful toward filtering out irregularities.

Filtering Systems for Film. In a very judicial appraisal of the relative advantages of film and disk, P. H. Evans[92] speaks of the disk system as giving better speed constancy. He was of course referring to the experience up to the time of writing. There can be no question that film presents a more difficult problem. Synchronous drive and the maintenance of free loops require that it be propelled by sprockets. In the earlier systems of driving the film, it seems to have been regarded as sufficient to provide constant rotational speed for the sprocket (often called the "sound sprocket") which carries the film through the point of recording or reproduction. To obtain such constant sprocket speed it was practically necessary to use mechanical filtering to take out irregularities originating in the gearing.[4,70]

But the spring-driven sprocket was very sensitive to jerks from the film, so that it was necessary to employ extra sprockets with slack film between to isolate the filtered sound sprocket. It was also necessary to have an unusual degree of precision and concentricity in the sound sprocket. (Fig. 4A).

But there remained the question of what imperfections there might be in the film perforations, or how much it had shrunk since the holes were punched. Shrinkages up to 1% were not uncommon.

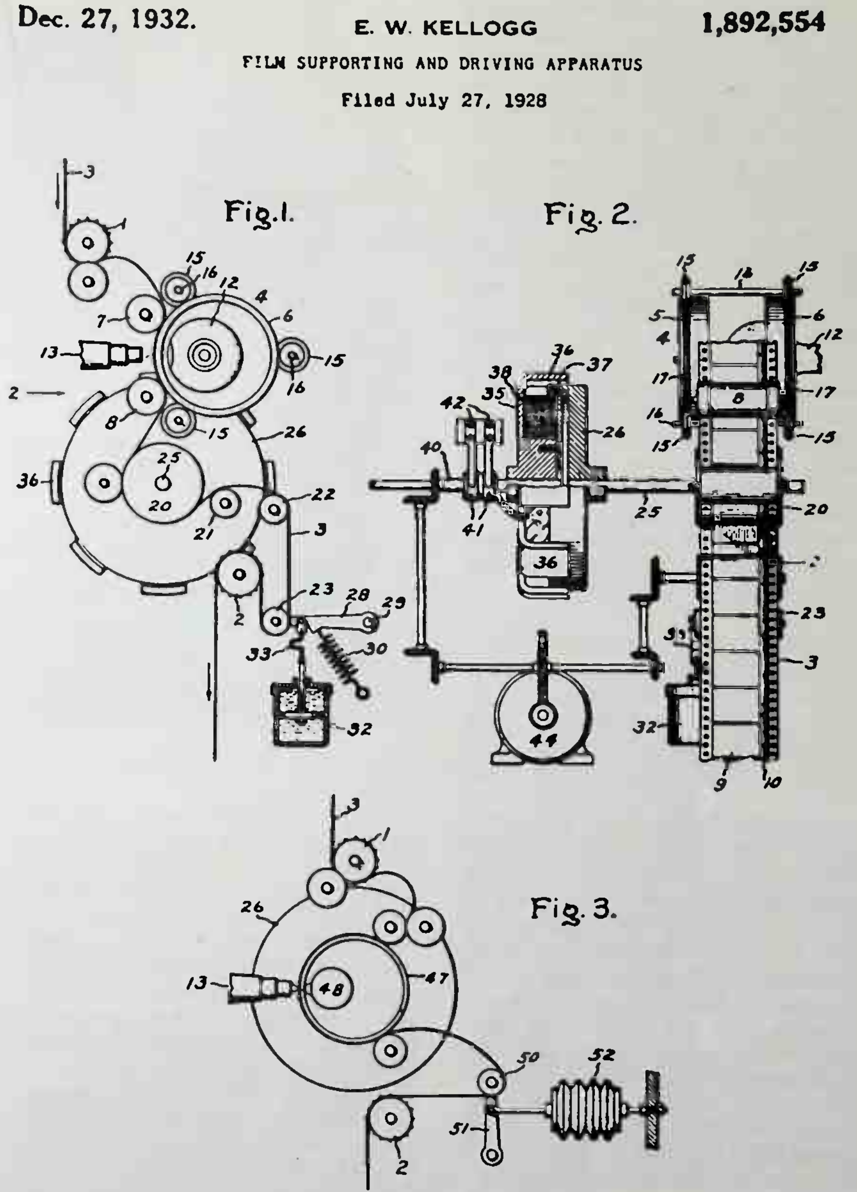
A sprocket can propel a film at uniform speed only when the pitch of the teeth and that of the holes match perfectly.* Otherwise there are continual readjustments of the film on the sprocket, producing in general 96-cycle flutter, plus random small variations. A paper by Herbert Belar and myself[82] shows graphically the startling breaking up of single tones into a multiplicity of side tones by a 96-cycle speed change such as might result from a shrinkage of about ½%.

Recorders, since they are working with fresh film, may give very little 96-cycle flutter at the sprocket. The Western Electric recorders of the earlier 1930's were designed on this basis.[4,93] The large sprocket was of precise construction and a nearly perfect fit for unshrunk film. It was on the shaft with a flywheel, and driven through damped springs. Another sprocket (unfiltered) drew the film from the magazine and resisted the pulls from the take-up magazine.

The engineers who designed the recorders supplied by RCA took no chances with sprocket teeth. In the first General Electric recording machines the film was carried past the recording light on a smooth drum (with a flywheel on its shaft) and a soft-tired pressure-roller prevented slipping.[67] Between the drum and the sprocket which fed the film through the machine at synchronous speed were flexible loops of film which (so long as they remained under sufficiently low tension to retain their flexibility) would not transmit appreciable disturbances from the sprocket to the drum. Because of uncertain shrinkage the drum must be free to choose its own speed. The simplest expedient was to let the film pull the drum, like a belt. Machines built this way worked so well at times that they delayed the effort to design something on sounder principles. My own part in the development of a better machine lay originally in the recognition that the stretch of film which pulled the drum, in combination with the inertia of the flywheel, constituted an oscillatory system, although its period varied so greatly that the irregular action did not look like that of any oscillator we were accustomed to seeing. Another trouble was that the film loop was not free enough for isolating the drum. The cure for the bad effects of oscillatory action would be to provide damping. One way to provide this would be by bending the film around a flexibly supported idler roller,[105] connected to a dashpot. Another measure would be to use eddy-current damping at the flywheel by mounting a copper flange on the flywheel, spanned by a set of magnets. To use stationary magnets would provide damping but would also produce a steady drag, making a really flexible film loop impossible.[102,102a] By mounting the magnets so that they could be driven somewhat above flywheel speed, it became possible to provide a forward torque as well as damping, thereby relieving the film of all but a small part of its tension. (Figs. 4B, 5).[94–96]

The first magnetic-drive machine (an experimental model) (Fig. 6) employed both the damped idler roller and the rotating magnetic damper, but the latter was so effective that the first was superfluous. By adjusting the magnet current the film loop could be caused to run anywhere between a very slight deflection and a nearly semicircular bend. A production model (the R-4) recorder was designed in 1929 and was in production in 1930.[94] It was followed by other models (PR23 in 1933[98] and PR-31 in 1947[99]) employing the same principle.

* Sprocket propulsion of the film through the light beam has certain advantages for printers, as will be explained in the section on printer improvements. This mechanical section, however, seems the logical place for a brief review of studies by J. S. Chandler and J. G. Streiffert of the Eastman Co., directed to the reduction of sprocket-tooth flutter.



Film 5. Schematic representation of the magnetic drive for film motion, showing also provision for damping by use of a movable roller with dashpot.

The magnetic drive probably carried the idea of isolation of the film drum from disturbing forces farther than it has been carried in any other film-recording machine. Its extreme effectiveness as a filter system was demonstrated by Russell O. Drew and myself at the SMPE's 1940 spring convention.[96]

Although only a few were built, I should mention another recorder, the R-3,[70] designed by C. L. Heisler of the General Electric Co., which preceded the magnetic type. This had the smooth drum with flywheel to carry the film past the recording point, and the sprocket drive to hold synchronism. The drum was driven through a continuously adjustable-speed friction drive, which might be compared to a cone pulley, and the speed adjustment was automatically controlled by the length of the loop of film between the sprocket and drum, which loop was measured by the position of a movable deflecting roller.

Effect of the Tri-Ergon Patents.[35,114] Mention has been made of the development, beginning in 1918, of a sound system by Vogt, Massole and Engl, to which the name Tri-Ergon was given. They obtained very broad patents in Germany and were allowed some extremely broad claims in the United States. The patent which figured most seriously in litigation was No. 1,713,726 in which one claim covered the use of a flywheel on the shaft of the roller which carried film past the translation (recording or reproducing) point. Another claim covered carrying the film on a short roller and scanning it at the overhanging edge, and a third (based on a showing of flexibly mounted rollers pressing against and deflecting the stretches of film on either side of the drum) called for a spring-pressed roller engaging the film between the sprocket and the roller (drum). Patent attorneys in the RCA group and Western Electric felt very confident that the broad flywheel claims could be safely disregarded because anticipated in many old sound-recording and reproducing devices, but the patent departments would not approve constructions using the overhung film for scanning until after about 1930, when W. L. Douden of the RCA patent department discovered an older disclosure of the same idea in a patent application of C. A. Cawley* (to which RCA obtained rights).

Film-Transport System of Soundheads. So the first reproducing machines to be marketed avoided the overhanging film

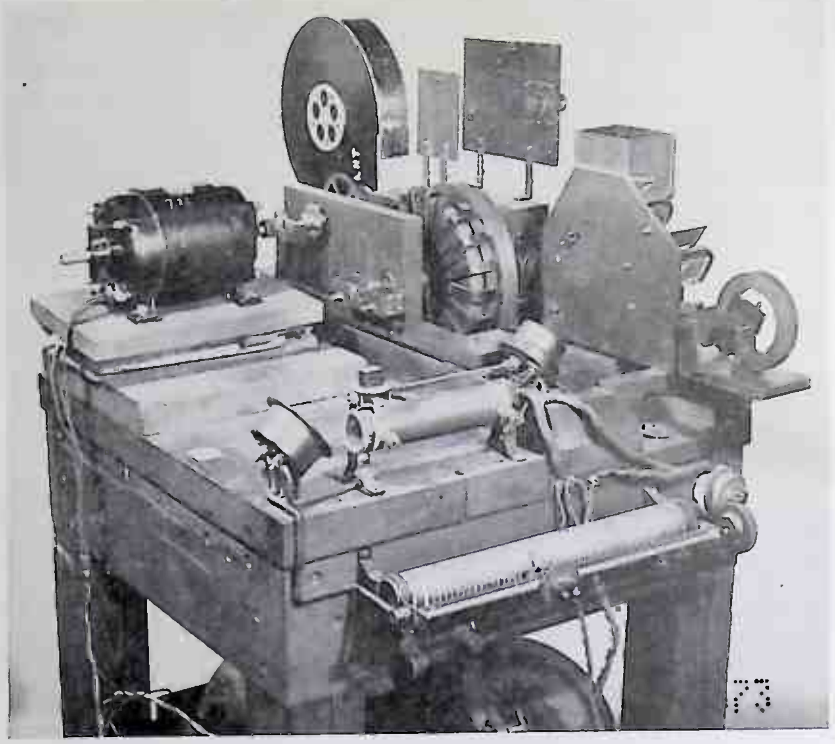


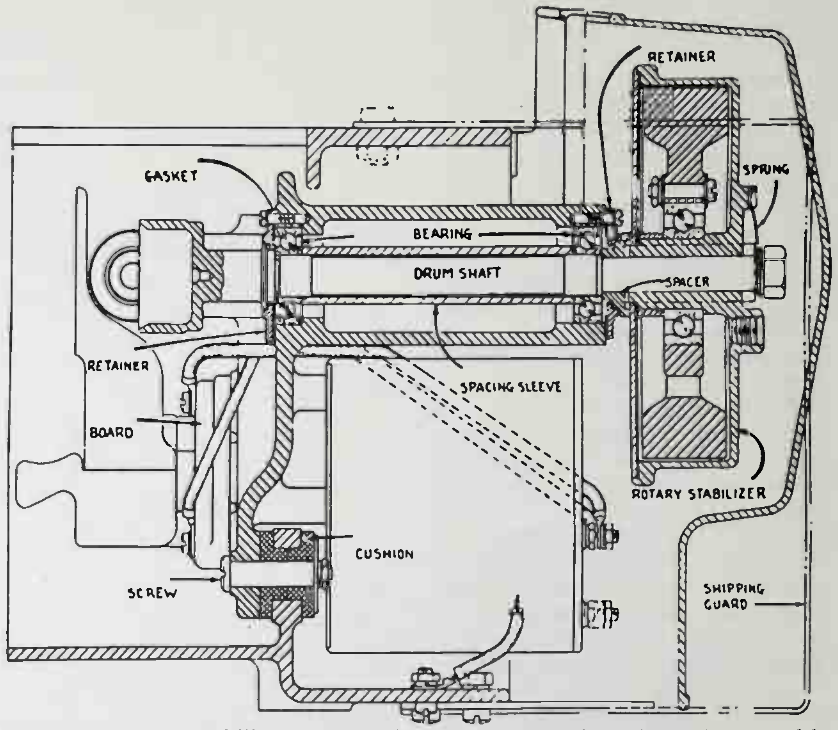
Fig. 6. Original model of magnetic-drive recorder.

feature, and instead pulled the film through a sound gate, where the scanning light passed through it and into the photocell. Friction in the gate made this arrangement much less favorable to constant speed than the use of the overhung principle. For constancy of film speed no further measures were used than to try to provide good sprockets to pull the film through the gate, and to filter the motion of the sprocket by use of a flywheel, and driving through springs. To damp this filter, the RCA PS-1 used grease-pads acting on the flywheel (Fig. 4A) and the Western Electric used a balanced pair of oil-filled sylphon bellows which acted as a dashpot supplementing the driving springs.[53] A practical improvement over filtering the sound sprocket was to drive a heavy flywheel on the sound-sprocket shaft by multipleV belts directly from the motor, and then by gearing take from this shaft whatever power is needed to drive the projector. The heavy flywheel and tight coupling to the motor gave the sound-sprocket drive such high mechanical impedance that its speed constancy was not materially disturbed by the irregularities of the projector load.

The Rotary Stabilizer. The discovery of the Cawley patent application by Douden made the RCA Patent Department consider it safe to build machines in which the reproducing light passed through the film where its edge overhung a short roller. With this privilege the way was open to make the film motion in reproducing machines comparable with that which had been attained in the magnet-drive recorder. However a less expensive construction was very desirable. The damping in the recorder was by eddy-current coupling between the flywheel and a coaxial magnet running at nearly the same speed. The functional equivalence of eddy-current coupling and viscous-fluid coupling was well recognized. I had tried some experiments with viscous coupling to a coaxial member which was not independently driven but was free to pick up the flywheel speed. The inertia of the viscously coupled member would tend to keep its speed constant so that a change in flywheel speed would cause relative movement and hence energy loss.[100] But I gave up in view of the feebleness of the damping I obtained.

It remained for C. R. Hanna of Westinghouse to make an analysis of the system. He showed that in order to get critical damping of the mass which is rigidly connected to the drum, the viscously coupled mass must have eight times as much moment of inertia, and the coupling coefficient must have the right value.[69] In 1932 and 1933 E. W. Reynolds and F. J. Loomis of the RCA Victor Co. in Camden did the job right.[101] The directly connected mass was an oil-tight shell of aluminum alloy inside which was a heavy cast-iron flywheel supported on a ball bearing whose friction was negligibly small in comparison with the oil coupling. Small clearance between concentric surfaces and a suitable oil gave the desired coupling. The inertia ratio was less than 8:1, but damping somewhat short of critical is satisfactory. By use of high-grade ball bearings the drum with attached stabilizer was caused to run with so little tension on the film which pulled it that the loop had plenty of flexibility for effective filtering.[102, 102a] (See Figs. 4C and 7.) The rotary stabilizer introduced in 1933 proved so satisfactory that it has been retained with little change for twenty years. A device on similar principles, called the "kinetic scanner" was used in Western Electric soundheads early in 1936[103] (Type 209).

In 1941 Alberscheim and MacKenzie[81] and Wente and Müller[104] described damped flywheels in which the entire viscously coupled mass was liquid. In order that there might be sufficient viscous resistance to movement of the liquid with respect to the container, partial obstructions were placed in the annular channel. This type of damped flywheel was used in the recorders and reproducers of the stereophonic system developed and demonstrated by the Bell Telephone Laboratories. Study has been given to the problem of finding suitable fluids. A low-temperature coefficient of viscosity is desirable, and if the entire coupled mass is liquid, high density is valuable.



Rotary stabilizer construction of F. J. Loomis and E. W. Reynolds.

Fig. 7. Cross section, showing construction of the "rotary stabilizer."

Filters Using Movable Idler Rollers. The use of this type of filter was avoided in this country because of the danger of infringement suits on the basis of either the Tri-Ergon patent (No. 1,713,726)[35] or the Poulsen and Peterson patent (No. 1,597,819). Both of these show rollers elastically pressed against the film to deflect it from a straight path and thereby provide flexibility. Neither patent shows or mentions provision for damping, and yet the great merit of such an arrangement is in the simplicity with which damping can be obtained and not in the extra flexibility, for plenty of flexibility can be had by simply freeing the film of too much tension.[102,102a] The flywheel may be solid and the arm on which the film-deflecting roller is mounted can be connected to a dashpot. (Even a cruder frictional device may give good results, but resistance of the viscous type is better.) A laboratory model of a soundhead using this type of filter was built about 1928 by the writer and performed very well, but did not receive patent approval (Fig. 4D).

After the patent obstacle to the use of the sprung-idler type of filter was ended, soundheads employing this principle were brought out by the Century Projector Corp. and the Western Electric Co.[105] and a recorder by Western Electric.[106] RCA adopted this film-motion system for 16mm machines and lightweight recorders R-32 and R-33,[107] but for 35mm soundheads continued to use the rotary stabilizer, the advantage of the movable-idler design being not so much a matter of performance as of lower manufacturing costs, an item which is contingent on schedules and tooling costs. Recently the flexibly mounted idler filter system has been utilized by RCA[108] and others[109,110] in soundheads for use with multiple magnetic soundtracks.

Filter System With Drum and Sprung Sprocket. A film-motion system developed by engineers of M-G-M is described by Wesley C. Miller.[111] Recording or reproduction takes place on a drum with solid flywheel, and the drum is driven from a sprung sprocket isolated from other sprockets by loose loops. The film passes from the sprung sprocket, around the drum and back to engage the opposite side of the sprung sprocket, and this portion of the film is maintained under tension by a roller pressing against a free span of the film. The tight film affords the required traction between film and drum. Adjustable friction pads between the sprocket and its shaft cause frictional resistance whenever the deflection of the sprocket driving springs changes, thereby damping the system.

Excellent film motion was obtained in these machines (Fig. 4E).

Minimizing Sprocket-Tooth Flutter. J. S. Chandler, in 1941[112] showed it to be possible to so shape sprocket teeth that the film speed would fluctuate between a maximum and minimum value which are the same at perfect fit and spread progressively with increasing misfit, but with a net flutter which can for a moderate range of shrinkage be quite small. However the realization of the calculated flutter values demands perfect perforation uniformity and freedom from any sticking on the teeth as the film is fed on or stripped off.

A further development in improving sprocket action is described by J. G. Streiffert.[113] The driving faces of the teeth are radial and the film is supported on a cylindrical surface which is slightly eccentric with respect to the sprocket. The film is fed on at a point where the teeth project only slightly above the film support, and as it travels around its arc of engagement the film gets closer to the roots of the teeth. The radius from the sprocket center to the film thus keeps decreasing, and therefore the velocity of the tooth face at the plane of the film decreases. The effect is essentially as though the tooth speed and the tooth pitch decreased correspondingly. The design is such that the tooth enters the hole with a margin of clearance and with the effective velocity (since the working radius is here near maximum) slightly greater than that of the film. The tooth face therefore gains with respect to the film, closing up the clearance, and as soon as it touches the edge of the perforation begins propelling the film. While it is doing so the next tooth is catching up. Each tooth in turn propels the film from the moment that it reaches the perforation edge until the next tooth, which at the instant is moving slightly faster, touches the edge of its

* The Cawley application had been filed Jan. 28, 1921, but had been held up on technicalities. It was put into suitable shape and issued Sept. 29, 1931 as a parent patent, No. 1,825,438, and three divisional patents, of which No. 1,825,441 contained the claims to the overhang feature.

perforation. Thereafter, the effective speed of this tooth, which continues to decrease, is less than that of the film, so that a gap or clearance develops between the tooth and the leading edge of the perforation. The design is such that the film is not stripped from the teeth until, in all cases, sufficient clearance has accumulated to avoid possible interference during the stripping.

Assuming that the film velocity is equal to that of the driving-tooth face at the radius where it touches the film, the film speed will fluctuate by the amount by which the effective tooth speed decreases as it travels one tooth pitch. This can be a very small change, especially if there is a large number of teeth and the eccentricity no more than needed to take care of a reasonable shrinkage range.

The region on the circumference of the sprocket where the propelling action takes place varies with the shrinkage of the film. Thus with unshrunk film the propulsion will be relatively near the place where the tooth enters the perforation, while with shrunk film it will be where the teeth are projecting farther, so that the point of contact is nearer the root of the teeth.

One way of describing the action of the system is to say that a film of any given shrinkage finds the appropriate radius where the pitch of the teeth equals the pitch of the perforations, and this is the region where propulsion takes place.

The Eastman Co. has used this system with excellent results in experimental printers. The Streiffert paper gives wowgrams of negative recordings made on such a sprocket, and also of contact prints, and for comparison wowgram (or flutter recordings) of prints made in a conventional sprocket-type printer, showing a major reduction in flutter with the new sprocket.

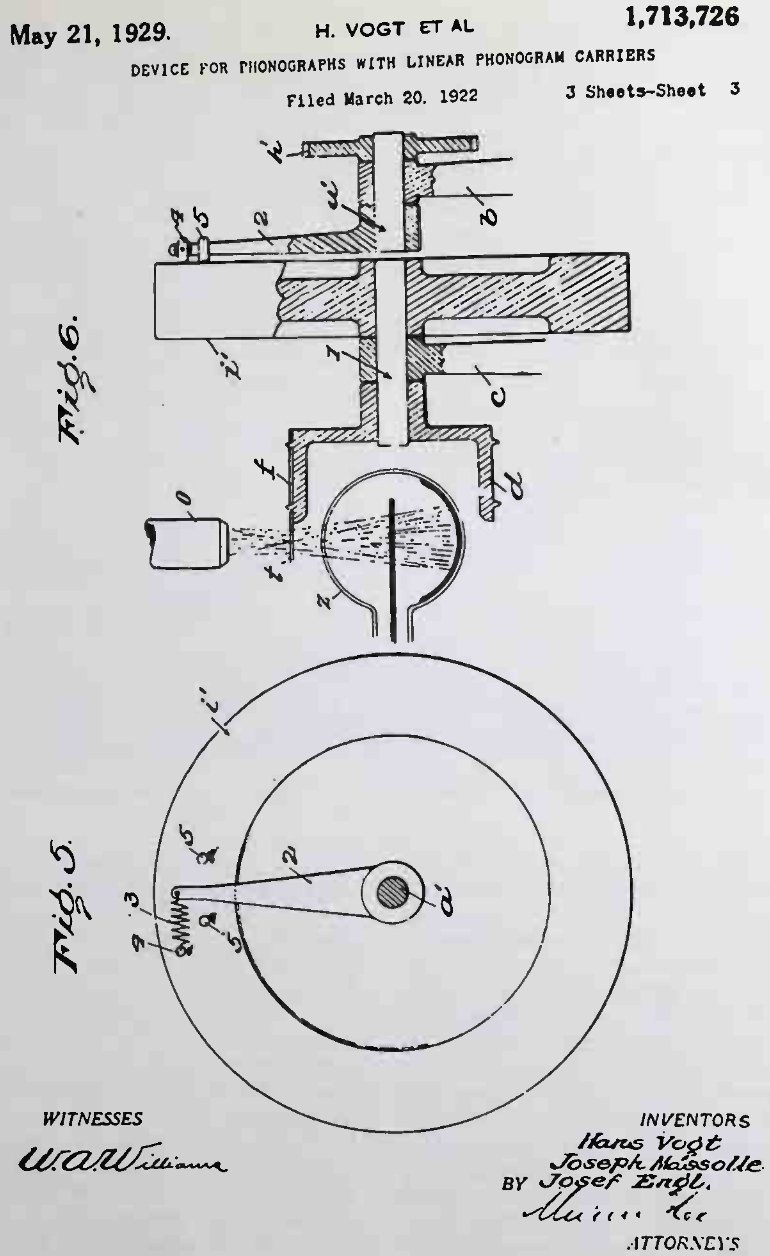


Fig. 8. Tri-Ergon showing (U. S. Pat. 1,713,726) of flexibly mounted rollers deflecting the film between sprockets and drum.

Litigation. Despite the efforts to avoid infringement of such claims of Tri-Ergon patent No. 1,713,726* as appeared to have any likelihood of being held valid, the American Tri-Ergon Corp. brought suit against Altoona Publix Theatres Inc., who were using an RCA Photophone (PS-1) projector attachment or soundhead. The case was tried at Scranton, Pa., in the U.S. District Court for the Middle District. The apparatus had been sold with a guarantee against patent liability, and the suit was defended by RCA, Electrical Research Products Inc. giving technical assistance in the defense. The court ruled (Feb. 10, 1933) that seven of the claims were valid and infringed. The case was appealed and reviewed by the U.S. Circuit Court of Appeals for the third circuit (in Philadelphia), which affirmed (June 13, 1934) the findings of the lower court. The defendents then appealed to the U.S. Supreme Court, which at first refused to review the case, but finally decided to do so, and on Mar. 4, 1935, ruled that the seven claims in the suit were all invalid (294 US 477).[114]

This removed the threat to the equipment manufacturers of what might have been almost crippling damages, for had the findings of the lower courts been sustained the plaintiffs would have been in a position to bring suits for damages for infringement by most of the recording and reproducing equipment in this country, and covering a period of over five years.

The American Tri-Ergon Corp. applied on Feb. 18, 1937, for a reissue patent with modified claims, and this was granted Jan. 11, 1938, as Re. No. 20,621. On Oct. 25, 1946, RCA reached an agreement with American Tri-Ergon Corp. whereby it was granted rights under both the original and reissue patents.

Two other patents placed restrictions on the film-motion systems which American engineers could safely employ, namely Poulsen and Peterson No. 1,597,819 (filed July 9, 1924, and issued Aug. 31, 1926) and Poulsen No. 2,006,719 (filed Germany Sept. 1, 1930, and U. S. Aug. 19, 1931, and issued July 2, 1935). These patents to Danish inventors were owned by British Acoustic Films Ltd., which brought infringement suits against RCA Mfg. Co. and against Electrical Research Products Inc. The trial (in Wilmington, Del.) was before the U.S. District Court for the District of Delaware (43 USP–Q69). The arrangement shown in the patent comprised a drum propelled by the film, the film being passed around a flexibly mounted idler roller. Some of the claims in suit described the invention as "means contacting the film for increasing its flexibility." The apparatus in suit was the RCA PS-24 (rotary stabilizer type), which has no flexibly mounted roller, but was alleged to have the equivalent in that the film loop was so formed by the fixed rollers as to be very flexible. The court ruled Sept. 22, 1939, that the claims in suit were not infringed and not valid (the flexibly mounted idler having been disclosed in the earlier Tri-Ergon patent).

Plaintiffs appealed and the case was reviewed by the Circuit Court of Appeals of the Third Circuit which affirmed the findings of the lower court (46USP–Q107, June 27, 1940).

To forestall possible future trouble RCA obtained rights under these patents by agreement with British Acoustic Films Ltd., Dec. 21, 1944.

* This patent, issued May 21, 1929, was filed in the U.S. Mar. 20, 1922, and had a German filing date of Mar. 24, 1921.

Immediate Requirements for Sound

Our historical story thus far has been confined almost entirely to the three fundamental elements, sound pickup (or microphone), a recording and reproducing system and loudspeakers. These represented the difficult phases of the problem, but before sound could become commercial certain items of equipment had to be made available and techniques established. Before discussing the advances in the art that followed commercialization, we shall mention some of these items.

Standard Track Position and Width.[115] Agreement between the makers of variable-width and variable-density systems was reached in 1928. The reproducing light spot must cover more than the extreme width of the clear area of a variable-area track, with both ends on black areas, but should fall entirely within the width of a variable-density track. This requirement is met with margins of safety, by recording density tracks 0.100 in. wide, while the scanning spot is 0.084 in. long. The modulated area of a variable-width track is limited to 0.071 in. with the black parts extending to the 0.100-in. width. The track center line is to be 0.243 in. ± 0.002 in. from the edge of the film.

Printers. Continuous contact printers previously used for pictures only could be adapted to sound by providing masks by which light could be confined to either the picture or the soundtrack area. Except for certain newsreel negatives, the sound and picture were on separate negatives, so that the print film had to be run through the printer twice. Even when the sound was on the same negative as the picture, the offset was not usually the required 14.5-in. and independent light controls were needed. Combination printers were soon developed which were essentially two printers in cascade, so that the print was complete with one passage through the machine.[116–120]

Bloops. The development engineer can overlook many defects so long as he knows their cause and that the apparatus he is testing is not at fault, but before sound pictures could be shown the public, these faults had to be corrected. The noise which a splice makes as it passes through the scanning beam can be made almost inaudible by cutting off the light gradually instead of suddenly. This was accomplished at first by painting a black spot with sloping edges over the splice. Later, black patches which could be quickly cemented in place where the splice crosses the sound track were made available. These are called "bloops." They are of trapezoid shape, masking off the entire sound track for a distance sufficient safely to cover the splice and with end slopes designed to change the light gradually enough to keep the noise just below noticeability at normal gain settings. The design of bloops is discussed in several papers.[121–123] To prevent a disturbance due to a printed-through negative splice, Sponable[124] described a punch which made a hole in the negative, resulting in a suitably-shaped black spot on the print.

Electrical blooping of splices in negatives has come into extensive use. When a negative splice goes through the printer an auxiliary light exposes (through the base) a suitable area of the print film, an edge notch or other means being employed to control the blooping light.

Lewin (Apr. 1947)[124a] describes a system of silencing splices in re-recording positives in which the output is momentarily suppressed in response to a punched hole.

Blimps.[125] Cameras which were entirely satisfactory for silent pictures were much too noisy for making sound pictures. Much quieter cameras were developed eventually,[126–128] but for immediate requirements it was necessary to reduce the noise radiated by existing cameras by building shells around them with thick layers of sound-absorbing material. These were called "blimps," or sometimes "bungalows." To smother the sound and still give access for the necessary operations was enough to tax the skill and ingenuity of the designer. Even with the quieter cameras it is still common to resort to partial or complete sound-insulating housing.

Sound Stages.[129,130] The requirement of freedom from noise necessitated the building of sound stages in which extreme measures were taken to exclude noise of outside origin. Many of these had double concrete walls and double floors, with sound absorbing material between, the inner walls and floor being supported on cushion mounts to prevent transmission of earth tremors. The roof and ceiling structures were designed on the same principle.

The high absorption (or short reverberation time) desirable for recording purposes helped control noises originating inside, but so far as possible all sources of noise were eliminated. Noisy arc lights gave way to incandescent or other quiet lamps, and all mechanisms were made to operate as noiselessly as possible. Ventilating systems required extreme measures.

In recording dialogue, the better the suppression of general room reverberation, the farther (within limits) from the action can the microphone be placed, thus affording more uniform coverage

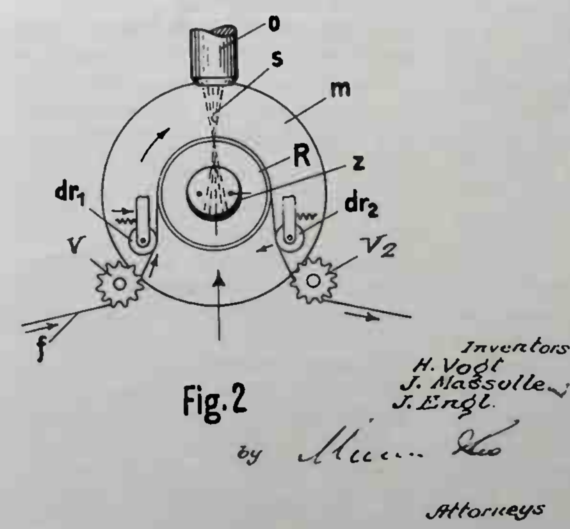


Fig. 9. Tri-Ergon showing of filtered sound-sprocket and overhanging soundtrack.

and making it easier to keep the microphone out of the field of the camera. If some echoes are wanted the "set" can frequently be designed to produce enough. Artificial reverberation using echo chambers in the recording channel or equivalent devices has many applications.

In contrast to the requirements for speech, the recording of music calls in general for rooms with considerable reverberation.[131–133]

Theater Acoustical Treatment.[134,135] The acoustical treatment of auditoriums has probably received more study than any other phase of architectural acoustics, perhaps because the desired characteristics are most difficult to attain. The reverberation must be sufficient to make music pleasing and to help equalize sound intensity in the various parts of the space, but must be short enough not appreciably to impair clarity of speech. A high order of directivity in the loudspeakers plus application of absorbent materials to any large surfaces toward which they are directed has helped with this part of the problem.

In general every theater or auditorium, many of which were built before the era of sound pictures, presents its own problems and calls for individual study. For new theaters there is optimum shape to consider as well as best distribution of absorption. The multiple loudspeaker systems (discussed later) besides making new effects possible have given the acoustical designer somewhat more freedom.

Booms and Dollies. In order that microphones might be suspended as near the action as might be wanted but just above the field of the camera, and in order that their positions might be readily changed, microphone booms of various types came quickly into use. The more elaborate of these were much like derricks on platforms, with rubber-tired wheels on which they could be moved quickly and almost noiselessly.

Similarly, rubber-tired, battery-operated camera dollies enabled the cameraman rapidly and quietly to change the position or height of his camera.

Equipment of the kind just described underwent improvements through the years, but the main features were available from the start of commercial sound pictures.

Monitoring and Level Control.[136–139] Another line of equipment the essentials of which were made available as soon as recording machines, and which has been improved from time to time, was that providing for monitoring and level controls and (especially in re-recording operations) for adjusting the relative levels from several sources, or "mixing." Volume indicators[140,141] of several types were in use in broadcasting stations, and the design of mixing controls was well established.

The man responsible for recording judged the quality by means of a monitoring loudspeaker. He could check quality as represented by the current supplied to the light modulator, or by means of a photocell, in terms of the light reaching the film.[1,64,254] In the case of the RCA Photophone system there was a card on which the modulator projected a light-spot, the movements of which showed the amplitude being recorded on the sound track.[142]

It was for a time held by some that the monitoring speaker should be of the same type as a theater speaker, but high-quality monitoring speakers of the direct-radiator type were soon made available, and these were much better suited to the small rooms where the controls were located. In terms of frequency range covered, the cabinet-type monitoring speakers kept pace with the improvements in theater speakers (see section on loudspeakers). High-quality headphones have also found wide use in monitoring.[143,211] Whatever type of listening device is used, it should obviously be designed to give the recordist about the same range and tonal balance that a theater patron would get.

Screens. Our sense of the direction from which sounds come is too keen for us to be fooled by loudspeakers placed alongside or above the screen. Sound must come from directly behind the screen to give a good illusion. This is one of the lessons that was learned early. Screens of the types developed for silent pictures caused excessive loss and distortion if placed between the loudspeaker and the audience.

Mention has been made of a sound-transmitting screen developed by E. I. Sponable in 1927.[6] One of the first papers in the SMPE *Journal* dealing with screens for sound pictures was that in 1930 by H. F. Hopkins.[144] His curves of measured transmission indicate good results with screens having perforations whose total area is 4% or 5% of the screen area, and show definite advantage in a thin (0.013-in.) screen rather than a thicker (0.030-in.) material. With such screens the loss of brightness need be no greater than the proportion of the area taken out by the holes. Allotment of about 8% of the area to holes has been common, for example about 40 holes of 0.050-in. diameter per square inch.[145]

Processing, Variable-Density.[146] In the story of the work at Western Electric and Bell Laboratories I said that it was recognized by Wente and by MacKenzie that for the correct, or linear, relation between negative exposure and print transmission, the product of the negative and print gamma* should equal unity.† This is in accordance with principles set forth in early SMPE papers by L. A. Jones[62] and by A. C. Hardy.[61] Since practice in making pictures had been to develop the print to a gamma of approximately two, and both sound and picture would receive identical development, the sound negative should be developed to a gamma of about 0.5 or slightly higher. Picture-positive film was used for a number of years for sound negatives. Developers of the types used for picture negatives tend to give low contrast and fine grain, and the use of such developers helped to give the desired low value of gamma for the sound negatives.[147,148] MacKenzie[64] gives some information about the harmonic distortion which results from departures from the unity product, and thus gives an indication of tolerances with respect to development.

With the advent of sound, with its requirement for more strict control of development, control by use of sensitometric test strips, and by specified time, temperature and developer formulas[149–152] supplanted dependence on visual judgments of operators, where that practice had prevailed.[153,154] Maintenance of developer activity received much attention,[155–158] and stop baths assumed increased importance.[159,160] Rack-and-tank methods, where these had been followed, gave way to continuous machine processing.[161–163]

How generally the distinctions between specular and diffuse density,[164] and between exposure modulation by varying time (light valve) and varying intensity (as by glow lamp)[165] were understood at first is a question, but these points were well covered in the literature. The Eastman Capstaff Densitometer,[166] which was developed primarily for measuring picture negatives for contact printing, reads diffuse densities. This would be appropriate for measuring the densities of sound negatives for use in contact printers, but not for densities of sound-track prints, for it is the specularly transmitted light which reaches the photocell in a reproducer.

The widely used Eastman IIb Sensitometer, brought out about 1932,[167] which gives an accurately standardized series of test exposures in the form of a step tablet with exposure time increasing in the ratio $\sqrt{2}$ per step, and ranging from about 0.004 sec to 4 sec, has been of utmost value in maintaining controls. However, it does not simulate sound-track recording conditions, where the intensity is extremely high and the time for average exposure was approximately 1/18000 sec (1/36000 sec with a later light-valve system and in present practice about 1/90000 sec) and still shorter for low exposures. The 1934 paper by Jones and Webb[165] gives an indication of the magnitude of the error. The Eastman Sensitometer on the other hand gives exposures which approximate sufficiently well those which a print receives, and are thus suitable for determining gamma of contact prints. For many purposes it has been satisfactory to draw conclusions by applying correction factors, if needed, to the readings of these instruments.

In the course of a few years densitometers employing photocells were developed which had the advantages of greater accuracy and much faster operation than the Capstaff visual-balance type.[168–172] For exposing sound negatives for sensitometry purposes, the light valve itself, with suitable calibration, can be used. The subject is again discussed under "Intermodulation Test."

While the conditions for low distortion were to keep both negative and positive exposures on the straight parts of the H & D characteristics, studies reported in 1931 by D. MacKenzie[172a] showed that low distortion was still possible while using the "toe" range of both films ("toe recording") or that of the positive only ("composite"). Toe recording using positive stock for the sound negative might, if the recording-system light was limited, be preferable to resorting to faster and coarser-grained recording stock. In the case of single-film systems (sound recorded on the picture negative) where the development of both the negative and positive soundtracks is fixed by picture requirements, MacKenzie found that the composite system offered best promise of low distortion. Both the toe and composite systems give higher output than a classical or straight-line system, but poorer signal-to-noise ratios.

It took a number of years to bring about the full transformation from the methods (depending much on visual judgments) which had been employed for making silent pictures, to the close controls and scientific precision needed for satisfactory and consistent sound. The constant and close checking of every element exerted a pressure for improvement along the whole front, including the manufacture of the film, in which departures from uniformity were quickly detected. The story is interestingly told by J. I. Crabtree.[146] An early account is given by J. W. Coffman.[153]

Processing, Variable-Area. Since the ideal variable-area track is part clear and part black with a sharp boundary between, there is no question of preserving correct shades of gray, but in general the higher the contrast (or gamma product) the better. As in the case of variable-density tracks it must be assumed that the print development will be that which is wanted for the picture, and that has been taken in general to give a gamma of about 2.0. Variable-area negatives as well as the prints are processed in high-contrast developers. The variable-area system is noncritical with respect to gamma product but, for a given positive emulsion and processing, there is for any given negative a best setting of printer light.

A comprehensive study of available sound-recording films and their processing was published by Jones and Sandvik.[173] Another study was made by J. A. Maurer.[174] From his curves it appeared that negative densities of 1.3 or higher were desirable, and the prints which gave maximum outputs were the ones having densities (in the dark areas) about equal to those of the negatives from which they were made. This held true for negative densities ranging from 0.6 to 1.3 and higher. The maxima however were very broad.

In November 1931, Dimmick[175] reported the results of a series of determinations of conditions for maximum output from a 6000-cycle recording, using Eastman positive 1301 for negatives and prints, and 4, 6, 8 and 10 min in D-16 developer. The study covered an adequate range of the four variables — negative (recording) exposure, negative development or gamma, printing exposure and print development. The results showed that wide ranges in each of the variables could be used with comparatively small loss of output, but for any negative there was a print density at which output was greatest. It made comparatively small difference (except near the extremes) whether a given density of either negative or print was reached with small exposure and longer development or more exposure and less development, but in general the maxima were broader with the higher values of gamma, especially that of the print. The two highest gamma values in the series, 2 and 2.18 of both negative and print, in general gave best results, with negative densities (measured in the black areas) in the range 1.5 to 2, and print densities a little less in each case than that of the negative.

While maximum high-frequency output is of less consequence than avoidance of cross-modulation (which is discussed in the section on distortion) it is of interest that recommended practices based on the test just described come very close to those found to be best in later experience and after current testing methods had become established. The cross-modulation test did not come into general use until 1938.[176]

For a number of years a print density of 1.4 or slightly higher, with appropriate corresponding negative density, was taken as a practical objective. As galvanometers and optical systems were improved and finer grain films came into use, the tendency was toward higher densities for both negatives and prints, especially for the negatives.

Evolution in a Growing Industry

Greatly Expanded Developmental Activities. The development work prior to commercialization of sound was carried on largely in laboratories supported by manufacturers of supplies or equipment, or in independent laboratories, and it was done on the basis of hope for returns which might be realized either through patent royalties or through sales of equipment or both.

Once sound pictures began to be made and shown, developmental work was on a different basis. Research and investigations of numerous incompletely solved problems took on rather the character of plowing in profits, with greatly increased total expenditures for research and participation by all the major picture-producing organizations.

Of all of the problems, the most fundamental and greatest in magnitude was learning how to use sound pictures, or the evolution of a new art. This is discussed by J. E. Abbott.[177] The expression "growing pains" aptly describes the less successful phase of this evolution. Capacity for readjustment is one of the qualities of greatness in individuals and in organizations, and the motion-picture industry came through splendidly.

When any industry becomes large, and especially if its requirements are as diverse as those of sound pictures, it provides a market for numerous specialties and services. Many of these requirements are met by comparatively small organizations and others by branches of companies having many other activities and products. A few such items will serve to illustrate: special lamps, arc carbons, screens, cameras, acoustic treatment materials and service, chemicals, printers, testing equipment and studio apparatus. Many important improvements and contributions to technical advances are due to those who develop and supply such auxiliary equipment.

Mingled with the natural rivalry between picture producers has been a

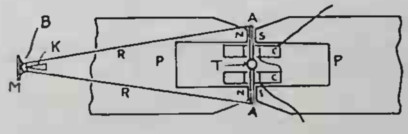


Fig. 10. Arrangement employed by G. L. Dimmick in 1929 galvanometer, for multiplying the rotation of the mirror.

* Gamma is the slope of the straight portion of a curve plotted with density $\left(\text{or log of } \frac{1}{\text{transmission}}\right)$ as ordinates and log exposure as abscissae. This is known as the Hürter and Driffield, or H & D curve. Gamma product is a measure of overall contrast as compared with that in the original exposure.

† In practice, because of some loss of contrast due to stray light in optical systems, best results with pictures had been found with somewhat higher gamma product.

spirit of cooperation and sharing of experience and knowledge which has greatly accelerated progress. In 1930 this society began issuing the monthly *Journal* instead of the quarterly *Transactions*, an appropriate step to accommodate the rapidly expanding literature of sound-picture technology, covering almost every phase of the making and showing of motion pictures. The Academy of Motion Picture Arts and Sciences also played an important part in promoting interchange of information. Engineers and technicians from the sound-picture laboratories have reported experiences with various problems related to processing and controls and to sensitometry, while film and photographic suppliers have spared no efforts to enable those using their product to get the best possible results.

So many have been the contributions to the art and science along these lines, that I find it quite beyond my ability to do more than pay this general tribute and to mention a very few developments which have seemed to me to be of outstanding importance. I trust that I may be forgiven for showing partiality to the types of development with which I am most familiar, and also if I unjustly fail to mention many important advances.

Some of the Improvements After 1930

Galvanometers for Variable-Area Recording. The galvanometers used in the first variable-area recorders supplied by RCA Photophone were practically standard oscillograph vibrators, as these had been built at General Electric. They were oil-immersed and responded well up to 5000 cycles or above. An improved smaller model was brought out in 1930,[178] completely sealed instead of having an open oil well and with no external adjustments. This used molybdenum ribbon (much stronger than the bronze) and was tuned to about 6000 cycles.

When recording was started at the RKO studios in Hollywood, one of the men from the General Engineering Laboratory who had had much experience with oscillographs, F. B. Card, joined the RKO staff. The RKO engineers soon decided that their sound would be better if the frequency range were extended. The ribbons of the oscillographs had been of phosphor-bronze. A small supply of duralumin ribbon was obtained, and with this Card succeeded in re-stringing the RKO galvanometers, with sufficient tension to tune them to nearly 9000 cycles. A thinner damping fluid was then appropriate, a change almost necessary to realize the benefits of the higher natural frequency.

G. L. Dimmick came to the General Electric Co. in 1929, and one of his first projects was the development of a new galvanometer which was promptly used in newsreel equipment.[178] He used a magnetic driving system of the balanced rocking armature type and by an ingenious mechanical arrangement, shown in Fig. 10, made his mirror rotate through about ten times the angle of the armature. The important advantage of this galvanometer was that the mirror was about ten times the area of that of the previous galvanometers. A few years later Dimmick designed a new galvanometer on the same principle (Fig. 11) but improved in numerous details.[142,179,180] This became the RCA Photophone standard for all photographic recording. These galvanometers were tuned to about 9000 cycles. Damping was by means of a block of rubber, the action of which was analogous to that of the rubber line of H. C. Harrison,[*,49] but since it

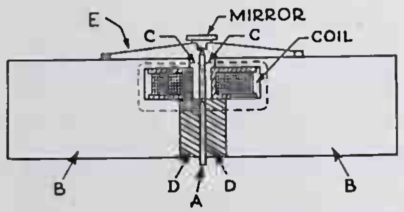


Fig. 11. Cross section of improved recording galvanometer, G. L. Dimmick. A—armature; B,B—pole pieces; C,C—working air gaps; D,D—nonmagnetic spacers; E—tensioned bronze ribbon.

had to work only at high frequency it could be of quite small dimensions. Dimmick found that he could increase the effectiveness of such damping blocks by incorporating tungsten powder in the rubber to increase its density (Fig. 12).

Further improvements in the galvanometer were reported by Dimmick in July 1947.[181] By the substitution of better magnetic materials he was able to

* See June *Journal*, p. 296, third col.

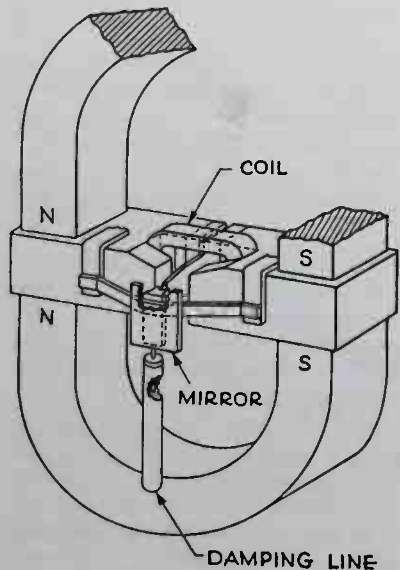


Fig. 12. Construction of RCA recording galvanometer (shown in section in Fig. 11).

reduce hysteresis almost to zero, to increase the sensitivity, and to avoid a slight saturation effect which had been present in the previous design.

A More Efficient and Versatile Optical System. In order that the galvanometer in one of our experimental optical systems might be closer to the slit, and thus send more light through it, I arranged a galvanometer to work on its side, so that it would move the light spot up and down across, instead of parallel to, the slit. I used a light spot with a sloping edge at an acute angle such that the change from zero to full-length slit illumination was accomplished with a movement equal to only one-fourth of the slit length.[*,182] Dimmick improved on this by making the light spot symmetrical with respect to middle of the slit, and having two sloping edges (see Fig. 13, and in Fig. 14 compare C with B). An advantage of the transverse-movement system was that it became very simple, by changing the masks which were imaged on the slit, to produce a variety of tracks which had their special applications.[142,180,183,184]

The combination of larger mirror and reduced distance between galvanometer and slit practically eliminated the diffraction trouble that had, with the small mirrors, impaired the formation of clean, sharp, high-contrast images at the plane of the slit.

Ground-Noise Reduction—GNR.[185] Scratches and dirt on film and graininess of emulsion cause a background noise which is particularly conspicuous when the modulation is low. The noise is reduced by reducing the transmitted light. At the same time, for a given modulation level there is no need for the average transmission to be more than about half the maximum. The noise may thus be reduced when the reduction is most needed, by decreasing the average light when the level of the recorded sound is low. This can be accomplished by biasing the light modulator toward zero, and then using a current derived by rectifying some of the modulation current to increase the mean light transmission when this is needed. The reduced transmission when the modulation is low means a darker track in a variable-density system, or a narrower clear area in a variable-area system, and in either case the noise is reduced. The early patent on this idea was to E. Gerlach.[186] This was assigned to Siemens & Halske, but to the best of my knowledge it

* All galvanometers have practical limits to the angle through which they can swing the light beam. And the required light spot movement sets a minimum to the distance between galvanometer and slit. The light which a galvanometer can send through the slit is proportional to the mirror area and the inverse square of its distance from the slit, up to the point at which the objective lens is "filled."

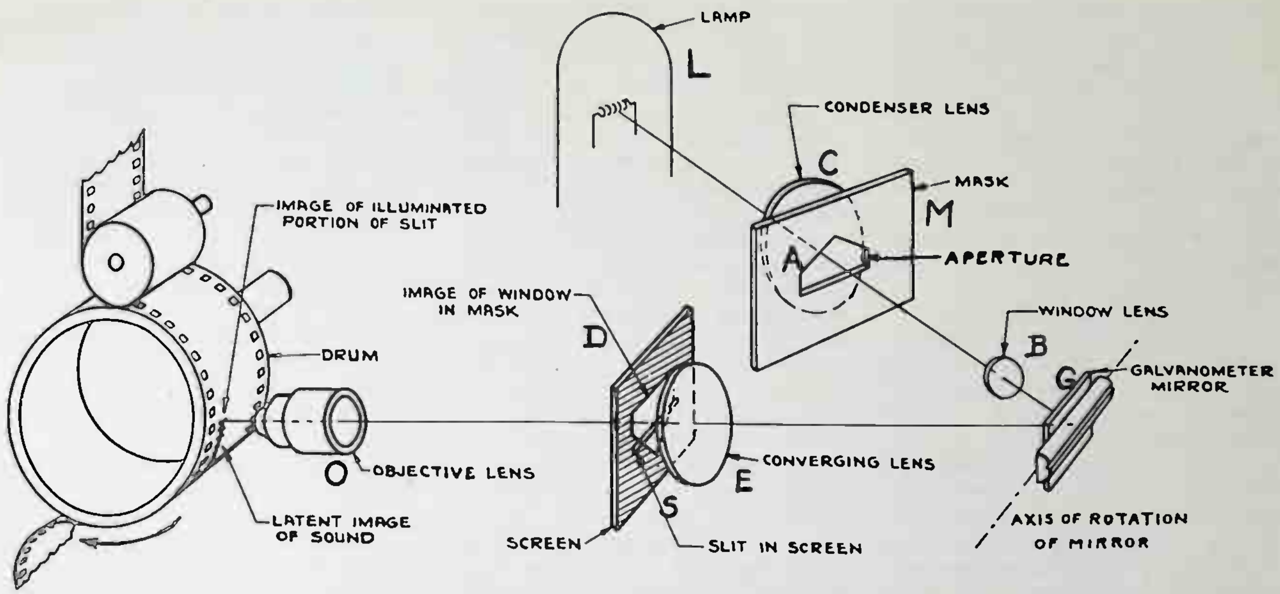


Fig. 13. Variable-area recording optical system. (Lens B images A on slit-plate D.)

occurred independently to L. T. Robinson in Schenectady and C. R. Hanna at Westinghouse,[187] and a system applicable to light valves was developed at Bell Laboratories.[188] The first trials at Schenectady with variable-area tracks did not indicate an impressively large reduction of noise, and there was an objection (in the case of the earlier, unilateral, variable-area tracks) in that the bias threw all of the low-level modulation over to one edge of the track, where many reproducing light beams were of reduced intensity. The project was revived by Hugh McDowell of RKO who got around the objection by screening off the surplus light by means of a shutter instead of biasing the galvanometer.[189] This method and the results obtained were reported to the Academy of Motion Picture Arts and Sciences by Townsend, McDowell and Clark in 1930.[190] Thereafter a commercial form of shutter was designed[191] and ground-noise reduction became standard in the RCA system (Fig. 15).

With the introduction slightly later of the symmetrical track, the objection just mentioned to depending on galvanometer bias instead of a shutter no longer applied, but there was still some danger of saturating the galvanometer. Therefore a double-vane shutter was developed to mask down the light from both sides.[192]

The Bell Laboratories system is described by Silent and Frayne.[188] In the variable-density system there is no objection to accomplishing the result by biasing the valve.

In case of a sudden increase in modulation level the rectified current would change so rapidly as to cause an audible sound. The current for the shutter or for bias is therefore passed through a filter which reduces the rate of change. Limiting the speed with which the average light can increase means some clipping of the first few modulation peaks.

The light valve is a very low impedance device, and since the bias current may have to be sustained for considerable periods of time transformers cannot be used for impedance match coupling to the output tube of the ground-noise reduction amplifier. In the system most widely used with light valves the modulation current is rectified, passed through the timing filter, and used to modulate a 20,000-cycle oscillator, the output of which is amplified and rectified for supply to the valve. The design of the timing filter is much simpler when it can operate at interstage impedance.[193]

As compared with variable-area, the variable-density system is characterized by ground noise which is less in the nature of "ticks" and "pops" and more a continuous hiss. Another difference is that in the density system the noise falls more rapidly with reduced light transmission, so that a given amount of noise reduction is obtained with a smaller change in transmission. The continuous hiss type of noise is especially noticeable if it comes and goes, which changing bias causes it to do. This is

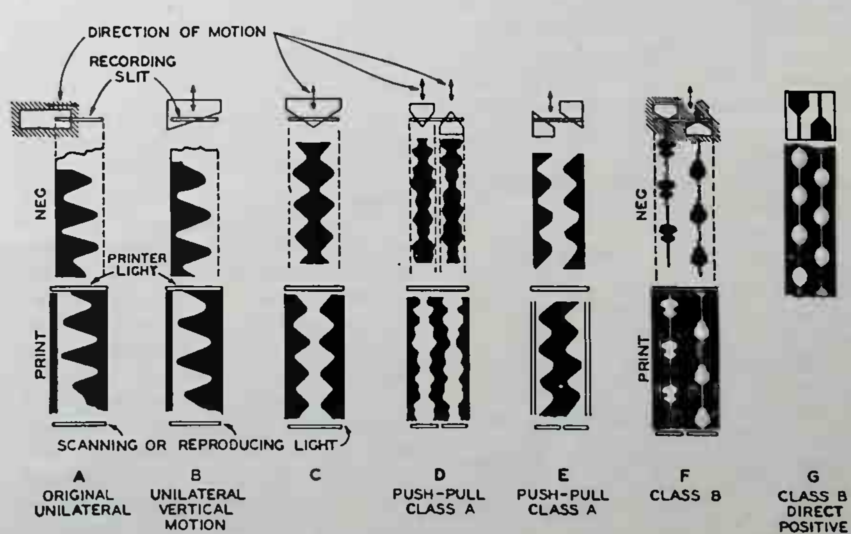


Fig. 14. Various types of area soundtracks, and light-spot shapes which produce them.

Fig. 15. Soundtrack produced by McDowell ground-noise reduction shutter.

often called the "hush-hush" effect, and becomes noticeable if the valve opening does not immediately fall when the modulation drops to a low level. At the same time the fact that smaller changes in transmission suffice to control the ground noise makes it possible to change transmission (or bias) more quickly without causing "thump." The filters for density systems are therefore designed for much faster timing than those of area systems, particularly in closing down when modulation falls.

While clipping of initial modulation peaks is a less serious problem in density than in area recording (because of faster opening and larger margin) engineers working with both systems have given much study to minimizing such clipping. Increased margin will decrease the frequency of occurrence of clipping, but this would be at the price of more noise. In all systems, opening (increasing light) is about as fast as it can be made without becoming audible, while the closing is much slower so that very brief reductions in level will not produce incessant closing and opening.[194]

A nonlinear characteristic has been given to RCA ground-noise reduction amplifiers, which causes the opening, as a function of modulation amplitude, to rise more steeply at first and then more slowly.[185] Relatively low-level modulation is then sufficient to cause an increase in margin and thus reduce subsequent clipping. It is when the modulation is nearly zero that close margin is urgent.

Certain characteristics of speech sounds have an important bearing on the design of ground-noise reduction systems. The positive pressure peaks are higher than the negative, and it is important to maintain correct polarity from microphone to valve, or to shutter and galvanometer.[192,195] R. O. Drew and I made an investigation to determine how often speech sounds build up rapidly.[196] Instances in which maximum amplitude was reached in less than three or four waves (voice fundamental) were surprisingly rare.

The ideal solution to the problem of avoiding initial clipping would be to anticipate increases in sound level. This is discussed by J. G. Frayne.[197]

Anticipation by use of a second microphone was used experimentally in the stereophonic system described in the October 1941 *Journal* (p. 351) for operation of the compressor.[329a] A system employing a 14-msec electrical delay network is described (March 1950) by Whitney and Thatcher.[198]

When reported it had been in use for over a year by Sound Services Inc. with very favorable results. Besides reduced clipping, advantage was taken of the system to increase ground-noise reduction in density recordings by 5 db, and in area recordings to work with a bias line only 1 to $1\frac{1}{2}$ mils wide. A low-distortion network good to 8000 cycles necessarily employs many sections (several hundred). Cost is probably the reason that this expedient has not been widely employed. A direct-positive variable-area recording system with the anticipation effect provided by means of an auxiliary exposure was designed and demonstrated by Dimmick and Blaney and has been used in the Warners' studios.[199]

In the re-recording operation anticipation should be a simple matter, involving only a double reproducing system (with scanning points a fraction of an inch apart) with separate amplifiers.* Mueller and Groves (June 1949)[200] mention use of this system at Warners. I understand however that little advantage has been taken of this possibility, presumably on account of costs. The explanation is probably that when initial clipping occurs in a well adjusted re-recording system, there was probably also some at the same spots in the original recording and the possible gain from anticipation in the re-recording operation is hard to detect. If clipping, in systems using ground-noise reduction, causes an appreciable impairment to sound quality, the ideal solution is to use for original recording a system whose ground noise is inherently so low that it needs no such expedient, and then introduce the ground-noise reduction when re-recording to the photographic tracks. Similar considerations apply to the use of compressors. Among the systems which in more or less measure meet this specification are Class B area recording, wide-track push-pull (with fine-grain film and fast-acting noise reduction) direct positives with the auxiliary exposing light,[199] and direct-playback disks. A real answer to this problem seems to have come in the recent adoption of magnetic recording.†, [200, 352–388]

Pre- and Post-Equalization. The practice of electrically exaggerating the high-frequency components in recording, as compared with the low-frequency components, in order to compensate for inevitable losses, began early in the recording of sound. Studies of the distribution of energy in program material had indicated that this could be done without resulting in overloads in the high-frequency end of the scale. A large amount of high-frequency pre-emphasis had been employed in cutting transcription disk records. In film records as well as disk records, ground noise can be made less noticeable by decreasing the gain at high frequency. Therefore, in addition to such relative attenuation of high-frequency sounds as was caused by the imperfections of reproducing systems and loudspeakers, the practice became prevalent of producing a drooping characteristic in the electrical circuits.

In order that all films might have good balance in all theaters, a committee of the Academy of Motion Picture Arts and Sciences made recommendations for a standard reproducing characteristic.[201] With such a standard adopted, the producers of sound pictures would have incentive to use recording characteristics which would give good balance when their films were played in a theater with the typical or standard reproducing characteristic.

The problem is discussed by J. K. Hilliard[202,203] and by Morgan and Loye.[201]

Better Lamps. Among the lamps used in the late 1920's for recording were some of the ribbon-filament type. These were ideal from the standpoint of uniformity, but required an inconveniently large current (18 amp) and did not have as long life at a given temperature as was obtainable with lamps of the helical-filament type. A series of lamps for sound recording and reproduction was made available by the lamp companies. The filaments were close-wound helices, of relatively heavy tungsten wire, and to permit operation at high temperatures with satisfactory life, the pressure of the inert gas (argon) with which the bulbs were filled was increased above that employed in lamps from which less intensity was required.

In the series of lamps described by F. E. Carlson of General Electric in 1939[205] the recording lamps were rated as operating at above 3100 K (color temperature) which is several hundred degrees higher than that of common incandescent lamps. Somewhat later, krypton-filled lamps were introduced, permitting still higher temperatures. The krypton, being heavier, more effectively retards evaporation of tungsten. (See the section on basic inventions.)

The light from a helical filament varies somewhat, depending on the angle from which the lamp is viewed. At the suggestion of L. T. Sachtleben of RCA, the helix of the lamp used in the variable-area recordings was curved, the convex side being presented toward the lenses. The helix with the curved axis gives definitely better uniformity.[183,205,205a]

Improvements in Light-Valves and Density Optical Systems. In June 1932 Shea, Herriott and Goehner[66] described the development of improved duralumin ribbon (stronger and with straighter edges) and better methods of adjusting and anchoring the ribbons at the ends of the free span. The new anchoring system practically eliminated the frictional hysteresis which had been found in the earlier design, thus reducing waveform distortion and making for greater stability. Stability became increasingly important as the mean spacing between the two ribbons in the valve was reduced. The ground-noise reduction system called for reducing the ribbon spacing when the modulation was low, and when it was found possible in view of required exposures to reduce the unbiased spacing from 0.002 in. to 0.001 in. this was done, while still keeping the optical reduction from valve to film at 2:1.

At high frequency and high modulation, the images of the ribbon edges move with velocities comparable with the speed of film travel. This results in a waveshape distortion which would convert a sine wave into a saw-tooth wave. Fortunately such a combination of high frequency and amplitude is not often encountered in program material, and the harmonics generated would probably not be reproduced, nor noticed if they were. However, it is desirable to minimize this "ribbon-velocity" distortion, and the reduced slit width helps in that respect and also in giving better resolution or high-frequency response.

In order that harmful effects (harm to sound quality if not to the ribbons themselves) might not result when the modulation current drives the ribbons to the point of touching or hitting each other (light-valve "clash") light valves have been built with the ribbons slightly offset, or in two planes.[207] There appears to have been difference of opinion about the necessity of this precaution. It should be remembered that in a density system the downward light modulation normally stops considerably short of zero, to avoid photographic nonlinearity. In other words touching of the ribbons would represent considerable overload. The two-plane design of valve has become generally standard for variable-density recording.

The optics of light-valve recording systems[207] have been modified by the addition of a small horizontal cylindrical lens close to the film, which results in greater optical reduction between valve and film and therefore a narrower image (0.0002 in.). This makes for improved high-frequency recording and for further reduction of ribbon-velocity distortion. One of the factors which has made the narrower image possible without sacrificing exposure is that new lamps of higher intensity have become available.

Light-valve optics have been adapted to making variable-area tracks, for example as used in the stereophonic system described in 1941.[208] The valve is turned with the ribbons vertical, or parallel to direction of film travel. The lens system, which employs cylinders, magnifies the ribbon motion ten to one. This means that the lens must be close to the ribbons, hence with little depth of focus. Therefore in this application the ribbons are in the same plane, and an electrical current-limiting expedient prevents clash.[209]

A strong magnetic field is advantageous for the sake of sensitivity and damping. In the design described by Wente and Biddulph[208] an air gap flux density of 32,000 gauss is attained, an achievement which testifies to the excellence of the permanent magnet materials and the high flux capacity of the pole-piece material. There is some further discussion of light valves in the section on variable density vs. variable area.

Microphones. While condenser microphones had excellent characteristics they were more expensive and required more servicing than magnetic microphones, and it was practically necessary to provide a stage of amplification close to the microphone. On the other hand, the electrical impedance of a magnetic microphone is such that a transformer may be used if wanted, and the output transmitted at a convenient impedance. A magnetic microphone of the flexibly mounted rigid-diaphram, moving-coil type is described by Jones and Giles in December 1931.[210,211] It is a pressure-type (rather than velocity or pressure-gradient) microphone. Damping is obtained by flow of air when the diaphram vibrates, back and forth between two cavities, through passages which are of such small dimensions as to make air viscosity effective in dissipating energy.

In June 1931 H. F. Olson described the velocity microphone,[212] consisting of a ribbon of very thin aluminum (0.0001 in.) in a magnetic field between pole pieces which are adjacent to the edges of the ribbon, so that when the ribbon moves in a direction normal to its surface a voltage is induced in the ribbon. A transformer is used to step up this voltage, which is then applied to the grid of an amplifier tube. Transverse corrugations are formed in the ribbon, which prevent it from curling and give it lengthwise flexibility. It is mounted under only such tension as is needed to keep it between the pole pieces. Olson shows that theoretically such a microphone should give uniform frequency response, and that it should have a polar directivity curve like a figure 8 (cosine law), the directivity being the same throughout the frequency range. Experimental results are also given confirming the theory. The velocity of movement of the ribbon is proportional to the velocity of air movement, so that it is often called a "velocity microphone."

Since a microphone of this type responds less and less as the direction of the sound departs from normal, it picks up much less reverberation (random in direction) than a nondirectional microphone having the same sensitivity for sound of normal incidence. The ratio of direct to reverberant sound in many cases sets the limit to how far from the source the microphone can be placed, and under such circumstances a ribbon microphone can get satisfactory pickup some 70% farther from the source than a nondirectional microphone, such as one of the pressure type.[213] Advantage has been taken of the directional characteristics of the ribbon microphone to exclude certain sounds or disturbances (for example camera noise), for it is deaf to sounds originating in the plane of the ribbon.

If the output of a pressure microphone is combined in correct phase and amount with that of a velocity microphone, the combination becomes unidirectional, having a cardioid-shaped directivity curve. It has a dead-spot 180° from the direction of maximum sensitivity. The forward directivity is much less sharp than that of a velocity microphone, and such a unidirectional microphone is better suited for picking up sound over a wide angle, as for example from a large orchestra. The cardioid directivity pattern has the same advantage as the figure 8 pattern in picking up less noise from random directions than a nondirectional microphone.

Before making a unidirectional microphone, Olson worked out an arrangement for converting a velocity microphone into a pressure microphone. He placed close behind the ribbon a combination shield and absorber consisting of an open-ended tube of the same cross-sectional area as the active area of ribbon. He distributed through the tube tufts of absorbent fiber. The length of the tube was made sufficient to dissipate wave energy. The impedance of the mouth of the tube then becomes equal to that of so much free air (to plane waves) but air in which there is no other sound to react on the ribbon. The ribbon is then actuated only by the pressure on the exposed side.[214]

Having successfully made this conversion, Olson applied the same treatment to only one half of the length of the ribbon, leaving the other half to act as a bidirectional velocity microphone, and the combination has the cardioid directional characteristics.[215]

Bell Laboratory engineers also developed bidirectional or velocity microphones, and unidirectional types, but differing from the Olson type in employing a second microphone more nearly like their standard pressure microphone. The unit and its applications were discussed by Marshall and Harry in September 1939.[216]

All studios use directional microphones for situations where the maximum ratio of direct to random sound is wanted.

* In a set-up for mixing numerous sounds, it would for practical purposes be sufficient to equip only the dialogue film-phonograph with double scanner and amplifier.

† Some experimenting has been done with a double magnetic pickup for anticipation. Cross-talk between the two heads was a problem. The difficulties will undoubtedly soon be worked out.

Loudspeakers[4,25,73,222]

Single Range Loudspeakers. For several years after the industry had adopted sound the theater loudspeakers were of the kinds already mentioned, (1) the directional baffle-type, using a coil-driven cone, much like those used in direct radiator speakers, with a short, straight-axis exponential horn of large throat area, and (2) those using long exponential horns[217,218] with small throats, and coil-driven metal diaphragms. In view of their length the horns were coiled or otherwise bent to a form which took up less space.

Multi-Range Loudspeakers and Improved Single-Range Speakers. The idea of providing separate devices to radiate high and low frequencies is undoubtedly of early origin. When practically all radiators had strong fundamental resonances, the double or triple unit could spread the range of reasonably high response over a wider frequency band. With the advent of coil-driven, untuned diaphragms, resort to separate radiators was a measure for improving efficiency, in that the design did not have to be a compromise between what was best for low and for high frequencies. A triple-horn speaker designed with special consideration to efficiency and load capacity was advocated and demonstrated by C. R. Hanna of Westinghouse in 1927.[219]

However theater speakers of the single-unit type were so far improved (by the coil-driven unit of Wente and Thuras in 1926,[57] and by the adoption of directive baffles for the GE-RCA cone-type speakers in 1929) that they handled quite well the frequency range then obtainable from film or disk.

As recording improved, the benefits from extending the loudspeaker range became more noticeable. After various improvements in the recording system including the new galvanometer, the symmetrical track, ground-noise reduction (by galvanometer bias), ribbon microphone for sound pickup, and a film-phonograph using the magnetic drive, Dimmick and Belar gave a demonstration of extended frequency range at the SMPE 1932 Spring Convention.[179] They did not resort to two-way (divided-range) speakers, for the straight-axis, directional baffle units, which had 6-in. cone diaphragms with aluminum voice-coils, had good response even at 10,000 cycles. The range was extended downward (to 60 cycles) by using slow expansion exponential horns (of the large-throat or directional-baffle type) 10 ft long, with mouth openings 75 in. square.

Multi-Range Speakers of Bell Telephone Laboratories. A divided-range speaker system was used by H. A. Frederick in the demonstrations of vertically cut disk records in the fall of 1931.[220] The high-frequency units were of the type described by Bostwick in the October 1930 *Journal of the Acoustical Society of America*, and in the May 1931 *SMPE Journal.*[221] These were equipped with small horns better to load the diaphragms. The low-frequency units were of the direct-radiator or flat-baffle type, using (as I recall it) approximately 12-in. diameter dynamic cone units,[38] a number of units being distributed over a large baffle. A curve indicates a response within ±5 db from 50 to 10,000 cycles.

A triple-range system is described by Flannagan, Wolf and Jones,[222] whose review of the development of theater loudspeakers is comprehensive and of much interest. The system is also discussed by Maxfield and Flannagan in the January 1936 *Journal.*[222a] The mid-range units were essentially like the previous single-range speakers, using the Western Electric No. 555 driver units. The radiators for the high-frequency range (3,000–13,000) were the same as used in the Frederick demonstrations. The authors state that for the range below 300 cycles large coil-driven conical diaphragms in a large flat baffle gave better results than designs using horns.

In April 1933 the Bell Telephone Laboratories gave a demonstration of reproduction of orchestra music in "*auditory perspective*,"[223,224] the orchestra being in Philadelphia and the reproduction in Constitution Hall in Washington, D.C. Three microphones picked up the music at three well-separated positions, and at the other end the independently transmitted and amplified currents were supplied to three correspondingly placed loudspeakers.

In this demonstration no recording and reproduction entered to affect frequency range, and it was essential for the purpose to provide abundant sound power and frequency range. A dual-range system was decided on.

The low-frequency unit was designed to work from 40 to 300 cycles, and consisted in a large-diaphragm, moving-coil unit, working into the 8-in. diameter throat of a horn which expanded exponentially to a mouth 60 in. square in a total length of approximately 10 ft.

The high-frequency driver unit, which covers the range 300 to 13,000 cycles is shown in cross section as Fig. 10 in the Flannagan, Wolf and Jones paper.[222] Particular attention is given in the design to the air space and passageways leading from the diaphragm surface into the horn or group of horns.

If a single straight-axis exponential horn is used, the tones of highest frequency are radiated in a direction close to the axis, while those of lower frequency are spread through much larger angles. This defect is avoided by dividing the total cross section of the passage into a number of smaller passages each of which is a small exponential horn. In this case there were sixteen horns for each driver. These are nested with their mouths adjacent and with their axes pointed in different directions to cover a total angle of about 30° vertically and horizontally. Since the horns are of equal length, the waves, whatever their frequency, unite at the ends of the horns to form a practically continuous spherical front, which is the condition for uniform distribution throughout the 30° angle. Two of these 16-horn nests were placed side by side to give the desired total of 60° horizontal coverage.

First Commercial RCA Two-Way Theater Speakers. In the RCA line of theater equipment a dual-range loudspeaker system was briefly described by J. Frank, Jr., at the 1935 fall meeting.[225] The speakers demonstrated by Dimmick and Belar, in 1932, using 6-in. cone diaphragms with aluminum voice coils, and 10-ft horns, were not seriously lacking in frequency range and were used in a number of deluxe installations, but they had two drawbacks. There were many theaters without sufficient room to install the long straight axis horns, and in addition, the high-frequency sound components were not well distributed. When a wave front reaches a point in an exponential horn at which the dimensions of the passageway are about a wavelength, its ultimate angle of spread will not greatly exceed the angle between the walls at that place. From this consideration it follows that a rapid flare horn would distribute the high-frequency sounds through considerably larger angles. Moreover with short rapid-flare horns, it is not impractical to multiply the number of units and thereby further control the sound distribution. In the theater speaker described by Frank, there were three high-frequency horns diverging in direction, the driver units being 6-in. cones with aluminum voice-coils. These units were rated to operate effectively from 125 to 8,000 cycles, and a separate folded horn unit took care of sounds in the 40 to 125 cycle range.

One of the practical advantages of a direct-radiator (flat baffle) loudspeaker is the small space it requires. However a horn makes it possible to radiate more sound from a given-sized diaphragm without increasing the amplitude of motion, and is therefore desirable for increasing the sound output capacity. It also affords some control of the direction of radiation. But to radiate low frequencies the rate of expansion (ratio of increase in cross section per unit distance along axis) must be small, which for a given total ratio of expansion means length. One way to provide a long passageway without requiring excessive depth of space back of the screen is to coil up the horn. Drawings of coiled horns are shown in ref. 1, p. 298, and on p. 251 of the March 1937 *Journal.* The bending of large sound passageways is objectionable. Instead of expanding continuously as in the ideal horn, short waves suffer repeated reflections by the walls, causing some irregularities in the response and making the direction of radiation of high-frequency sounds rather unpredictable.* On the other hand if the horn is to handle only low-frequency sounds, the shapes of the bends are not at all critical, and the condition is easily fulfilled that the difference between the shortest and longest paths around a bend is a small fraction of a wavelength.

In a common form of low-frequency horn (in the sense of an approximately exponentially expanding passageway) the driver unit (or units) is at the middle of the back of a box-shaped space, and the passage is forward for a short distance, dividing and forming two passages which turn back and then forward and expand to form a pair of large adjacent rectangular openings, which together form the mouth of the horn. This roughly describes the low-frequency unit of the theater speaker system reported by Frank, the drivers in that case being a pair of 8-in. coil-driven cones.

Shearer System.[226] In 1936 Douglas Shearer, sound director for M-G-M, gave demonstrations of improved sound, using loudspeakers described by J. K. Hilliard in the July 1936 *Journal.* The high-frequency radiators in this system were similar in many respects to those used for the Auditory Perspective demonstrations (see figure in Hilliard paper). The frequency range to be covered was 50 to 8000 cycles, and the division or cross-over was at 250 cycles.

The low-frequency unit was a folded horn, with four 15-in. cones in a vertical column. For simplicity of construction the expansion was all in the horizontal

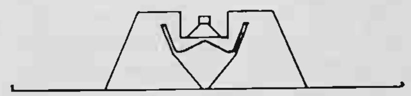


Fig. 16. Horizontal cross-section showing sound passages of folded low-frequency horn.

plane, accomplished by suitably arranged vertical partitions. The horn cross section was divided into two expanding passageways, whose final openings together form a 68-in. square (Fig. 16). This was surrounded by a flat baffle 10 × 12 ft, to reduce end reflections and improve the loading of the units. The mean length of each passageway was 40 in. (very short as low-frequency horn designs go), nor was the expansion ratio large, the throat area being sufficient to accommodate the four 15-in. cones.

* This effect can be largely reduced by careful design of re-entrant (zig-zag passage) horns.

A nest of high-frequency horns, similar to that used for the Auditory Perspective demonstrations, three high by six horizontally, covered a horizontal angle of about 100°. With the lengths of the high- and low-frequency horns nearly the same, there would be little time difference in the arrival of the sounds at the plane of the mouths, thus simplifying the avoidance of a "phasing" error which has been found to have detrimental effects with transients. However, in all divided-range speaker systems the best relative positions of the high- and low-frequency units have been determined by careful trials. This problem of "phasing" is discussed by Maxfield and Flannagan,[222a] by Hilliard[226] and others. It appears to be not wholly a question of minimizing the mean time difference, although that is a part of it.

In all divided-range speaker systems, dividing networks[227] have been used to separate the high- and low-frequency portions of the amplifier output and direct each to the appropriate speaker units. The networks consist in general of simple filter sections, and their design has received much study.

Commercial Two-Way Systems. Commercial models of dual-range or "two-way" theater speakers were brought out in 1936, employing the multicell high-frequency horn system, and low-frequency units much like those described by Hilliard. The high-frequency driver units of the RCA system differed from the ERPI and Lansing designs in that the diaphragms were of molded phenolic instead of aluminum. This resulted in a more rugged, if less sensitive, device. The reduced sensitivity and greater "roll off," or falling off at high frequency, can be readily compensated electrically, and do not mean any serious increase in amplifier output power, because the high-frequency components of the sound represent only a small part of the total sound power.

The ERPI "Diphonic" speaker system is described in the Flannagan, Wolf and Jones paper.[222]

The description by Hilliard of the Shearer low-frequency unit may be taken as in general typical of the commercial speakers of 1936–7.

Divided channel or "two-way" speakers came into wide use during the several years following 1936.

In some later designs of low-frequency units, the sound passageway was not folded, and consisted only of a short flaring connection between the driver units (which presented a large total radiating surface) and the large opening in the flat baffle.* It is of interest that the evolution of low-frequency sound sources has been toward a closer resemblance to the cone and baffle speakers of 1925, but greatly magnified in size, and with some "directive baffle" effect better to control sound distribution.

Higher crossover frequencies than the 250-cycle point of the Shearer system have prevailed, 400 cycles being the choice in many of the postwar units. In 1948 Hopkins and Keith[228] described the design of a two-way theater speaker in which the crossover had been raised to 800 cycles, observations having been made that the irregularities which are apt to occur at the crossover frequency are less prejudicial if the crossover is above the frequency range of maximum energy (250 to 500 cycles for orchestra music).

A photograph of a loudspeaker designed to use with "Cinerama" is shown in the May 1954 SMPTE Progress Report, p. 343. This is more or less typical of recent practice. The horn (or directive baffle expansion passage) of the low-frequency units as well as that of the high-frequency unit is designed to give exponential expansion of the total cross section by side walls which are radial, the floor and ceiling of each passage being curved to compensate. The reflex, phase-inversion principle (mentioned under "Monitoring Speakers") is employed to utilize radiation from the backs of the diaphragms, for the extreme bass. Note in the illustration the outlet slots on either side of the horn mouth.

Alternatives to Multicellular Horns. A somewhat simpler way of achieving the directive characteristic for which the multicellular high-frequency horns were designed has been developed in the post-war period by RCA and others. Horns are used with linear expansion in the horizontal plane (i.e. walls radial with respect to the throat), while in the vertical plane the rate of expansion is such as to bring the total expansion of the cross section to an exponential relation.

Another expedient for gaining the desired spread of high-frequency sounds was described by Frayne and Locanthi at the May 1954 convention of this Society.[229] If the waves issuing from a straight-axis exponential horn can be made to assume a spherical instead of nearly flat front, they will spread as desired. An acoustic equivalent of a concave optical lens is placed in the mouth of the horn, in order to retard the off-axis parts of the waves relative to the central part. The reduced velocity of propagation is achieved by means of a series of closely-spaced perforated sheet-metal baffles, the number of layers being progressively greater toward the edges. This system was reported to have

* This would raise the "cutoff" frequency of the horn, but where the total expansion ratio is comparatively small that is not necessarily very significant. Below its "cutoff" frequency, an exponential horn does not impede sound transmission. It merely fails to multiply the volume displacement as it does above cutoff.

given smoother distribution than the multicell horns.

Monitoring Loudspeakers.[4,230,231] Wider-range monitoring speakers kept pace with theater speakers. While a single conical diaphragm can be designed so that the center portion radiates high frequencies and the outer area radiates low frequencies, best results have been obtained by using separate diaphragms and separate voice coils, or in other words resorting to the dual-range system. The upper- and lower-range units may be adjacent or concentric. In the latter case the low-frequency diaphragm becomes a directive baffle for the high-frequency radiator.

Permissible cabinet size tends to set the lower limit of the frequency range, air reaction on the back of the diaphragm creating the problem if the back is enclosed, or inadequate baffling if an open-back cabinet is used. In order to utilize the radiation from the back of the low-frequency diaphragm, a second opening is often provided (for example below the diaphragm) and the space in the cabinet used to provide a folded horn between the diaphragm and the opening, or else to serve as a simple chamber which acts in conjunction with the inertia reactance of the air at the second opening as a phase inverter. This does not greatly augment the low-frequency output except near the resonance, set by the elastic reactance of the cavity and the inertia reactance of the combination of openings (one with the diaphram and one without). Sound-absorbing material is often used in the cabinet to reduce the magnitude of other resonances. If the horn-type back-wave system is used, its augmentation of output is limited at the lower end when the phase shift through the horn becomes less than about a quarter cycle, and at the upper end by the fact that it is deliberately designed to have a low-pass filter characteristic.

Developments Which Extended the Frequency Range

Better Light-Modulating System for Variable Area. Mention was made in the section on loudspeakers of a demonstration by Dimmick and Belar[179] of sound with extended frequency range. Aside from the improved loudspeakers and the ribbon microphone (whose response was practically uniform from 60 to 10,000 cycles) there were a number of advances that contributed to the result. The new galvanometer and optical system made so much more light available that it was feasible to reduce the width of the recording beam to ¼ mil, thus improving resolution. No small factor in giving clean high frequencies is avoidance of flutter, particularly rapid flutter such as 96 cycles. In the demonstrations, both recording and reproduction were on magnetic-drive machines.* (The rotary stabilizer was not yet available.)

Ground-noise reduction was by galvanometer bias with a single narrow line of transparent film when the modulation was zero, and it is my recollection that a measurement indicated a ratio better than 50 db between signal at full modulation and the ground noise when biased for zero modulation.

Nonslip Printer and Improved Sprocket-Type Printers. With the sprocket-type contact printers in almost universal use, a certain amount of slipping of the negative with respect to the print film is almost inevitable. The curvature at the sprocket compensates for a certain negative shrinkage at which the ratio of radii of the shrunk negative and unshrunk print stock is just equal to the ratio of their lengths. The sprocket diameter is designed to make this compensation correct for an average negative shrinkage, but it will be only approximate for negatives whose shrinkage differs from this assumed average. By mechanically stretching whichever film is shorter, two films can be moved together through an appreciable distance in perfect nonslipping contact. A printing system in which this was done was developed by the Technicolor engineers for their color transfer, but it was not used for sound prints. An extended investigation of the losses and irregularities in high-frequency response which result from slippage and imperfect contact during printing is reported by J. Crabtree in the October 1933 and February 1934 *Journals.*[232]

In 1934 C. N. Batsel described a nonslip contact printer[233] employing a principle proposed and demonstrated earlier by A. V. Bedford for a different application.[70,234] Bending a film stretches one face and compresses the other, and it is only necessary that the contacting surfaces of the two films be made equal. In the nonslip printer, the negative is rolled through the machine at fixed curvature, and the print stock (held against the negative at the printing point, where it is propelled solely by friction) is made automatically to assume the curvature at which identical numbers of sprocket holes are fed through in a given time.

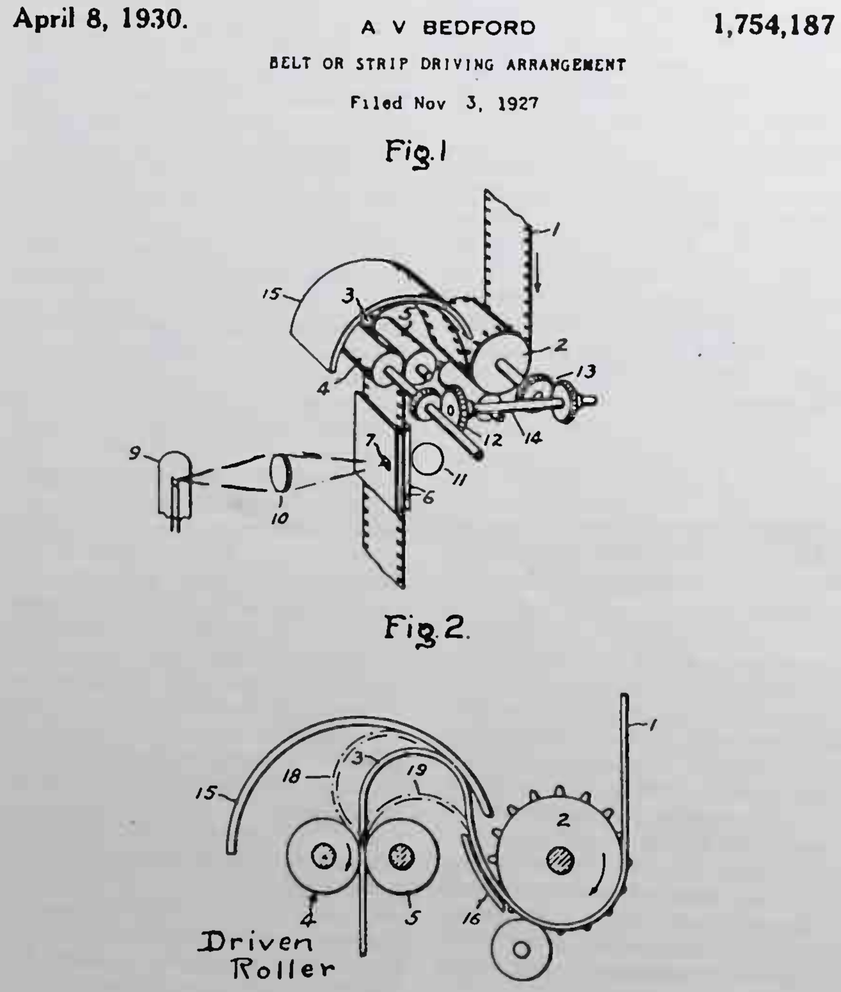


Fig. 17. Control of film speed by flexure (A. V. Bedford). Principle later used in nonslip printer.

* The special film-phonograph used in the demonstrations was a prototype of those used in the Disney *Fantasia* reproduction (see Fig. 7, p. 136, *Jour. SMPE*, Aug. 1941). The film was pulled by the magnetically driven drum over a curved supporting plate where the tracks were scanned, and was steadied at the other end of the plate by another drum with flywheel. This was an anticipation of the double flywheel system now used with magnetic tape and film.

The nonslip printer was not built for sale by RCA, but free license and drawings were offered, and numerous laboratories built and used them, notably Consolidated Film Industries in Hollywood and Ace Film Laboratory in Brooklyn.[235]

A printer based on the identical principle was developed independently by R. V. Wood.[236]

Although nonslip printers were not made in large numbers, and did not to any great extent displace the conventional sprocket type of contact printer, they served to demonstrate how much improvement was possible through better printing, and in this way stimulated makers and users of sprocket printers to improve their machines and maintain them in the best possible condition. In a number of laboratories it appeared that better prints resulted if the teeth were removed from the side of the sprocket next to the soundtrack, leaving only a smooth supporting rim. The theory behind this is not clear unless it is that such disturbances of contact as are due to film sticking on the teeth are thus kept further from the soundtrack.

There were two reasons which limited the use of the nonslip printer. Since there is a very narrow region of close contact, the printing beam had to be narrow. Under such conditions any speed irregularities cause variation in print exposure, which sometimes becomes audible in density prints. While there is no valid excuse for such speed variations, this trouble contributed to the conclusion that the nonslip printer was unsatisfactory for density printing. Film laboratory operators were naturally averse to maintaining one type of printer for area and another for density. A more fundamental fault of the nonslip printer in making density prints is due to the fact that negative and positive perforations are not maintained in register. The more active developer circulation close to sprocket holes tends to darken these areas, resulting in a slight 96-cycle hum. The effects on the negative and print are compensatory provided the printing is done with the holes in registration, but cumulative if out of register. With a purely friction drive at the printing point the relative positions of positive and negative perforations are unpredictable and may drift slowly between one condition and the other. This results in a hum that comes and goes and is therefore more conspicuous than if steady. Area tracks are far less sensitive to such development variations.

Optical printers with independent filtering of negative and positive film motion would be subject to the same uncertainty with respect to relative perforation positions, but this problem comes up only with 35mm-to-35mm printing, for which optical printers are rarely used.

In spite of such improvements as have been made in developer turbulation in the processing machines, the sprocket type of printer does not appear to have any strong competitor for density prints. In such a situation the efforts of several Eastman engineers[112,113] to improve sprocket action are well justified. The shrinkage-compensating sprocket described above under "Mechanical Systems" should result in great improvements. The new film-base materials with their reduced shrinkage should also make for better results with printers of the sprocket type.[391]

Optical Printers. Cost and quick production have been dominating considerations in much 16mm production, and contact printing on the sprocket has prevailed, the printers being simpler than other types and available in large numbers to meet the heavy wartime demands. However, registration of negative and positive perforations is not a factor in the quality of 16mm prints of either density or area tracks. There is thus no obstacle on this score either to nonslip printers or to optical printers with independently filtered film motion, and both types have been used, the nonslip for printing from 16mm negatives, and the optical principally for printing on 16mm film from 35mm negatives,[237–239] but also with a 1 : 1 optical system for printing from 16mm negatives (Precision Film Laboratories, New York).[240]

The Eastman Co. has designed several models of optical reduction printers in which the 35mm and 16mm films are on sprockets on opposite ends of a single shaft and a U-shaped optical system is employed.[241] This system minimizes chances of slow "wows," but depends for good results on the degree of excellence attainable with sprocket action. The shrinkage-compensating, radial-tooth sprocket described in the section on "Mechanical Systems"[113] was developed with printer applications particularly in view.

Ultraviolet Light for Recording and Printing

For truly distortionless recording the ideal light spot would be of infinitesimal width. In practice, as various improvements made light of greater intensity available, the nominal width of the recording spot was reduced (to 0.00025 in.), but a limit is set by diffraction. With light of shorter wavelength lenses of the same dimensions can form a more nearly perfect image. This is one of the considerations that led to a study by Dimmick of the possibilities of recording with ultraviolet light. Even more important than the improved image definition was the fact that ultraviolet light is more rapidly absorbed in the emulsion than is visible light. The increased absorption must be compensated for by increased intensity, but the net effect is that the exposure is confined more nearly to the surface where the silver halide grains are more completely used, and the result is less graininess. Most important of all is that the light scattered by the emulsion does not spread as far sidewise, and thus enlarge the image, as does visible light.[242,243]

Dimmick reported the results of tests with ultraviolet in August 1936.[244] He had found it possible to obtain adequate exposure with the regular incandescent lamps and Corning No. 584 filters.

Conversion for ultraviolet recording involved provision for the filter, redesign of the objective lens, and substitution of glasses with less ultraviolet absorption for other lenses in the system. Many of the variable-area recording systems were converted. Single-film newsreel systems (with the sound recorded on panchromatic negatives) were especially benefited by use of ultraviolet.[245,246]

The results obtainable by the application of ultraviolet exposure to variable-density films were studied and reported by Frayne and Pagliarulo in June 1949.[247] They found major improvement in definition and reduced distortion, but only about 1 db reduction in ground noise. Further experiences with ultraviolet are reported by Daily and Chambers.[248]

Many printers were modified to print sound with ultraviolet. High-intensity mercury arc lamps were widely used for the printers.

Ultraviolet exposure gives lower gamma for the same development than does white light. Advantage has been taken of this in certain cases of density recording to reduce the gamma of the sound print without departing from optimum development for the picture.

Coated Lenses. Each time any optical or light-source improvement made an increase in image intensity possible, the benefit could be realized in terms of finer resolution and better high-frequency response.

When the results were published of some tests in which the reflection from glass surfaces had been materially reduced by applying quarter-wavelength coatings of certain low-index minerals,[249] Dimmick procured equipment, mastered the techniques of applying such coatings by evaporation in vacuum, and compared the merits of various materials and methods of hardening the deposited layer. Having found a satisfactory procedure, he treated all the glass-air surfaces in the RCA recording optical system. Although the loss at each glass-air surface is only about 5%, there are sixteen such surfaces (excluding the lamp bulb), and by reducing each to a magnitude of between 1 and 2%, a gain of more than 50% was possible in the brightness of the image on the film. Not only is the image thus brightened, but the stray light due to "lens flare" is reduced to a small fraction of its previous magnitude. A review of the subject, containing further references, has been published by W. P. Strickland.[260]

The first RCA coated-lens optical system to be field-tested was a "variable-intensity" system taken to Hollywood by Dimmick early in 1938 and used experimentally during some months at the Fox studios, C. N. Batsel being the RCA liaison engineer.[261] I believe that I am right in saying that the demonstrations of this system were partly responsible for arousing interest in the benefits to be had by such treatment of glass surfaces.

Within a short time the lenses in a number of the RCA recording systems were given low-reflection coatings, and the practice of coating lenses in both area and density recording systems spread rapidly.[262]

Dichroic Reflectors. As a by-product of the work on low-reflection coatings, Dimmick studied multilayer coatings of alternate high- and low-index materials.[263] With these he was able to produce plates having very high reflectivity in one part of the visible spectrum and very high transmission in another (65% reflection at 7500 A and better than 95% transmission between 4000 A and 5000 A).

The first motion-picture application of his "dichroic reflector" was for reflecting the red light of the recording light beam to a caesium phototube for monitoring, while losing practically none of the blue and violet light that produces the exposure.[254]

During the war there was heavy demand for dichroic reflectors, designed to various specifications, for the armed forces, and also great demand for low-reflection treatment of many optical devices. Selective reflectors of this type have also played an important part in color-television developments.[255]

Heat-Transmitting Reflectors. Another outcome of the development work on dichroic reflectors is a design in which the transition from high to low reflectivity occurs just beyond the end of the visible spectrum, so that as a mirror it would compete well in efficiency and whiteness with the best silver or aluminum mirrors, while the invisible heat rays are better than 75% transmitted.[255,256] No comparably effective means of separating the useful visible from the infrared radiation has been hitherto available, and we may look forward to extremely valuable applications.[257,258]

Measures for Reduction and Control of Distortion

There is hardly space here to mention the many refinements, including principles of good design, by which distortion has been kept low in microphones, amplifiers, recording devices, optical systems and reproducing systems. We shall confine ourselves to discussing the sources of distortion associated with photographic soundtracks. These may be listed as:

(1) background noise caused by graininess (especially in density tracks) or by scratches or dirt in the track area (most serious in area tracks);

(2) noise (hum) and amplitude modulation resulting from sprocket hole proximity;

(3) fluttter or "wows;"

(4) high-frequency loss;

(5) irregular high-frequency loss, and modulation of recorded sound, due to printer defects;

(6) waveform distortion due to misalignment of the aperture (slit) either in recording or reproducing, and to uneven reproducing-slit illumination;

(7) waveform distortion and production of spurious tones due to nonlinear input-output relations (more frequent in density recording than in area);

(8) waveform distortion and production of spurious sounds due to finite width of the recording slit and image spread in the film emulsion (a problem in the area system); and

(9) distortion due to overloading.

Measures for improving sound records with respect to background noise are discussed in part under "Ground-Noise Reduction," but also include extreme care as to cleanliness (especially in processing), to hardening and lubricating (waxing) the film,[259] to maintenance of projector condition, and to the several expedients by which the maximum reproduced volume can be increased.

Fine-grain film has made perhaps the greatest single contribution to improvement in ground-noise ratio, especially for density recordings, but its benefits have been in so many directions that only the present topic heading (Reduction and Control of Distortion) seems broad enough to include it.*

Various developments which improved high-frequency response have already been mentioned, and to these should of course be added the use of fine-grain film. The topic "Improved Printers" has also been included in the section on better high-frequency response. Reduction of flutter is dealt with in the section on "Mechanical Systems." Analytical studies of the distortion that results from misalignment (or azimuth errors) of recording or reproducing slits in the case of area recordings, have been published by Cook[260] and by Foster.[261] In the case of density tracks the effect of misalignment is practically equivalent to widening the scanning slit, the results of which in terms of high-frequency loss are given in a paper by Stryker.[262] Both Cook and Foster have published analyses of the distortion in area recording due to slit width.[263,264]

Calculations of the distortion resulting from certain cases of uneven slit illumination and with several types of area track have been published by Cartwright and Batsel.[265] A test film has been made available for checking light uniformity.[266] It has a number of very narrow recordings of a tone in a series of positions across the track, and their relative outputs when played indicate light intensity at the corresponding positions in the scanning beam.

Sprocket Holes and Irregular Development. The proximity of sprocket holes to the soundtrack causes various difficulties. Mechanically the variations in stiffness of the film cause it to bend in a form resembling a polygon instead of a true circle.[232,267] This causes 96-cycle flutter, less perfect contact in printers, and 96-cycle modulation of high-frequency tones if optical systems have small depth of focus and are not exactly focused for the average emulsion plane. More serious from the sound standpoint is the effect of the holes on development.[268,269]

In film development there is always a tendency to nonuniform developer activity due to local exhaustion. For example, over and near a large dense area the developer is slightly weaker in its action than over the middle of a lightly exposed area, and this in turn is weaker than the average of the bath. If the partly exhausted developer as it diffuses out of the emulsion is not immediately carried away from the film by currents in the liquid, it may creep along the film surface and weaken the development in such areas. If there is little fluid movement except that caused by the travel of the film, the direction of travel is often evident from the appearance of pictures.[270,271] Circulation of the liquid is somewhat freer near the holes and edges of a film than elsewhere, and therefore the development and average density greater. This was discussed in connection with printers.

The irregular development troubles are reduced by very active stirring or turbulation of the developer while the film is passing through.[272] Improvements on this score are reported by Leshing, Ingman and Pier.[273]

* It has been said that the entire evolution of sound reproduction can be divided into two parts: (1) the learning how to make a noise and (2) the reduction of distortion.

Uniformity of development is further improved if the emulsion has such characteristics that its "gamma infinity" is only slightly higher than the desired gamma. Film manufacturers have been successful in producing fine-grain emulsions which come much closer to this desideratum than earlier types.

Waveform Distortions. The sensitometric studies and analyses about which we have already spoken are a part of the large body of literature bearing on the subject of distortion of the nonlinear type in density soundtracks. There have also been a number of experimental studies reported in this *Journal* giving distortion measurements with both density and area tracks. Two of the earliest of these were by Sandvik and Hall (October 1932)[274] and by Sandvik, Hall and Streiffert, October 1933.[275]

One of the most effective measures for reducing such distortions (second in importance only to avoiding overload) is by the use of push-pull soundtracks, and this applies about equally to density and area systems, although the causes of their distortions (Nos. 7 and 8 in the list) are quite different.

Two systems of testing for distortion (one for area and one for density systems) have assumed great importance since their introduction for ascertaining optimum ranges of exposure of negatives and positives. These are known as the "cross-modulation"[176] and the "intermodulation"[277] tests, and will be briefly described.

Overload. Overload in any element of the entire sound system other than in the soundtrack (either in the original recording or in the re-recording) means that the system has not been properly designed. It is easy to say that the way to avoid distortion due to overloads is not to overload, but that is asking too much. The price in terms of reduced average level is from a practical standpoint excessive. The levels set in recording are a compromise between too low an average level (with ground noise therefore more conspicuous) and too many and too bad-sounding overloads.*

In sections to follow, a number of expedients are mentioned for compassing a wider range between background noise and maximum reproduced level. These devices may be used either to make the background quieter or the loudest noises louder, and if properly applied can accomplish the latter with reduced instances of soundtrack overload.

Cross-Modulation Test to Determine Best Processing for Variable Area. While the variable-area system is not critical to gamma product it is subject to distortion, particularly at high frequency, due to the spread of the exposure outside the area of the actual image, caused by the scattering of light in the emulsion. This image spread increases with exposure. The distortion is largely in the form of a change in average transmission when a high-frequency wave is present. It is often referred to as "zero shift." It was particularly troublesome in some of the earlier 16mm recordings. It is possible however, by selecting the proper printing exposure, to make the image spread in the print neutralize that in the negative.[276]

*Most movie fans have probably forgotten what a gun sounds like.

In January 1938, J. O. Baker and D. H. Robinson described tests and equipment which became the standard method for ascertaining the correct relative negative and print densities.[176] For 35mm tests a 9000-cycle tone is 100% modulated at 400 cycles, but there must be no 400-cycle component in current supplied to the galvanometer. For a given negative exposure and development, a series of prints having different black area densities are made. They are then reproduced and the magnitude of the 400-cycle output measured and plotted (in db below the 9000-cycle level, or other reference) as a function of print density. The curve goes through a sharp minimum at optimum print density. With a negative of greater (black area) density the optimum print density is found to be higher. One of the most frequent problems is to select a recording exposure which will give maximum (or at least satisfactory) cancellation of cross modulation, when printed for a certain desired print density (say 1.4 to 1.6).*

Experience shows that if the 400-cycle cross-modulation is 30 db or more below full modulation, the sound is satisfactory. In similar manner if the print exposure and development are specified, the optimum negative density can be determined. For 16mm processing control, a 4000-cycle tone is modulated at 400 cycles. In push-pull recording the cross-modulation in the two parts of the track largely cancel each other and tolerances are extremely broad.

At the 1954 spring convention Singer and McKie[278] reported good results with an electrical compensating circuit, so that cross-modulation can be practically neutralized when making direct positives. This involves determining what function of frequency, exposure and (possibly) amplitude comes nearest to expressing the magnitude of the cross-modulation products.

*This test is made simpler than it might appear by the fact that when a high-contrast negative is used, the print black-area density is scarcely affected by differences in the black-area density of the negative, hence the printer light setting can generally be kept the same for the entire series of negatives.

Intermodulation Test for Variable-Density Control.[277] In the June 1939 *Journal* Frayne and Scoville describe a test for control of variable-density processing. In the earlier years of recording, the conditions for linear relation between negative exposure and print transmission were figured out by conventional sensitometric methods, with correction for the ratio of projection (or semispecular) density to diffuse density. But this is tedious, and a testing system which duplicated soundtrack recording and reproducing conditions was found preferable. In 1935 F. G. Albin[279] described a dynamic check on processing which consisted in making several recordings of a tone, all at the same fairly low amplitude, with a series of increasing mean valve-ribbon spacings. Comparison of the relative outputs from the print showed whether the slope of the curve of print transmission vs. negative exposure was constant or not. This was something of a forerunner of the intermodulation test. In the latter a tone of 1000 cycles or higher is superimposed on a low-frequency tone of, say, 60 cycles. Normally the level of the high-frequency tone is 12 db below that of the low tone, and the combined amplitude just short of overload. The print is run through a reproducing system, and the reproduced high-frequency tone separated out and measured for fluctuations in amplitude. Nonlinearity (distortion) becomes evident as fluctuations of the high-frequency tone, in general at rates equal to the frequency of the low tone or multiples of that frequency.

Fine-Grain Films.[280–282] While the many components of sound recording and reproducing systems were being improved, the great film companies were busy. In the period 1930–1940, the Eastman Co. offered at least seven new emulsions,[283] each of which offered some advantages for soundtrack application as compared with the currently used picture positive (Eastman No. 1301). In 1932 Eastman brought out No. 1359, intended for variable-density recording It was however similar to No. 1301 in contrast. In 1936 No. 1357 was brought out for variable-area recording, slightly faster than No. 1301 without sacrifice with respect to grain and resolution. With ultraviolet exposure it proved popular for variable-density recording. For release prints, however, the positive No. 1301 continued the almost universal choice except as similar films of other manufacture were used.

What was primarily wanted was finer grain, for this would mean cleaner, sharper images for variable-area tracks, less background noise in variable-density tracks, and better resolution and high-frequency response in all tracks. But fine grain in general means lower speed, or more exposure required, and too great sacrifice in this direction would make new films unacceptable. Another difficulty was that with many of the earlier fine-grain emulsions the image had a brownish tone, which was objectionable in the picture, and so ruled these emulsions out for the release prints. Measures were found for correcting this fault in Eastman fine-grain positive No. 1302. The problem of exposure was less serious in the case of fine-grain films for release and for duplicating than for recording stock, although some changes in printers were required. Special films were introduced about 1937 for duplicating purposes.[284]

One of the chief problems in variable-area recording was due to image spread (beyond the exposed area), but it was found possible to make image spread in the print compensate for that in the negative. The criticism of one of the trial emulsions intended for variable-area negatives (No. 1360) was that it was *too* good in this respect for the currently used print stock.

Re-recording is the universal practice. Graininess and other imperfections in the original negative are transferred to the print, passed on to the second negative by the re-recording operation, and in the final printing are added to the graininess of the release print. The use of direct positives offered promise of eliminating at least one step which was contributing its share of noise. In 1937 Eastman introduced its No. 1360, and Du Pont its No. 216, both fine-grain films designed for making variable-area direct positives.

Changing to a new release stock represented a greater hurdle than the use of special films for negatives and for the re-recording print, so the general thought was to evolve the best possible films for these applications, and take what improvement could be gained in this way. As the story turned out, use of the fine-grain films did not become general until they were used also for the final release prints.

Film graininess contributes to background noise in much greater measure in the density system than in the area system, but the requirements seem to have been harder to meet. Many tests were run and studies made during the years 1938–1940, and a special committee representing those particularly interested in variable-density recording reported in the January 1940 *Journal*,[289] in which tentative specifications were given to guide in the development of the needed fine-grain films.

Du Pont No. 222 was one of the first fine-grain films to gain considerable acceptance, especially at the M-G-M and Paramount studios, where it was used for original negatives, re-recording prints, re-recorded negatives and for a limited number of release prints. High-pressure mercury lamps were substituted where needed to give the required printing exposure. Experience at the Paramount studios is reported by C. R. Daily in the same issue.[286] Figures published in these reports indicate that while a re-recording print might be 8 to 12 db better in signal-to-noise ratio than one made on the previous types of film, only 4-to-6 db net gain was to be expected in the release print if this were on the usual stock. This indicates why the real gains from use of fine-grain film were not achieved until they were used for the release prints as well as in the earlier stages. (See also references 287–290.)

Some difficulty was encountered in processing the negatives to a suitably low gamma, but it was accomplished by use of special developers. (Uniformity of development is almost impossible if the time is cut so short that the gamma is far below the "gamma-infinity" for the given emulsion and developer.) In order to permit somewhat higher values of gammas in the fine-grain negatives the practice was adopted in some laboratories of holding down that of the fine-grain release prints by printing the sound with ultraviolet light, which gives lower gamma for a given development than does white light.

In 1937 the Eastman Co. brought out a high-resolution film (No. 1360) for variable-area recording and in 1939 a recording film (No. 1366) for variable density. In 1940 the No. 1302 fine-grain positive which has been widely used for release prints, became available. In 1944 the Eastman Co. brought out a high-contrast fine-grain film (No. 1372) for variable-area recording and a low-contrast fine-grain film (No. 1373) for variable-density recording, which could be processed in regular picture-negative developers instead of requiring special low-energy developers. The same year Du Pont brought out a new fine-grain recording negative (No. 236) for variable density, whose contrast could be more easily controlled than previous types because of low gamma infinity.

The fine-grain recording films are all much slower than the previous recording films, (No. 1302, for example, requiring about four times the exposure of No. 1301), and recording optical systems, although improved by coated lenses, could not with tungsten lamps provide sufficient exposures for the combination of fine-grain and ultraviolet light, but use of the fine-grain films with white-light exposure went so far in affording the benefits of high resolution and low noise that the combination seemed hardly needed. On the other hand, printing fine-grain positives with ultraviolet light from high-intensity mercury arcs has been widely employed.

It is of interest that the Callier coefficient (ratio of specular to diffuse density) is less with fine-grain than with coarser-grain films, and the result is that a higher control gamma (in which the measurements are of diffuse density) is needed to give the optimum picture on the screen—for example 2.5 for No. 1302 as compared with 2.1 for No. 1301.[290]

Variable-Density vs. Variable-Area.[291,292] The rivalry between advocates of variable-density and variable-area recording is nearly as old as commercial talking pictures and has unquestionably promoted progress in both systems. The fact that the industry did not standardize on one or the other track does not seem to have been any handicap, for all theaters could play either type of track.

During the earlier years of commercial sound, the advantage seemed to be on the side of area for music but density for speech intelligibility. With both at their best there was little to choose in clarity of speech reproduction, but the density system seemed able to take more abuse without too serious loss of articulation. Before control of zero-shift was well in hand (see cross-modulation test[176]) this type of distortion seemed to do more damage to the quality of the high-frequency reproduction than the nonlinearity to which the density system is more subject. With improved techniques and equipment this difference disappeared.

Another factor which for a time was prejudicial to the area system was that, as compared with density recordings, low-level passages seemed to be excessively weak when the controls were set for satisfactory normal and high-level passages. One theory was that, due to some fault in equipment or system, there was "volume expansion," or actual exaggeration of level difference. To those most familiar with area recording such a theory appeared untenable. Measurements did not bear it out. Monaural or single-channel listening itself makes level differences seem greater than direct binaural listening. Could it not be that the density recordings by their nonlinearity produced a compression effect? I recall that one of the most confident exponents of this explanation was M. C. Batsel. One of the fundamental differences in the systems is that an area track overloads very abruptly, whereas in a density track the upper and lower limits of light transmission are approached more gradually. With such a characteristic, considerable overload (i.e. beyond the range of true linearity) can be tolerated without too objectionable effect, and it was thought that considerable advantage had been taken of what might be called permissible overloading in the density recordings as compared with those made on the area system.

A nearly equivalent effect in an area recording can be obtained by rounding the corners of the V-shaped opening in the aperture plate (whose image is the light spot on the slit as shown in Fig. 13). Transition curves are introduced at the outer (full-slit illumination) ends, while the vertex is drawn out to a finer point. Such curved apertures were tried and largely overcame the cause of criticism.

However, the use of electronic compressors (long employed in broadcasting studios) appeared preferable to sacrificing the range of low-distortion light modulation. Compressors were introduced in the recording channels with very satisfactory results. The first such compressors were tried at RKO and Warner studios and compression became the standard. If the original recording is on a wide-range (or low-background-noise) system, most of the compression is introduced in the re-recording to the release-print negative, but I am told that even with magnetic recording some compression is used (or at least available) in the initial recording.

Any simple and general statement of the relative signal-to-noise ratios of the two systems is impossible, since each has its own cause and type of ground noise: grain hiss in density, and ticks and pops due to dirt or scratches in the area system. Area starts with an initial advantage of about 6 db greater output, and if the film remains in good condition its signal-to-noise ratio is better. The variable-area system was chosen by the Bell Laboratories engineers for their stereophonic demonstrations.

As various improvements became available to the industry they worked to the benefit now chiefly of one system and now of the other. Fine-grain films helped both, but helped density more.

The higher output from variable-area tracks led to the proposal by Levinson[293] that intercutting of the two types of track be used as means for increasing reproduced volume range.

The RCA light-modulation system has been modified to record variable-density, by use of an out-of-focus image or "penumbra" on the slit in place o the usual sharply focused triangular spot.[183,294] This system modulated the light reaching the film by changing its *intensity*, whereas the light valve produced a spot on the film of constant intensity but varying height, and therefor varied the *time* of exposure.

The light valve also was modified to make variable-area records by turning it 90° so that the ribbons were vertical (parallel with the direction of film motion) and the motions of their edges imaged at 10 : 1 magnification on the film. This was the system employed for the stereophonic demonstrations.[208,295,296] Variable-area optical systems using light valves were shortly thereafter offered as optional equipment to licensees.

By using valves having one, two or three ribbons as needed, the light-valve system has been made to produce many of the various types of area track that are produced by the RCA galvanometer optical system by changing apertures.[183,184,297]

The strict adherence to one system (area or density) on the part of the major picture companies has in considerable measure broken down, and several are employing both systems, depending on the type of operation required.[298] For example M-G-M, while continuing to use double-width push-pull density for initial recordings,* releases on variable area. One of the reasons for changing to area for release prints is that projectionists often neglected to change their gain settings when switching back and forth between area and density tracks, with a result to the disadvantage of the film with lower output (note also interest in Perspecta Sound.)

Mention is made, in the discussion of control tones, of a system in which the control tone is of subaudible frequency and superimposed on the recorded sound. Modulation of the sound by the control tone is much more easily avoided in area recording.

Compressors.[299] In the previous section comparing the area and density systems, the first urgent need of electronic compressors was described, but their use has proved so advantageous as to have become quite general, both in recording and re-recording channels.[300] Electronic compressors had long been in use in radio broadcasting.[304,304a] In their application to sound recording, if there is no provision for re-expansion, they do not actually increase the reproduced volume range although they may seem to. A common characteristic compresses 20 db into 10 db at a uniform rate. This would rob any passage of expression only to a slight degree. As the overload point is approached it is common to make the compression more drastic, for example 20 into 3 db. This is sometimes called a limiter type of compressor. If (for example by means of control tone) provision is made for re-expansion, the compression may be such as to keep the loud passages, whatever their original magnitude, at nearly full track amplitude. (See section on "Stereophonic.")

Since initial consonants usually carry little power compared with vowels, the compressor may wait to act for the vowel, with resulting relative exaggeration of consonants (especially sibilants). This has been corrected by a high-frequency pre-emphasis in the modulation which controls the compressor.[301–303]

Anticipation in compressor systems is desirable (see section on "Ground-Noise Reduction"). Fast action of compressors is essential and avoidance of "thump" is one of the problems.[305,306]

Experiences in recording with compressors are reported by Aalberg and Stewart of RKO[300]. Experience with compressors at the Warner Bros. Studios is reported by B. F. Miller.[301]

Squeeze Track.[307] In view of the limited range of loudness which a soundtrack permits, and the very great range encountered in our ordinary experience, many expedients were tried for the purpose of increasing the range obtainable from the film. One of these was the "squeeze track" described by Wesley C. Miller of M-G-M.[307] If a variable-density track is reduced in width, both the noise and the modulation are reduced. The ground-noise reduction system described by Silent and Frayne in May 1932[188] may not have been as yet available, but the two devices are not equivalent and can be complementary. There is a practical limit set by film characteristics as to how far the ground-noise reduction, by reducing mean negative exposure, could be carried, and this was at about 10 db of reduction, but the noise can be still further reduced by narrowing the track. The controls in recording or in re-recording are used to avoid too low levels on the track, but some of the range can be restored by narrowing the track in the release print. There were several ways in which this could be done; one, for example, was by preparing a masking film and running the print through the printer a second time with the track width determined by the mask.

"Reversed Bias" System.[308] An expedient for obtaining greater output for bursts of high level sound was described by Hansen and Faulkner[308] of Twentieth Century-Fox. In effect the light-valve bias operates in the usual way for normal and low levels, but for passages of extra high level the mean ribbon spacing is increased to as much as twice normal while still being fully modulated. There is some resulting distortion, but substantial increase in print output. The loss of some relatively high-frequency output is probably not objectionable.

Wide Track and Push-Pull. By doubling the width of the track the noise would be expected to increase 3 db but the usefully modulated light would increase 6 db. While the improvement is not large, it is worth while, and has been widely employed for original recordings and for special purposes, when the sound print does not have to carry a picture.

Push-pull systems have also been developed for both density[309] and area[183] systems and for wide tracks[310,311] and standard-width tracks. Standard-width push-pull systems have been considered for general theater reproduction, but their actual use has been limited to

* Probably later adopted magnetic recording.

places where the required special reproducing systems did not involve large investment. The push-pull system does not give any improvement in signal-to-noise ratio except for the few film blemishes or other disturbances which affect the light through both sides alike. The principal advantage of the push-pull system is that it reduces distortion, and in the density system it may thereby somewhat increase the permissible ratio of light modulation. As applied to area recording it almost eliminates zero-shift distortion. With both systems, push-pull operation permits the use of faster-acting ground-noise reduction, since the "thump" caused by rapid change in transmitted light is largely cancelled. Milestones in the progress toward better sound are usually fixed in our minds by major demonstrations. Such a demonstration was given at the 1935 spring convention. Using wide-track push-pull density recordings, improved two-way (divided-frequency-range) loudspeakers and amplifiers with abundant reserve power, the engineers of the M-G-M Sound Department, under the direction of Douglas Shearer, gave impressive demonstrations.

Push-pull wide-track is regarded as the last word in photographic recording and has for years been used in many studios for original recording.

If the original recording is wide-track push-pull, the positive (which is edited and then used as the master for re-recording to the release-print negative) would be a print of the original. If the original is magnetic, the favored practice has been to re-record to a wide-track push-pull direct positive for the edited film. In eliminating the step of printing, the direct positive minimizes quality losses.

Control Tone.[312] The most effective way to reproduce the great range of sound volume encountered in natural sounds is to resort to compression for the recording and to expand in reproduction. It has always been the practice for the recordist to use his controls to maintain the recorded levels between the limits set by overload and a satisfactory margin above noise. Were a record made of each change in recording gain, and the projectionist given a cue sheet by which he could at the right times make the inverse changes in reproducing gain, the natural levels could be restored. Such a method of operation was at one time contemplated. However, manual controls in recording are too slow, and manual restoration of level too unreliable. Automatic control of gain on the other hand has been used very effectively. If space can be found on the film for an extra soundtrack, a continuous tone can be recorded with either its amplitude or its frequency automatically correlated with the gain of the recording amplifiers, so that it provides a complete record of the recording gain throughout the recording.

For example, a voltage derived from the modulation to be recorded (and thus a function of its initial level) can operate on the recording amplifier to change its gain as in any automatic compressor, and the same voltage can operate simultaneously on the control tone to make the appropriate change in either its amplitude or frequency. In reproduction the output of the control-tone track is then used to provide a voltage which is directly related to that which altered the recording gain, and can therefore be used to produce inverse, or compensating, changes in the reproducing gain.

The use of an extra soundtrack or recording to be used in the manner just described was, I believe, first proposed by C. F. Sacia of the Bell Laboratories, and described in U.S. Pat. No. 1,623,756. One of the first in the RCA group to become interested in control tones was Charles M. Burrill who experimented with tones superimposed on the modulation, but of either too high or too low frequency to be reproduced by the audio system, especially a subaudible tone such as 20 cycles. For film recording he proposed scanning the sprocket-hole area, and varying the magnitude of the resulting 96-cycle tone by blackening more or less of the film between perforations. Thus, if these areas are left clear, the 96-cycle tone is comparatively feeble, while the maximum is produced if they are black.

The sprocket-hole control track was developed for application by H. I. Reiskind and adopted by Warner Brothers for their "Vitasound" system, which will be described in the section on multiple-speaker systems.[326]

Control tones of the subaudible type have in recent years found use for producing spread-sound and stereophonic effects which are particularly appropriate for accompanying large-screen pictures. (See Perspecta Sound, under "Multiple-Speaker Systems" below.) The chief advantage of this method of recording the control tones is that it requires no changes in the recorders and scanning systems, the changes being confined to the electrical circuits. It meets the important practical requirement that reproduction must be acceptable on standard equipment (with no provision for control-tone use). Variable-area recording is better suited than variable-density for use with subaudible control tones, since with suitable processing more accurate linear relationship between input and output can be maintained throughout the range of modulation, thus avoiding modulation of the audio by the control tones. The maximum level which can be recorded must be reduced 2 or 3 db, to make room for the control tones, so that the sum of the two will not exceed the permissible maximum. Tones of 30, 35 and 40 cycles have been used for the controls.

A control-tone system used by adherents of the variable-density system was described by Frayne and Herrnfeld.[313] Between the normal soundtrack and picture areas, space was found for a soundtrack 0.005 in. wide. The authors give their reasons for believing that frequency modulation of the control tone would be more reliable than amplitude modulation. With a frequency range of one octave and using a bandpass filter, they found that in spite of the narrowness of the track it provided a 38-db signal-to-noise ratio. The system was designed to afford changes of gain for the sound up to 30 db, thus expanding the dynamic range by that amount. The soundtrack was of the standard-density type.

In certain systems in which three or more independently recorded soundtracks are used, the sound and picture are on separate films. This gives plenty of room for the control track, but the practice has been to allot only one track to the control and to superimpose the tones, separating them in reproduction by bandpass filters.

Class B Push-Pull. Of the possible systems of photographic recording, the Class B area track undoubtedly carries furthest the principle of low print transmission when the modulation is low. Such a system was described by Dimmick and Belar in 1934,[314] and favorable experience in its use was reported in 1939 by Bloomberg and Lootens of Republic Pictures,[315] where it had been adopted for original recordings and been used in the making of fourteen pictures at the time of the report. They have continued to the present time using the Class B push-pull system for all original recordings. They also reported the methods used for test and adjustment. In the Class B system one side of the track carries only the positive parts of the waves and the other only the negative. When there is no modulation there is no clear film, and no ground-noise-reduction system is needed; therefore the transient or initial clipping which ground-noise-reduction systems can produce are avoided. Two triangles of light (see Fig. 13) are formed at the plane of the slit (one for each half of the track) and their vertices just touch the slit with no current in the galvanometer. Exposure takes place only as one or the other triangle illuminates more or less of the slit. A feature not described in the papers just mentioned is shown in a September 1937 paper by Dimmick.[183] In the mask which is imaged at the plane of the slit to form the triangular light spots, a narrow slit extends from the vertex of each triangular opening, so that light is never completely cut off at this point. There is thus formed a continuous line at the middle of each half track, not wide enough to let an appreciable amount of light through the film, for it is so narrow that it is largely fogged in, but it prevents a slight wave-shape distortion which might otherwise be produced by image spread, just as the tip of the triangle crosses the slit.

Direct Positives. Recording direct positives, instead of recording negatives and then making prints from which to make re-recordings, would seem to offer important advantages in simplifying operations and reducing time loss. There would also be an advantage in reduced ground noise since one less film is used and thus one less source of graininess. The high-frequency losses due to the printer would also be avoided. In the section on "Ground-Noise Reduction," it was stated that a ground-noise reduction system applicable to variable-area direct positives had been developed with the feature of anticipation, so that initial clipping need never occur. It was further possible under these conditions to work with closer margins, and so reduce the width of clear film. This system and results of tests with it are described by Dimmick and Blaney in 1939,[199] with a further report by Blaney in 1944.[316]

The chief obstacle to making variable-area direct positives had been that an entirely satisfactory way of avoiding distortion due to image spread had not been found. In the negative-positive process image spread in the negative could be compensated by that in the print, but the direct-positive system left no place for that solution. There is, to be sure, a certain density for any film at which there is no image spread, and below which the blackened area is slightly less than the exposed area. With the recording films available for a number of years, the density which gave no spread, or zero shift, was so low that light passing through the darkened areas would result in considerable noise. In some of the later fine-grain recording films, however, the balance occurred at densities which were satisfactorily high, thus making direct positives feasible under conditions in which they had not been before. However, the direct-positive system described by Dimmick and Blaney was not limited to the use of special films. They used a push-pull system which goes so far toward neutralizing the effects of image spread that excellent results were obtained with the standard fine-grain recording stock for variable-area.[316,317]

In variable-density recording the same anticipation feature is not applicable, but the other advantages of direct positives for original recordings apply. The problem of recording a positive lies in the requirement that the film transmission shall not be a reciprocal of the exposure (hyperbola) which a non-reversing film tends to produce, but must have an inverse relationship expressible by a downward-sloping straight line, covering a large range of transmission. A method of correcting *in the reproduction* for the nonlinear transmission characteristic of a variable-density negative was described by Albersheim in 1937,[318] but this would limit the usefulness of the direct positives. Electrical compensation in the recording was described by O. L. Dupy of M-G-M in 1952.[319] An approximation to the desired relation is possible by recording on the toe of the H&D curve, but high-level output is not possible from such direct positives without serious distortion.

A radically different solution of the problem was described by Keith and Pagliarulo in 1949.[320] Superimposed on the audio current supplied to the valve was a 24-kc bias current of twice the magnitude required to modulate fully the light-valve opening at normal unmodulated spacing. The ribbons are in different planes so that they can overlap without clashing. The authors reported 8-db higher output from the direct-positive than from a standard-density print, and 6-db higher signal-to-noise ratio than a standard print without ground-noise reduction. Direct positives, generally push-pull, have come into wide use for editing and re-recording service.[321,322]

Electrical Printing. Successful recording of direct positives opens the way to "electrical printing," or putting soundtracks on release prints by a re-recording operation instead of by contact printing. This is discussed by Frayne.[323] While such a method would necessarily be more expensive, the elimination of the flutter and irregularities resulting from the action of most contact printers is a strong argument in its favor. Frayne finds possibilities of greatly improved sound by this method, particularly for 16mm color prints of the reversible type such as Kodachrome or Ansco-color.

Engineers have long been attracted by the possibility of improved resolution and reduced distortion in 16mm positives by direct recording,[324] but until the more recent developments, direct positives (except on reversal films) were not a success.

Limits of Volume Range. While ideal sound reproduction would seem to call for duplication of original sound levels it is questionable whether illusion or seeming naturalness is improved by going as far as this in the direction of loudness. In the applications of control-tone systems, where very high levels are attainable, the extremes have been scaled down by 10 db or more. With respect to the desirability of carrying reproduction to extreme low levels, W. A. Mueller[325] has shown that there is a definite practical limit set by general theater noise.*

Multiple-Speaker Systems

It has often been observed that musical reproduction gives greater satisfaction if it comes from several sources. The Music Hall at Radio City in New York has been equipped to make this possible, and thus has afforded numerous opportunities to verify the advantages of multiple sources. With reference to the effort to obtain "dynamic range," Garity and Hawkins[333] state that: "Three channels sound louder than one channel of three times the power-handling capacity. In addition, three channels allow more loudness to be used before the sound becomes offensive, because the multiple source, and multiple standing-wave pattern, prevents sharp peaks of loudness of long duration."

That dialog should be reproduced on a speaker as near as possible to where the actors are seen is never questioned, but music and many sound effects such as thunder, battle noise and the clamor of crowds are far more impressive and natural if coming from sources all around the listener. The effects obtainable are discussed by H. I. Reiskind[326] who also describes the equipment used in the sprocket-hole control-track system.

Vitasound. This system, used by Warner Brothers, is described by Levinson and Goldsmith.[327] It is the simplest of the systems employing spread-sound sources and control tone. It uses three similar loudspeakers, the usual screen-centered dialog speaker and two side speakers outside the screen area. The design of the system is based on the theory that the volume range which the film (with the usual ground-noise reduction) affords is adequate for dialog and such other sounds as come from the center speaker only, and that higher sound levels will be wanted only for music and sound effects for which the spread-sound source will also be wanted. The control-tone output is therefore used, first to switch in the side speakers, with no change in total sound power. Further increase in control-tone output raises the gain on all three speakers, up to a total of 10-db increase.

Stereophonic Sound. Mention has been made in the section on loudspeakers of the transmission of music and other sounds "in auditory perspective" from Philadelphia to Washington in 1933.

* Background noise makes the useful sound seem fainter than sound at the same level in a quiet place. This holds even though the background noise is not loud enough to interfere nor even to make the listener conscious of its presence.

Three microphones were spaced across the stage and their outputs separately transmitted to three similarly located loudspeakers at the auditorium where the sound was reproduced. The various orchestral sounds seemed to come from the appropriate places, and a moving source such as a man walking across the stage and talking seemed at the receiving end to move about. In 1941 and earlier,[328,329] similar demonstrations were given, with the difference that this time the sound was recorded and reproduced. As in the previous case, every effort was made by the engineers of the Bell Telephone Laboratories and Electrical Research Products Inc. to minimize all forms of distortion and to reproduce the sound in the full dynamic range of the original.

Since it appeared that for the immediate purposes of the project the films could be maintained in good condition with respect to abrasions and dirt, and that under such conditions the advantage in terms of signal-to-noise ratio is with the variable-area system, the recordings were of the variable-area type, made with an adaptation of light-valve optics. The recorder and reproducing machine carried the film through the translation points on smooth (toothless) drums, on whose shafts were damped flywheels of the liquid-filled type described by Wente and Müller.[81,104]

On the basis that the film track was capable of giving a signal 50 db above noise, while the orchestral music to be reproduced had a range of 80 db, compressors were used in recording, and expandors in the reproducing system designed to compensate exactly* for the compression in recording. The compressors made no change of gain until the signal neared full track level and above this made the gain the inverse of the sound level, thus keeping the recorded level just below track overload. The level of the amplified microphone output controlled the magnitude of an oscillator tone, which tone was simultaneously recorded in a fourth track and applied to the recording-circuit compressor. In reproduction the same tone controlled the expandor gain. Because the levels at the three microphones did not necessarily rise and fall together, the compressors in the three channels were independently controlled by their own modulation levels. This necessitated use of three control tones, but these were recorded superimposed in one track and separated in reproduction by filters.

In order to give the compressor time to operate when a sudden increase of sound level occurred, an "anticipation" system was employed, using two microphones. The compressor operation was determined by output of a microphone closer to the source, while the second microphone supplied sound to be recorded.

As further insurance against audible background noise, the system of "pre-emphasis and post-equalization" of high-frequency components of the sound was employed. The usual ground-noise-reduction system by light-valve bias was not used, being made less necessary by the compressor-expandor (or "Compandor") system.

The October 1941 issue of the *Journal* carries discussions by Fletcher,[329] Steinberg,[330] Snow and Soffel,[331] Wente and Müller,[104] and by Wente, Biddulph, Elmer and Anderson,[332] of various aspects and elements of the stereophonic system.

Despite the unquestionable success of the stereophonic system in reproducing the subjective effects of sounds coming from the sources located as seen on the screen, and the impressiveness of the musical and sound effects which it was capable of handling, the motion-picture industry made no immediate move to adopt or apply it. A factor which undoubtedly militated against interest in utilizing the stereophonic system was that only for the patrons near the front of a theater did the screen subtend a large enough angle to make the difference between stereophonic and single-channel reproduction impressive. It appears to be another case of a development ahead of its time. With the recent advent (commercially) of the wide-film systems CinemaScope and Cinerama, stereophonic sound, supplemented by "spread sound," plays an essential role in providing the desired overall effects.

Fantasound: Fantasia. Walt Disney and his engineers had a somewhat different idea of what might be accomplished by means of multiple-track recordings with control tones. It might be appropriate to say that they proposed to make their spread-sound effects an art rather than a science. Duplication of an original distribution of sound sources was a secondary consideration, and the choice of directions from which sounds were to come was to be entirely at the discretion of the directors, musicians and technicians.

The story of the development of Fantasound and its evolution through numerous experimental forms beginning in 1937, was told by Garity and Hawkins,[333] with further reports by Garity and Jones,[334] and by E. H. Plumb.[335] Garity and Hawkins reported from tests that if a sound source (from loudspeakers) was to seem to move smoothly from one position to another, the output power from the two speakers should be held constant. This condition is not necessarily met by having the actual source move from near one microphone to near the other, but it can be met when the gain is reduced to one speaker and increased to the other by means of a knob or control tone.

The animated picture was designed specifically for the music, which was taken from great classics. In the initial orchestral recordings many microphones and separate recording channels were used. Recordings were selected or mixed in the re-recording to obtain desired effects such as predominance in turn of various orchestral groups (strings, brass etc.). Sound and picture were on separate films. The final sound film carried three 200-mil push-pull variable-area soundtracks, and three superimposed variable-amplitude control tones on a fourth track. The amplitude of the several control tones was determined by manual adjustments.

The theater equipment consisted of three loudspeakers at the center and to the sides of the screen and additional speakers at the sides and back of the auditorium. The latter could be brought into operation by relays responsive to notches in the edge of the film. They were used effectively for various sound effects and for the music of a large chorus. Abundant sound power and volume range were employed, the volume range being readily obtained by use of the control tracks.

The final Fantasound equipment was designed by RCA engineers. A special optical printer was used to print sound from separate negatives onto the multitrack positive, and special film phonographs in which a single system illuminated all four tracks, the transmitted light being received by four independent double-cathode photocells.

Fantasia was enthusiastically received but was not a financial success because of the heavy expense not only of its production but of "road-showing" with such elaborate equipment. The advent of World War II hastened its withdrawal, but Disney had performed a great service to the industry and art by pioneering in a sound-effects field which is now finding important applications.

Recent Multiple-Channel Systems.[336–342] During 1952 and 1953 three new systems of presenting pictures were introduced, known as "3-D" (Three-Dimensional), "Cinerama" and "CinemaScope." All of these represent more elaborate and expensive picture-projection systems, and with them would logically go whatever could be offered in the way of more impressive sound. The principles of stereophonic sound, plus the surround speakers distributed around the auditorium (as used in Fantasound), have been applied.

In the 3-D system two projectors are required, and a third synchronized machine, a film phonograph, is added for the sound. The sound system is stereophonic and uses three 200-mil magnetic tracks.[370,371]

The CinemaScope system[339] (developed by Twentieth Century-Fox) presents a much wider screen picture than does the standard system, and therefore presents an excellent opportunity for the use of stereophonic sound since the screen speakers are distributed over a large enough distance to make the shifts in the position of the sound source noticeable and needed for realism. The sound system developed to go with CinemaScope uses four magnetic stripes or tracks, two just inside the sprocket holes and two outside. To make more room for the tracks and leave as much as possible for the picture, sprocket holes of new shape have been designed. In the new perforation no valuable feature is believed to have been sacrificed, and the film will still run on standard sprockets. Three of the tracks provide the regular stereophonic sound (three speakers behind the screen) and the fourth carries sound effects for transmission by the surround speakers. For multitrack magnetic reproducing equipment see references 108–110.

Cinerama[340] (developed by Fred Waller and Hazard E. Reeves, and commercialized by Cinerama Productions Corp.) carries the wide-screen principle much further, employing three projectors to project adjacent edge-blended pictures on a curved screen. For the sound six magnetic tracks are used, and five speakers behind the screen, while the sixth track feeds a set of surround speakers.

The surround speakers, wherever these are used, are of the direct-radiator type, much less bulky than the screen speakers and with more limited frequency range.

Recent Multiple-Speaker Systems With Photographic Track.[339] A simpler multispeaker (or sound-placement) system has been developed by Robert Fine[342] and named "Perspecta Sound." Mention was made under the heading "Control Tones," of the use of control tones of subaudible frequency. In the Fine system there is only one soundtrack (variable-area, photographic) in the standard position, but three loudspeakers behind the screen, fed through separate variable-gain amplifiers. Three control tones, superimposed on the sound recording and separated in reproduction by filters, control the gains of the three amplifiers so that the apparent source can be made to shift across the screen, and the total sound level may be varied as desired to increase the dynamic range. The Perspecta-Sound recording if reproduced on a standard system (not equipped to make any use of the control tones) would, except for a slight reduction in level, be indistinguishable from a standard recording. M-G-M and Paramount have been particularly interested in Perspecta Sound and have equipped numerous theaters with it.[342]

Another sound-placement system was developed by the Dorsett Laboratories of Norman, Okla., and has been installed in a number of theaters in Texas and Oklahoma. It is described in the May 1954 Progress Report[339] as using a standard optical track, but with provision for shifting the sound from the center speaker to either of two screen speakers to right or left, or to peripheral (surround) speakers. This is accomplished by "switching cues in the form of a binary code marked into both sprocket-hole areas," and optically scanned. "Standard single-track optical release prints are cued for use with this system by the Dorsett representatives." (Quotations are from the Progress Report.)

Sound and Color. Since in color pictures the silver is removed from the emulsion, special handling of the soundtrack is required. In the Technicolor imbibition process the soundtrack is printed in the usual way, the remainder of the film is then cleared of all silver, and the pictures produced by dye transfer from relief masters.

With the multilayer color films such as Kodachrome, Kodacolor, Anscocolor and a number of more recent types it would be possible to expose the track with white light so that all three dyes contributed to the density, or with colored light so that the sound image would be principally in the top layer, and reproduce with a filter of the complementary color. The principal problem with this procedure is that existing projectors are nearly all equipped with caesium photocells whose sensitivity extends well into the infrared, in which range all of the dyes in use are transparent.[343] Substitution of phototubes of other types would mean an objectionable loss of sensitivity. In May 1946, A. M. Glover of RCA reported the development of a blue-sensitive phototube which could replace the caesium tubes in projectors without loss of output.[344,345] However, by this time the makers of color film had begun the use of edge-development processes that form a track of silver or of a metallic salt (opaque to infrared) not removed by the bleach which was part of the picture processing.[346] In one system the silver is in the top layer only, and the dyes in the other two layers contribute to the opacity. The metallic track became general in the handling of sound on color films.[347–350]

Sound on 16mm Film. I have confined my story to developments in the 35mm field in the belief that the general principles and practices which solved problems or led to improvements in one case were applicable in the other. It is probably a gross injustice to pass over the valuable work of engineers whose efforts were devoted to 16mm sound, but it seems to be a necessity.

From the start, the recording of sound at 36 instead of 90 ft/min. presented great difficulties. Only the fact that standards were much less exacting made the project practicable. But with each advance in the resolution of high frequencies in 35mm recording, the corresponding principles were applied to 16mm. One factor at least was in the favor of the low film speed, namely the providing of adequate exposure. By the time of the outbreak of World War II sound on 16mm film had improved to the point which made it acceptable for a great number of military applications.[351]

Recording sound on black-and-white reversal film was not too satisfactory, but years later when color films came into wide use, it again became necessary to put the sound on what was essentially a reversal type of film. Particularly successful for this purpose is recording on the individual prints or electrical printing. (See the section with that heading, and the December 1950 paper by J. G. Frayne.[323])

Black-and-white prints, with original sound recorded either on 16 or 35mm films, formed the great bulk of the product before and during the war. Fortunately, fine-grained emulsions could be used. Cost considerations dictated for the most part the use of 16mm negatives and contact prints, and unhappily nearly all of the printers were of the sprocket type. One laboratory adopted 1 : 1 optical printing as its standard.* Much study was given to the possibilities of recording direct positives. Still more development work was devoted to direct optical reduction from 35mm negatives, and this certainly helped resolution and high-frequency response,[239] but in the case of area tracks, neutralization of image spread in the positive by spread in the negative was not easily achieved.[276] The same factor made difficulty in the recording of direct positives, which had had advocates since early in the era of sound pictures. Since 35mm negatives for optical reduction, or 35mm re-recording positives to be used for making 16mm negatives give best results if recorded with different degrees and kinds of high-frequency pre-equalization, the organizations producing these masters have worked together to establish standards of recommended characteristics.[352]

Adding sound to 16mm films by use of a stripe of magnetic material promises to become of great popular and commercial importance.[353]

Drive-in Theaters.[354] The drive-in theater was first proposed and advocated

* Exactly, in terms of timing, but not necessarily fully restoring the 30 db of compression range.

* Precision Film Laboratories, New York, operated by J. A. Maurer.[240]

by R. M. Hollingshead, a businessman of Camden, N.J., not affiliated with motion-picture or electronic interests. The first such theater was built near Camden in 1933. In the earlier experiments with the system effort was made to put out enough sound power from a screen speaker to enable patrons to hear satisfactorily in their cars. This presented great technical difficulties, and also would have restricted theaters to locations where the noise would not be too objectionable. Several arrangements were tried, one with loudspeakers distributed over the field so that each speaker provided sound for two cars side by side. This was a great improvement from the noise standpoint, and the theaters previously equipped with screen speakers were converted. However, these "out-car" speaker arrangements still left something to be desired on the score of general noise. The "in-car" speaker, introduced by RCA in 1941, provided sound which was much more satisfactory to patrons and practically eliminated the neighborhood-noise problem. In the design of "in-car" speakers, the qualities of ruggedness, conveniently small size without too much sacrifice of sound quality, and immunity to damage by weather were design objectives.[355] The amplifier and audio-power distribution system had to be such that the individual levels would not be too much affected by the number of speakers in use. During the 1930's a number of drive-in theaters were built in New England, in the South, and (through the efforts of Philip Smith) in Indianapolis, Cleveland, Detroit, Milwaukee, St. Louis, Kansas City and Pittsburgh. Building activities were practically at a standstill during the war, but after its close the number increased rapidly. Underhill[356] gives the number of drive-in theaters at the end of the war as about 60 and four years later as 1000. He attributed the rapid growth just after the war in part to the greatly increased number of cars on the road, especially after the end of gasoline rationing, and in part to the shortage of building materials which made the construction of indoor theaters relatively difficult and expensive. The May 1954 Progress Report puts the number of drive-in theaters at the end of 1953 at 4000, and estimated that six months later there would be 4500.

The most serious problem of the drive-in theater is the provision of sufficient light for the large screen, a consideration, however, which lies outside the scope of this paper.

Magnetic Recording

The principle of magnetic recording was demonstrated in 1900 by Valdemar Poulsen, who called his device the "Telegraphone."[27,357] About 1917 the American Telegraphone Co. marketed a dictation machine under that name which performed well by existing standards but was complicated and could not compete commercially with cylinder machines. Recording was on 0.010-in. "piano" wire, contacted from opposite sides by a pair of offset steel pole-pieces. Erasure was by a saturating d-c field, and a d-c bias was combined with the recording-voice current in sufficient magnitude to maintain all the recorded magnetization of the reversed polarity.

An alternating magnetic field has for years been the accepted means of demagnetizing magnetic materials, and it seemed a paradox that a superimposed alternating current would assist in, rather than obliterate, recording. However, it was found in 1921 by W. L. Carlson and G. W. Carpenter, then working on a government project of recording telegraph signals, that a bias current of supersonic frequency could be very advantageous in magnetic recording of audio-frequency signals.[358]

The development of magnetic recording was carried on by the Bell Telephone Laboratories during the 1930's. A thin steel ribbon about $\frac{1}{8}$ in. wide was found better than wire, and alloys of superior magnetic properties were developed. Direct current erase and bias were used in the equipment.

The developments and publications of Marvin Camras[359,360] of Armour Research Foundation aroused widespread interest in magnetic recording and the results obtainable with high-frequency bias. The essential difference between a recording made with d-c bias, and one made with high-frequency bias is that with the latter (1) modulation is through zero, with resultant increase in maximum recordable amplitude, and (2) there is no remanent magnetism when the modulation falls to zero. The second feature is the condition for low background noise. The comparison is analogous to that between a Class B push-pull photographic record, and a unilateral recording without ground-noise reduction.

There was much developmental activity in wire recording just before and during the war, with numerous applications for the armed services. The extreme compactness of recordings stored as magnetized wire plus the ruggedness of the system with respect to mechanical injury made it of especial value for military uses. Better wires were developed, 4- to 5-mil diameter sizes being widely used. The Brush Development Co. developed a compound wire in which an alloy of superior permanent magnetic properties was plated on a brass core.[357] The National Standard Co. of Niles, Mich., developed a successful stainless-steel recording wire.

S. J. Begun[357] quotes an article in *Machinery* of January 1917 about the Telegraphone in which mention is made of a stripe of powdered iron to be painted on a motion-picture film to provide synchronous sound. It was more than 30 years later before use was made of this principle — one of the many examples of development in which the basic concept is only a small part of the invention, and the real contribution to progress the result of laborious experimentation and wise application of refinements and better techniques. Not until a sound quality was attained in magnetic recording surpassing that of any other known system did it become of great concern to motion pictures.

Development of magnetic coatings on paper or other base materials was undertaken by the AEG in Germany about 1928, but up to the outbreak of war nothing of outstanding quality had appeared. At the close of the war the American occupying forces brought back samples of a new German magnetic tape and equipment. The magnetic material was a finely divided iron oxide, mixed with a binder and coated on a thin cellulose base (total thickness about 0.002 in.). In the recording magnet a supersonic bias current was superimposed on the audio current. Magnetization was longitudinal, produced by C-shaped magnets of high-permeability alloy, with very short gaps where they contact the tape. Reproducing magnets were similar. In cleanness of reproduction, low ground noise and volume range the German system set a new high standard.

A period of intensive development followed the demonstrations. Some time was required before American firms could match the quality of the German tape, especially with respect to freedom from noise. Numerous papers were published describing basic properties of magnetic materials,[361,362] analysing the action of the supersonic bias,[363,364] and reporting tests on various tapes to determine optimum bias, relation of distortion to recorded amplitude, and amount of residual noise.[365] Among the first in this country to produce acceptable tapes was the Minnesota Mining & Mfg. Co.[361] Experimenters tried the magnetic black oxide (Fe_3O_4) in powder form and found it capable of producing higher output levels but noisier than the red oxide used in the German tapes. The red oxide (Fe_2O_3) in a certain crystal form ("gamma phase") is magnetic. Grinding to a particle size of about 1 μ or less was found essential, and prolonged mixing. Many tape-recorder-reproducer equipments were developed, both for amateur and professional use, the latter principally for broadcast stations.

Advantages of the Magnetic System. The compactness and portability of the magnetic-tape equipment, its freedom from dependence on laboratories, the immediate playback, the small storage space required, the relatively small cost of recording stock, plus the ability to re-use tape when the recorded sound on it is no longer wanted, the ability to work in daylight, and finally the excellent sound quality and dynamic range, combined to make the magnetic system a great advance from both the economic and performance standpoints.

Uses in the Motion-Picture Industry of Nonsynchronous Recording. The prompt interest of motion-picture producers in magnetic recording was shown in two lines of activity. The Basic Sound Committee of the Motion Picture Research Council held meetings in 1946 to which prospective suppliers of magnetic-recording equipment and record materials were invited.[366] The purpose was to formulate the requirements of the motion-picture industry in order that developments might be better directed toward meeting these requirements. The other activity was making experimental use of the nonsynchronous thin-tape recording equipment which was the first to be developed.[200] While the most important use which might be made of the magnetic system would be for the original synchronous recordings, there are numerous operations for which the lack of exact synchronism of magnetic tape would not be too serious an obstacle, such as for immediate playback after rehearsals, for training singers and actors, and for recording of music and sound effects which do not have to synchronize exactly with the picture. In musical numbers the picture is often secondary to the sound, and in many cases the practice had been followed of recording the sound first, without picture, and subsequently fitting the picture to the music. For this, initial recording on tape was applicable, the sound being re-recorded to film while musicians or vocalists performed synchronously for the camera.

Papers on Progress in Magnetic Sound. At the October 1946 convention of this Society Marvin Camras described and demonstrated synchronous magnetic sound, using 35mm film coated at Armour Research Foundation. However, the coated film was not offered commercially. At the same meeting H. A. Howell[362] of Indiana Steel Products Co. gave data on several magnetic materials and described a coated-paper tape which could be perforated if desired. While paper offered certain advantages which commended its use at the time, it has not been possible with coatings on paper to attain the quality and low noise that are realized with bases of clear plastic.

In February 1947 the Du Pont Co. furnished RCA with samples of coated 35mm film base. A conversion kit was developed so that a standard RCA photographic recorder (PR-23) could record and reproduce magnetically. The November 1948 issue of the *Journal* contains a description of the kit by Masterson,[367] measurements of the properties of the Du Pont film reported by O'Dea,[368] and bias studies by Dimmick and Johnson[365] of $\frac{1}{4}$-in. thin tapes made by Du Pont and by Minnesota Mining & Mfg. Co. as compared with one of the German tapes. The commercial version of the conversion kit was described by Gunby[369] at the October 1948 convention.. In the June 1949 *Journal* Mueller and Groves[200] describe experiences at Warner Bros. Studios with various uses of nonsynchronous magnetic recording, and such experience as had been had up to that time with synchronous magnetic recording. The practice of re-recording from magnetic to direct-positive photographic tracks is mentioned. This has sometimes been called "electrical printing," and soon became the prevalent method of providing a sound film to be edited and then re-recorded to the final release negative. The January 1952 paper by Carey and Moran[322] states that the practice of re-recording from a magnetic original to a 200-mil push-pull variable-area direct positive for editing had been followed by Universal International Pictures since January 1951.

The Progress Report of May 1951 shows RCA synchronous magnetic recording equipment, designed to make one, two, or three tracks on the film.* This was being used by Columbia, the sound being re-recorded to direct positives for editing. The same report tells of Westrex portable synchronous magnetic-recording equipment, suited (optionally) to 35mm, $17\frac{1}{2}$mm or 16mm film, describing this equipment as in wide use in this country and abroad. Triple-track Westrex equipment is described by Davis, Frayne and Templin[371] in the February 1952 *Journal*.

Portable magnetic-recording equipment (the complete channel weighing less than 100 lb) using $17\frac{1}{2}$mm film was described by Ryder[372] at the April 1950 convention. The May 1951 Progress Report states that since April 1, 1950, all Paramount production recording had been done on this equipment.

The Progress Report[373] of May 1952 states that by the end of 1951 approximately 75% of the original production recording, music scoring and dubbing in Hollywood was being done on magnetic-recording equipment.

Editing. While it is entirely possible to edit magnetic recordings with the help of quick-stop reproducers, sound-film editors have come to depend in part on the visibility of modulation in a photographic track. This is probably one of the chief reasons for the retention of photographic-sound records for this intermediate function. Push-pull recording is usual, and in order to add as little ground noise as possible, preferably wide-track. The use of direct positives saves time as well as printing losses. When the editing is completed, the negative for release printing is made by another re-recording operation.

To provide the visual advantages of the photographic track for purposes of editing while retaining the quality advantages of the magnetic system, an arrangement has been tried which registered by an ink line on the magnetic film the amplitude of the modulation being recorded.[372] Another expedient, described by Frayne and Livadary,[374] is to make simultaneous tracks on the film to be edited, one magnetic and the other photographic variable-area, the magnetic to be used for the re-recording.

A system in which both pictures and sound can be backed up at any time without loss of synchronism is described by George Lewin[377] of the Signal Corps Pictorial Center. This makes it possible for a narrator to correct or change his speech, erasing portions of the previous record as he substitutes the new.

Synchronous Thin Tape. While use of perforated film afforded the necessary synchronism, the thin tape has the important advantages of reduced size and weight of equipment, smaller space required for storage, lower cost, and better contact between magnet and record surface because of its greater flexibility. A number of systems have, therefore, been developed to drive tape in strict synchronism. These have for the most part used the principle of recording a tone on the tape in addition to the audio modulation and of controlling the speed of the driving system so as to hold within limits the phase relation between the reproduced tone and a reference frequency, usually the 60-cycle power which drives the camera. The recorded tone may be 60 cycles[375] (excluded from the audio by a high pass filter) or a tone slightly above the audio range for example 14,[376] 15[377] or 18[378] kc, and modulated at the 60-cycle rate.

An optical tone on the tape, as by stripes on the back of the tape would have certain advantages (one of which is that full voltage can be developed from standstill), but this has to the best of my information not been made commercial.

Experiments have indicated that with a suitably designed mechanical system, perforations and sprockets are not out of the question, even with tape as small as 0.002 in. thick by $\frac{1}{4}$ in. wide.[379]

* The use of three tracks on a 35mm film was begun by Columbia in November 1950. In single-channel recording the three tracks could be used in succession, or in recordings calling for combined voice, orchestra and sound-effects these could be independently and simultaneously recorded and later mixed for desired balance.

Striping. While magnetic sound has been thus far used mainly in the preliminary operations of making sound pictures, and release prints are still optical, there are some exceptions, for example the films for the CinemaScope system, on which the sound is all-magnetic. This calls for applying stripes of magnetic coating on the photographic film which carries the picture. There are also numerous applications, especially with 16mm films, for which it is desirable to add sound to an existing picture film, and other uses for which the stripe is applied to the unexposed film.

The application of the magnetic material in stripes of closely controlled width, position and thickness is no simple problem. The April 1953 *Journal*, Part II, carries a series of papers dealing with problems of magnetic recording. Striping is discussed by representatives of Reeves Soundcraft Corp.,[380] Minnesota Mining and Mfg. Co.,[381] Eastman Kodak Co.[382] and Bell and Howell.[383] The Minnesota Mining and Mfg. Co. developed a method of application of a stripe by transfer from a temporary supporting tape. Only heat (no solvent or wet cement) is required for the transfer. The Reeves engineers also have papers on the preparation of the magnetic material,[384] and the study of the minute surface irregularities which tend to lift the tape from the magnet causing sound "drop-outs."[385] In spite of the utmost effort to prevent the formation of such high spots they are not entirely eliminated, and polishing operations are helpful. Other papers in the group deal with wear on magnets,[386] measurements of magnetic induction,[387] and standardization.[388]

The May 1954 Progress Report mentions new high-output magnetic-oxide coatings introduced in 1953 and 1954 by Minnesota Mining and Mfg. Co., greatly increased use of striping, new machines for applying the stripes, and several designs of theater soundheads (to be mounted above the projector head) for reproducing multiple-track magnetic sound.

A Sound Committee Report by J. K. Hilliard in the June 1953 *Journal* tells of arrival at agreement for a standard of track positions for triple 200-mil magnetic tracks on 35mm film, of projects for standardizing theater magnetic-reproducing characteristics, and plans for making available various magnetic test films, corresponding to the long-used photographic test films.

Theater reproducing systems for magnetic tracks are designed to be mounted on the tops of projectors, and do not interfere with the optical reproduction.

New Safety-Film Base.[280,389] For years the motion-picture industry struggled to minimize the fire hazard of nitrate film. Safety-film base had early been developed and used for certain purposes for 35mm film, and was mandatory for amateur equipment (16mm), but the stability and mechanical properties of the safety base were so inferior to the nitrate base that it was not a satisfactory substitute.

The film companies worked long and diligently on the problem of improving film-base stock. In 1937 the Eastman Co. adopted an improved safety base,[390] and in 1948 announced a new safety base which combined the needed properties to replace the long-used nitrate base, being superior to the nitrate in heat resistance and low shrinkage. It is described by Fordyce[391]. With these virtues the new film has rapidly supplanted nitrate.

The improvement in sound quality due to reduced shrinkage may not be noticeable to the average listener, but it must inevitably mean better performance especially in the action of contact printers. The importance of this new base can hardly be exaggerated.

Acknowledgments

I wish to express my appreciation and indebtedness to the many who have supplied information, and reviewed portions or the entire text, submitting comments and indicating corrections. Special acknowledgment must be made to E. C. Wente who supplied me with much of the story at Western Electric, and to J. G. Frayne who reviewed a large part of the text, indicating corrections and answering many questions. A. C. Blaney gave particularly generously of time and thought in answering questions and filling in gaps of information. My thanks also go to E. M. Honan, and W. V. Wolfe, O. B. Gunby, E. W. Templin, Kurt Singer, Barton Kreuzer, and Walter L. Tesch for reviews and comments.

APPENDIX

Development of High Vacuum Amplifier Tubes from the Audion

A brief account of the improvement of the audion by Dr. Arnold of the Western Electric Co., and of the subsequent patent interference and litigation, appeared in *Jour. SMPE, 17:* 658–663, Oct. 1931. In view of the vital part played by the electronic triode, it seems appropriate to tell something more about the parallel developments at the General Electric Co. The results, in terms of practical, high-vacuum amplifier tubes, were attained independently and at very nearly the same time at the two laboratories. The Telephone Co. had immediate use for audio amplifiers and promptly put them to use on a large scale, whereas, at the General Electric Co., the work was much more nearly a pure research.

Study at the Research Laboratory of the General Electric Co. of the relation of "Edison Effect" current to residual gas was an outgrowth of incandescent lamp development. Because of serious effects in some cases of the minutest traces of impurities or of certain gases or vapors, the techniques of producing high vacua had been developed to a high degree. In the early days of the Fleming valve and the de Forest audion the opinion was widely held that these depended for their conducting properties on the presence of some gas. Experiments in 1907 by Prof. Soddy[395] of the University of Glasgow lent support to the belief that the space current would become zero if a true vacuum were obtained. On the other hand, the English scientists J. J. Thompson (who first proved that there was such a thing as an electron) and O. W. Richardson believed that high temperature would cause a conducting body to throw off electrons into the adjacent space, without dependence on the potent effects (in releasing electrons) of impacts on the cathode surface by positive gas ions. It was the question which of these theories was right that enlisted the interest of Dr. Irving Langmuir. By using the known techniques for driving occluded gases from glass and metal surfaces, the Gaede diffusion pump, and adding a liquid-air cooled trap in the exhaust line, Dr. Langmuir exhausted his experimental tubes to about the highest possible vacuum. He showed that (with adequate temperature of the tungsten cathode) a pure electron space current flowed, and that this current followed the theoretically predicted 1.5 power relationship to the anode potential.[397]

With high enough anode voltage to carry over all the electrons emitted by the cathode ("saturation current") Langmuir was able to verify O. W. Richardson's prediction ("The electrical conductivity imparted to a vacuum by hot conductors," *Phil. Trans. 201:* 497, 1903) that the rate at which electrons are "boiled" out of the cathode bears a similar relation to temperature that vapor pressure does in the case of an evaporating substance. The presence of small quantities of ordinary gases (other than the "noble" gases) was found to "poison" the tungsten surface and greatly reduce emission, but this effect disappeared at high enough temperatures.[397]

These studies were begun in August 1912, and continued through that year and the following. Three electrode tubes exhausted to high vacuum were found to be free from the voltage limitations and erratic behavior of the previous audions.[398] Another important outcome of the development of the pure electron discharge was the Coolidge X-ray tube (hot-cathode, high-vacuum type) in which the electron velocities at the anode (and thereby the frequency or penetrating power of the X-rays) could be accurately and reliably controlled, and carried to much higher values than had been possible in the previous tubes where gas ionization had limited the effective anode voltage. The use of chemical "getters" for improving the vacuum in sealed-off tubes was also much advanced by work at this time at General Electric. In October 1913, a patent application was filed for Langmuir on "Electrical Discharge Apparatus" (triode) in which conduction is entirely by electrons, the effects of gas ions being negligible.

The audions which de Forest supplied for radio reception broke into a glow discharge if anode potentials of more than 20 to 40 were applied, and all control by the grid then vanished. The very limited output with low-voltage operation was not serious for radio detectors, but made the tubes of little or no use as amplifiers. During the years 1909–1912 de Forest was employed by the Federal Telegraph Co. of California. The company wanted amplifier tubes and de Forest ordered some made with better vacuum, calling in some cases for re-exhaust, so that by August 1912, a tube was being used at 54 v and by November, one at 67½ v. This was held in some of the later court actions to be a clear indication of the direction from which improvement in the audion could be expected and thus, so far as invention or discovery goes, an anticipation of possible patent claims by others, directed to improving the audion by employing higher vacuum.

At the time that the work just described was going on, the Telephone Co. was making plans to establish transcontinental telephone communications in time for the opening of the San Francisco Panama-Pacific Exposition in 1915. There had been a long-felt need for a voice-current amplifier, much effort had been expended on the project, and devices based on various principles tried.

For the 3000-mile transmission the possession of amplifiers would be imperative and they must have low distortion to permit cascading. According to the accounts by Lovette and Watkins,[48] and by Wm. R. Ballard,[393] a visit by de Forest in October 1912 served to direct attention to the possibilities of the audion, other devices of promise having till then claimed the research efforts. De Forest gave demonstrations and left a sample for tests and study. The demonstrations were repeated next day for the benefit of research engineer Dr. H. D. Arnold, who was quick to recognize the potentialities of the audion, the requirement for high vacuum and the role of space charge in limiting and controlling electron current. He expressed entire confidence that an amplifying tube which would meet the requirements could be developed from the audion, and he was assigned the task. Progress was rapid, and tubes with much higher vacuum were quickly available, but vacuum of the desired value was not achieved until a Gaede molecular pump, which was ordered from Germany, arrived. Long-lived cathodes were of great importance for telephone applications, and cathodes of the Wehneldt oxide-coated type were developed to replace de Forest's tungsten cathodes.

More details about the developments at the Western Electric Co. will be found in the references already cited.[48,393] Doubting the patentability of the improvement brought about by higher vacuum, Arnold and his attorneys did not file any application until they learned that an application based on a similar development had been filed by the General Electric Co. The prolonged interference which followed is summarized in the footnote on page 657 of the October 1931 *Journal*, "After various conflicting opinions by successive tribunals, U.S. Pat. No. 1,558,436 was issued to Langmuir in 1925."

In 1926 the General Electric Co. brought suit for infringement against the de Forest Radio Corp. Again there were decisions, appeals, and reversals, ending with the ruling of the Supreme Court, May 25, 1931, that the patent did not involve invention.

Some light on the questions at issue may be found in opinions written by the successive courts.

> Court of Appeals of the District of Columbia (339 *Official Gazette* 56). Patent issued pursuant to finding of this court.
>
> Federal District Court for Delaware before which the General Electric vs. de Forest suit was tried (*Federal Reporter, Vol. 23, 2nd Series*, p. 698).
>
> U.S. Circuit Court of Appeals for the Third District (44 *Federal Reporter 2nd Series*, p. 931, and *U. S. Pat. Quarterly*, Oct.–Dec. 1930, p. 67). Majority opinion for plaintiffs and dissenting minority opinion.
>
> U.S. Supreme Court (283 U.S. 664, *Baldwin Vol. XI* 664).

Bibliography and References

1. Lester Cowan, *Recording Sound for Motion Pictures*, McGraw-Hill Book Co., New York, 1931; H. G. Knox, "Ancestry of sound pictures," Chap. 1.
2. H. B. Franklin, *Sound Motion Pictures*, Doubleday, Doran and Co., 1929.
3. Academy of Motion Pictures Arts and Sciences, *Motion Picture Sound Engineering*, D. Van Nostrand & Co., New York, 1938.
4. J. G. Frayne and H. Wolfe, *Elements of Sound Recording*, J. Wiley & Sons, New York, 1949.
5. W. E. Theisen, "Pioneering in the talking picture," *Jour. SMPE, 36:* 415–444, Apr. 1941.
6. E. I. Sponable, "Historical development of sound films," Pt. 1–2, *Jour. SMPE, 48:* 275–303, Apr. 1947; Pt. 3–7, *ibid.*, 407–422, May 1947.
7. F. H. Richardson, "What happened in the beginning," *Trans. SMPE*, No. 22, 63–114, 1925.
8. T. A. Edison, U.S. Pat. 200,521, 227,679.
9. Terry Ramsaye, *History of the Motion Picture*, Simon & Schuster, New York, 1925.
10. Terry Ramsaye, "Early history of sound pictures," *Trans. SMPE*, No. 35, 597–602, 1928.
11. A. Dickson and W. K. L. Dickson, *History of the Kinetograph, Kinetoscope and Kineto-Phonograph*, Albert Bunn, N.Y., 1895.
12. W. K. L. Dickson, "A brief history of the Kinetograph, Kinetoscope and Kineto-Phonograph," *Jour. SMPE, 21:* 435–455, Dec. 1933.
13. Will H. Hays, *See and Hear*, p. 40, Motion Picture Producers and Distributors of America, New York, 1929.
14. E. H. Amet, U.S. Pat. 1,124,580 and 1,162,433, 1915.
15. W. H. Bristol, "An electrical synchronizing and resynchronizing system for sound motion picture apparatus," *Trans. SMPE*, No. 35, 778–789, 1928; and "New synchronizing apparatus for 16 mm films with disk records," *Jour. SMPE, 14:* 361–365, Mar. 1930; also U.S. Pat. 1,234,127, 1917.
16. Oskar Messter, *Mein Weg Mit dem Film*, Max Hesse Verlag, Berlin Schöneberg, 1936; Bk. Rev., *Jour. SMPE*, 332, Sept. 1937; also Brit. Pat. 22,563, 22,564, 22,565.
17. T. A. Edison, U.S. Pat. 1,182,897, 1916; also Daniel Higham, U.S. Pat. 1,036,235, 1,054,203, 1,036,236 and 1,226,883.
18. E. W. Blake, *Am. J. of Science*, No. 16, 54, 1878.
19. A. C. Hardy and F. H. Perrin, *The Principles of Optics*, McGraw-Hill Book Co., New York, 1932.
20. V. K. Zworykin and E. G. Ramberg, *Photoelectricity and Its Application*, J. Wiley & Sons, New York, 1949.
21. L. R. Koller, "Characteristics of photoelectric cells," *Trans. SMPE*, No. 36, 921–939, 1928.
22. M. F. Jamieson, T. E. Shea and P. H. Pierce, "The photoelectric cell and its method of operation," *Jour. SMPE, 27:* 365–385, Oct. 1936.
23. L. de Forest, U.S. Pat. 841,386, 841,387, 879,532.
24. See refs. 392–399.
25. Frederick V. Hunt, *Electroacoustics*, J. Wiley & Sons, London and New York, 1954.
26. T. A. Edison, U.S. Pat. 307,031; and J. A. Fleming, U. S. Pat. 803,684.
27. V. Poulsen, U.S. Pat. 661,619, 789,336, 873,078, 873,083. "Der Telegraphon," *Annal. der Physik*, 3, 1900.
28. Wallace C. Sabine, *Collected Papers on Acoustics*, Harvard University Press, 1922; and *Acoustics and Architecture*, McGraw-Hill Book Co., New York, 1932.
29. See refs. 400 et seq. — on lamp developments.
30. T. Mishima, "Nickel-aluminum steel for permanent magnets," *Stahl und Eisen, 53:* 79, 1933; and U.S. Pats. 2,027,994 to 2,028,000 incl; also W. E. Ruder, "New magnetic alloys," *Proc. IRE, 30:* 437, 1942.
31. Ref. 6, also gives the references *Annal. der Phys.*, 103, 1901, and *Phys. Zeit.*, No. 34, 498, 1901; Ref. 5, p. 421, and Crawford, ref. 32, refer to an article by Rühmer in the *Scientific American*, July 29, 1901.
32. Report of the Historical Committee, *Jour. SMPE, 16:* 105–109, Jan. 1931; Merritt Crawford, "Pioneer experiments of Eugene Lauste in recording sound," *Jour. SMPE, 17:* 632–644, Oct. 1931.
33. R. W. Paul, "Kinematographic experiences," *Jour. SMPE, 27:* 495–512, Nov. 1936; *also*, Thomas Armat, "My part in the development of the motion picture projector," *Jour. SMPE, 24:* 241–256, Mar. 1935. (These are of interest, though they do not deal with sound.)
34. J. T. Tykociner, "Photographic recording and photo-electric reproduction of sound," *Trans. SMPE*, No. 16, 90–119, 1923.
35. Hans Vogt, Joseph Massole and Josef Engl, U.S. *Tri-Ergon* Pats. 1,512,681; 1,534,148; 1,555,301; 1,557,678; 1,558,032; 1,566,413; 1,590,185; 1,597,323; 1,608,261; 1,628,377; 1,634,201; 1,756,681; 1,825,598; 1,713,726; Re. 20,621; 2,140,003 and others.
36. L. de Forest, "The Phonofilm," *Trans. SMPE*, No. 16, 61–75, May 1923; "Phonofilm progress," *Trans. SMPE*, No. 20, 17–19, 1924; "Recent developments in the Phonofilm," *Trans. SMPE*, No. 27, 64–76, 1927; "Pioneering in talking pictures," *Jour. SMPE, 36:* 41–49, Jan. 1941.
37. Theodore W. Case, Honor Roll Award, *Jour. SMPE, 48:* 437–440, May 1947.
38. Chester W. Rice and E. W. Kellogg, "Notes on the development of a new type of hornless loud speaker," *Trans. AIEE, 44:* 461, Apr. 1925.
39. Carl L. Gregory, "Resurrection of early motion pictures," *Jour. SMPE, 42:* 159–169, Mar. 1944.
40. Theodore W. Case, "Thalofide cell, a new photoelectric substance," *Phys. Rev.*, Apr. 1920, and *J. Opt. Soc. Am.*, No. 6, 398, 1922; U.S. Pats. 1,301,227 and 1,316,350.
41. Theodore W. Case, "Aeo-Light," U.S. Pat. 1,816,825.
42. Edward B. Craft, Honor Roll Award, *Jour. SMPE, 48:* 440–443, May 1947.
43. J. I. Crabtree, "The work of Edward Christopher Wente," (1935 Progress Medal Award), *Jour. SMPE, 25:* 478–482, Dec. 1935.

44. E. C. Wente, "A condenser as a uniformly sensitive instrument for the absolute measurement of sound intensity," *Phys. Rev.*, July 1917; U.S. Pat. 1,333,744.
45. I. B. Crandall, "The air damped vibratory system: theoretical calculation of the condenser transmitter," *Phys. Rev.*, 449, June 1918.
46. E. C. Wente, "Electrostatic transmitter," *Phys. Rev.*, 498, May 1922.
47. W. C. Jones, "Condenser and carbon microphones, their construction and use," *Jour. SMPE, 16:* 3–22, Jan. 1931.
48. Frank H. Lovette and Stanley Watkins, "Twenty years of talking movies, an Anniversary," *Bell Telephone Magazine*, Summer, 1946.
49. J. P. Maxfield and H. C. Harrison, "Methods of high quality recording and reproducing of music and speech based on telephone research," *Trans. AIEE, 45:* 334, 1926.
50. H. A. Frederick, "Recent advances in wax recording," *Trans. SMPE*, No. 35, 709–729, 1928.
51. E. W. Kellogg, "Electrical reproduction from phonograph records," *Trans. AIEE, 46:* 903, June 1927.
52. E. O. Scriven, "Western Electric sound projecting systems for use in motion picture theatres," *Trans. SMPE*, No. 35, 666–678, 1928.
53. H. Pfannenstiehl, "A reproducing machine for picture and sound," *Trans. SMPE*, No. 38, 253–267, 1929.
54. H. M. Stoller, "Synchronization and speed control of synchronized sound pictures," *Trans. SMPE*, No. 35, 696–708, 1928.
55. I. M. Green and J. P. Maxfield, "Public address systems," *Trans. AIEE, 42:* 64, 1923.
56. E. C. Wente, U.S. Pat. 1,812,389, "Acoustic device" (i.e. loudspeaker).
57. E. C. Wente and A. L. Thuras, "A high efficiency receiver of large power capacity for horn type loud speakers," *Bell Sys. Tech. J.*, Jan. 1928, p. 140.
58. Samuel L. Warner, Honor Roll Award, *Jour. SMPE, 48:* 443–446, May 1947.
59. A. C. Hardy, "Optical system of the oscillograph," *J. Opt. Soc. Am. 14:* 505, June 1927.
60. A. C. Hardy, "The optics of sound recording systems," *Trans. SMPE*, No. 35, 760–777, 1928.
61. A. C. Hardy, "The rendering of tone values in photographic recording of sound," *Trans. SMPE*, No. 31, 475–491, 1927.
62. L. A. Jones, "On the theory of tone reproduction with a graphic method for the solution of problems," *Jour. SMPE, 16:* 568–599, May 1931.
63. E. C. Wente, U.S. Pat. 1,638,555, "Translating devices" (light-valve).
64. Donald MacKenzie, "Sound recording with the light valve," *Trans, SMPE*, No. 35, 730–747, 1928.
65. T. E. Shea, W. Herriott and W. R. Gochner, "Principles of the light valve," *Jour. SMPE, 18:* 697–731, June 1932.
66. C. A. Hoxie, "A visual and photographic device for recording radio signals," *Proc. IRE, 9:* 506, Dec. 1921; and U.S. Pat. 1,758,794 (filed Aug. 1927).
67. H. B. Marvin, "A system of motion pictures with sound," *Trans. SMPE*, No. 33, 86–102, 1928.
68. V. K. Zworykin, L. B. Lynn and C. R. Hanna, "Kerr cell method of recording sound," *Trans. SMPE*, No. 35, 748–759, 1928.
69. C. R. Hanna, U.S. Pat. 2,003,048 (conditions for critical damping).
70. E. W. Kellogg, "A review of the quest for constant speed," *Jour. SMPE, 28:* 337–376, Apr. 1937.
71. H. F. Olson, "Recent developments in theater loudspeakers of the directional baffle type," *Jour. SMPE, 18:* 571–583, May 1932.
72. Louis Malter, "Loudspeakers and theater sound reproduction," *Jour. SMPE, 14:* 611–622, June 1930.
73. H. F. Olson and F. Massa, *Applied Acoustics*, P. Blakiston & Son, Philadelphia, 1934.
74. C. R. Hanna, "The Mitchell Recording Camera, equipped interchangeably for variable area and variable density sound recording," *Trans. SMPE*, No. 38, 312–317, 1929.
75. Harry W. Jones, "The modern newsreel," *Jour. SMPE, 14:* 204–208, Feb. 1930.
76. P. M. Robillard and E. B. Lyford, "Recent developments in RCA Photophone portable recording equipment," *Jour. SMPE, 16:* 269–276, Mar. 1931.
77. *L'Industrie du Film Parlant*, Conservatoire des Arts et Métiers, Feb. 17, 1929. *Contains a resume covering the Gaumont and the Gaumont-Peterson-Poulsen sound systems. (The foregoing is from the Theisen historical paper—ref. 5.)*
78. H. E. Roys, "The measurement of transcription-turntable speed variation," *Proc. I.R.E., 31:* 52–56, Feb. 1943.
79. T. E. Shea, W. A. MacNair and V. Subrizi, "Flutter in sound records," *Jour. SMPE, 25:* 403–415, Nov. 1935.
80. E. G. Shower and R. Biddulph, "Differential pitch sensitivity of the ear," *J. Acoust. Soc. Am., 3:* 275–287, Oct. 1931.
81. W. J. Albersheim and D. MacKenzie, "Analysis of sound film drives," *Jour. SMPE, 37:* 452–479, Nov. 1941.
82. E. W. Kellogg and H. Belar, "Analysis of the distortion resulting from sprocket hole modulation," *Jour. SMPE, 25:* 492–502, Dec. 1935.
83. SMPE Committee on Sound, "Proposed standard specifications for flutter or wow as related to sound records," *Jour. SMPE, 49:* 147–159, Aug. 1947.
84. E. W. Kellogg, "Proposed standard for measurement of distortion in sound recording," *Jour. SMPE, 51:* 449–467, Nov. 1948. Further discussion of the Flutter Standard by the same author appears in an editorial in the *Trans. I.R.E.*, July–Aug., 1954.
84a. Frank A. Comerci, of the Material Laboratory, Bureau of Ships New York Naval Shipyard, Brooklyn, N. Y., "Perceptibility of flutter in speech and music," *Jour. SMPTE, 64:* 117–122, Mar. 1955. Discussion, 318, June 1955.
85. M. S. Mead, U.S. Pat. 1,854,949.
86. E. W. Kellogg and A. R. Morgan, "Measurement of speed fluctuations in sound recording and reproducing equipment," *J. Acoust. Soc. Am., 7:* 271–280, Apr. 1936.
87. A. Goodman, R. J. Kowalski, W. F. Hardman and W. F. Stanko, "Safeguarding theater sound equipment with modern test instruments," *Jour. SMPE, 34:* 409–423, Apr. 1940.
88. R. R. Scoville, "A portable flutter-measuring instrument," *Jour. SMPE, 25:* 416–422, Nov. 1935 Also: "Laboratory flutter-measuring instrument," *Jour. SMPE, 29:* 209–215, Aug. 1937.
89. F. P. Herrnfeld, "Flutter-measuring set," *Jour. SMPTE, 55:* 167–172, Aug. 1950.
90. U. R. Furst, "Periodic variations of pitch in sound reproduction by phonographs," *Proc. I.R.E., 34:* 887–895, Nov. 1946.
91. L. A. Elmer and D. G. Blattner, "Machine for cutting master disc records," *Trans. SMPE, 37:* 227–246, 1929.
92. P. H. Evans, "A comparative study of sound on disk and film," *Jour. SMPE, 15:* 185–192, Aug. 1930.
93. H. Pfannenstiehl, "High-precision sound-film recording machine," *Jour. SMPE, 29:* 202–208, Aug. 1937.
94. E. W. Kellogg, "A new recorder for variable area recording," *Jour. SMPE, 15:* 653–670, Nov. 1930. (*The title of this paper is misleading, in that the main features of the recorder were in no wise related to the type of sound track to be recorded.*)
95. E. W. Kellogg, U.S. Pat. 1,892,554, 1,899,571, and Re19,270.
96. R. O. Drew and E. W. Kellogg, "Filtering factors of the magnetic drive," *Jour. SMPE, 35:* 138–164, Aug. 1940.
97. See ref. 94.
98. A. G. Zimmerman, "Film recorders," *Jour. SMPE, 20:* 211–227, Mar. 1933.
99. M. E. Collins, "A deluxe film recording machine," *Jour. SMPE, 48:* 148–156, Feb. 1947. (*Model PR-31*)
100. H. A. Rowland, an early disclosure (1902) of the general principle of the damped flywheel is in U.S. Pat. 691,667, 1899, and 713,497. (See refs. 81 and 104.)
101. F. J. Loomis and E. W. Reynolds, "New Apparatus—A new high fidelity sound head," *Jour. SMPE, 25:* 449–460, Nov. 1935; and "New rotary stabilizer sound head," *Jour. SMPE, 27:* 575–581, Nov. 1936.
102. E. D. Cook, "The technical aspects of the high-fidelity reproducer," *Jour. SMPE, 25:* 289–313, Oct. 1935.
102a. Gerhard Schwesinger, "The compliance of film loops," *Jour. SMPTE, 57:* 320–327, Oct. 1951.
103. G. Puller, "Sound-picture reproducing system for small theaters," *Jour. SMPE, 27:* 582–589, Nov. 1936; Report of the Progress Committee, *Jour. SMPE, 27:* 29, July 1936.
104. E. C. Wente and A. H. Müller, "Internally damped rollers," *Jour. SMPE, 37:* 406–417, Oct. 1941.
105. C. C. Davis, "An improved film-drive filter mechanism," *Jour. SMPE, 46:* 454–464, June 1946.
106. G. R. Crane and H. A. Manley, "A simplified all-purpose film recording machine," *Jour. SMPE, 46:* 465–474, June 1946.
107. M. E. Collins, "Lightweight recorders for 35 and 16 mm films," *Jour. SMPE, 49:* 415–424, Nov. 1947. (*Models PR-32 and PR-33.*)
108. J. D. Phyfe and C. E. Hittle, "Film-pulled theater-type magnetic sound reproducer for use with multitrack films," *Jour. SMPTE, 62:* 215–220, Mar. 1954.
109. C. C. Davis and H. A. Manley, "Auxiliary multitrack magnetic sound reproducer," *Jour. SMPTE, 62:* 208–214, Mar. 1954.
110. S. W. Athey, W. Borberg, and R. A. White, "Four-track magnetic theater sound reproducer for composite films," *Jour. SMPTE, 62:* 221–227, Mar. 1954.
111. Wesley C. Miller, "M-G-M recorder and reproducer equipment units," *Jour. SMPE, 40:* 301, May 1943.
112. J. S. Chandler, "Some theoretical considerations in the design of sprockets for continuous film movement," *Jour. SMPE, 37:* 164–176, Aug. 1941.
113. J. G. Streiffert, "The radial-tooth, variable-pitch sprocket," *Jour. SMPTE, 57:* 529–550, Dec. 1951.
114. "Rulings of the U.S. Supreme Court in recent patent cases of the American Tri-Ergon Corp.," *Jour. SMPE, 24:* 529–550, June 1935.
115. Report of Standards and Nomenclature Committee, *Jour. SMPE, 14:* 126, 133, Jan. 1930.
116. Oscar B. Depue, "A machine for printing picture and sound simultaneously and automatically," *Jour. SMPE, 18:* 643–648, May 1932.
117. A. S. Howell, B. E. Steckbart and R. F. Mitchell, "The Bell & Howell fully automatic sound picture production printer," *Jour. SMPE, 19:* 305–328, Oct. 1932.
118. A. S. Howell and R. F. Mitchell, "Recent improvements in the Bell & Howell fully automatic printer," *Jour. SMPE, 22:* 115–126, Feb. 1934.
119. Roscoe C. Hubbard, "Printing motion picture film," *Trans. SMPE*, No. 28, 252, Feb. 1927.
120. J. Crabtree, "Sound film printing-I," *Jour. SMPE, 21:* 294–322, Oct. 1933; "Sound film printing-II," *Jour. SMPE, 22:* 98–114, Feb. 1934.
121. J. I. Crabtree and C. E. Ives, "A new method of blocking out splices in sound film," *Jour. SMPE, 14:* 349–356, Mar. 1930.
122. F. D. Williams, "Methods of blooping," *Jour. SMPE, 30:* 105–106, Jan. 1938.
123. W. H. Offenhauser, Jr., "Current practices in blooping sound-film," *Jour. SMPE, 35:* 165–171, Aug. 1940.
124. E. I. Sponable, "Elimination of splice noise in sound film," *Jour. SMPE, 26:* 136–144, Feb. 1936.
124a. George Lewin, "A new blooping device," *Jour. SMPE, 48:* 343–347, Apr. 1947.
125. L. E. Clark, "Some considerations in the design of sound-proof camera housings," *Jour. SMPE, 15:* 165–170, Aug. 1930.
126. Report of Progress Committee, *Jour. SMPE, 27:* 3–44, July 1936. Twentieth Century-Fox camera, p. 8.
127. H. R. Kossman, "A silent camera," *Jour. SMPE, 21:* 420–425, Nov. 1933.
128. D. B. Clark and G. Laube, "Twentieth century camera and accessories," *Jour. SMPE, 36:* 50–64, Jan. 1941.
129. J. P. Maxfield, "Acoustic control of recording for talking motion pictures," *Jour. SMPE, 14:* 85–95, Jan. 1930.
130. A. S. Ringel, "Sound-proofing and acoustic treatment of RKO stages," *Jour. SMPE, 15:* 352–369, Sept. 1930.
131. M. C. Batsel, "Recording music for motion pictures," *Jour. SMPE, 25:* 103–108, Aug. 1935.
132. R. H. Townsend, "Some technical aspects of recording music," *Jour. SMPE, 25:* 259–268, Sept. 1935.
133. M. Rettinger, "Scoring-stage design," *Jour. SMPE, 30:* 519–534, May 1938.
134. V. O. Knudsen, "Hearing of speech in auditoriums," *J. Acoust. Soc. Am., 56*, Oct. 1929.
135. W. A. MacNair, "Optimum reverberation time for auditoriums," *J. Acoust. Soc. Am., 242*, Jan. 1930.
136. S. S. A. Watkins and C. H. Fetter, "Some aspects of a Western Electric sound recording system," *Jour. SMPE, 14:* 520–530, May 1930.
137. B. Kreuzer, "Recent improvements in the variable-width recording system," *Jour. SMPE, 27:* 562–574, Nov. 1936.
138. W. P. Dutton and S. Read, Jr., "Some new RCA photophone studio recording equipment," *Jour. SMPE, 16:* 315–329, Mar. 1931.
139. S. Read, Jr., "RCA Victor high-fidelity film recording equipment," *Jour. SMPE, 20:* 396–436, May 1933.
140. S. Read, Jr., "Neon type volume indicator," *Jour. SMPE, 28:* 633–642, June 1937.
141. F. G. Albin, "Linear decibel-scale volume indicator," *Jour. SMPE, 29:* 489–492, Nov. 1937.
142. L. T. Sachtleben, "Characteristics of Photophone light-modulating system," *Jour. SMPE, 25:* 175–191, Aug. 1935.
143. L. J. Anderson, "High fidelity headphones," *Jour. SMPE, 37:* 319–323, Sept. 1941.
144. H. F. Hopkins, "Considerations in the design and testing of motion picture screens for sound picture work," *Jour. SMPE, 15:* 320–331, Sept. 1930.
145. Charles R. Underhill, Jr., "Practical solution to the screen light distribution problem," *Jour. SMPTE, 56:* 680–683, June 1951; Report of Progress Committee, *Jour. SMPTE, 56:* 575, May 1951.
146. J. I. Crabtree, "The motion picture laboratory," *Jour. SMPTE, 64:* 13–34, Jan. 1955.
147. H. W. Moyse and D. R. White, "Borax developer characteristics," *Trans. SMPE*, No. 38, 445–452, 1929.
148. H. C. Carlton and J. I. Crabtree, "Some properties of fine grain developers for motion picture film," *Trans. SMPE*, No. 38, 406–444, 1929.
149. E. Huse, "Sensitometric control in the processing of motion picture film in Hollywood," *Jour. SMPE, 21:* 54–82, July 1933.
150. L. A. Jones, "Photographic sensitometry," *Jour. SMPE, 17:* 491–535, Oct. 1931; Part 2, 695–742, Nov. 1931; Part 3, *18:* 54–89, Jan. 1932; Part 4, 324–355, Mar. 1932.
151. A. Küster and R. Schmidt, "The sensitometric control of sound records on film," *Jour. SMPE, 19:* 539–545, Dec. 1932.
152. J. B. Engl, "A new process for developing and printing photographic sound records," *Trans. SMPE*, No. 30, 257–266, 1927.
153. Joe W. Coffman, "Sound film processing," *Trans. SMPE*, No. 35, 799–808, 1928. Also "Art and science in sound film production," *Jour. SMPE, 14:* 172–179, Feb. 1930.
154. W. Leahy, "Time-and-temperature vs. the test system for development of motion picture negatives," *Jour. SMPE, 18:* 649–651, May 1932.
155. J. I. Crabtree and C. E. Ives, "A replenishing solution for a motion picture positive film developer," *Jour. SMPE, 15:* 627–640, Nov. 1930.
156. R. M. Evans, "Maintenance of a developer by continuous replenishment," *Jour. SMPE, 31:* 273–286, Sept. 1938.
157. H. L. Baumbach, "Continuous replenishment and chemical control of motion picture developing solutions," *Jour. SMPE, 39:* 55–66, July 1942.
158. C. E. Ives and E. W. Jensen, "Effect of developer agitation on density uniformity and rate of development," *Jour. SMPE, 40:* 107–136, Feb. 1943.
159. J. I. Crabtree and H. D. Russell, "Some properties of chrome alum stop baths and fixing baths," *Jour. SMPE, 14:* Part 1, 483–512, May 1930; Part 2, 667–700, June 1930.
160. J. I. Crabtree, L. E. Muehler and H. D. Russell, "New stop bath and fixing bath formulas and methods for their revival," *Jour. SMPE, 38:* 353–372, Apr. 1942.
161. R. C. Hubbard, "The straight line developing machine, *Trans. SMPE*, No. 18, 73–85, May 1924.
162. Alfred B. Hitchins, "Machine development of negative and positive motion picture film," *Trans. SMPE*, No. 22, 46–53, 1925.
163. H. D. Hineline, "Continuous photographic processing," *Jour. SMPE, 26:* 38–53, Jan. 1936.
164. Clifton Tuttle and J. W. McFarlane, "The measurement of density in variable-density sound-film," *Jour. SMPE, 15:* 345–351, Sept. 1930.
165. L. A. Jones and J. H. Webb, "Reciprocity law failure in photograph exposures," *Jour. SMPE, 23:* 142–159, Sept. 1934.
166. J. G. Capstaff and R. A. Purdy, "A compact motion picture densitometer," *Trans. SMPE*, No. 31, 607–612, 1927.
167. G. A. Chambers and I. D. Wratten, "The Eastman type IIb densitometer as a control instrument in the processing of motion picture film," *Jour. SMPE, 21:* 218–223, Sept. 1933.
168. F. L. Eich, "A physical densitometer for sound processing laboratories," *Jour. SMPE, 24:* 180–183, Feb. 1935.
169. W. W. Lindsey, Jr. and W. V. Wolfe, "Wide-range, linear-scale photoelectric cell densitometer," *Jour. SMPE, 28:* 622–632, June 1937.
170. D. R. White, "Direct-reading photoelectric densitometer," *Jour. SMPE, 33:* 403–409, Oct. 1939.
171. J. G. Frayne and G. R. Crane, "Precision integrating-sphere densitometer," *Jour. SMPE, 35:* 184–200, Aug. 1940.
172. C. M. Tuttle and M. E. Russell, "Note on the use of an automatic recording densitometer," *Jour. SMPE, 28:* 99–111, Jan. 1937.
172a. D. MacKenzie, "Straight-line and toe records with the light-valve," *Jour. SMPE, 17:* 172–202, Aug. 1931.
173. L. A. Jones and Otto Sandvik, "Photographic characteristics of sound recording film," *Jour. SMPE, 14:* 180–203, Feb. 1930.
174. J. A. Maurer, "The photographic treatment of variable-area sound-films," *Jour. SMPE, 14:* 636–649, June 1930.
175. G. L. Dimmick, "High-frequency response from variable-width records as affected by exposure and development," *Jour. SMPE, 17:* 766–777, Nov. 1931.
176. J. O. Baker and D. H. Robinson, "Modulated high-frequency recording as a means of determining conditions for optimal processing," *Jour. SMPE 30:* 3–17, Jan. 1938.
177. J. E. Abbott, "Development of the sound film," *Jour. SMPE, 38:* 541–545, June 1942.
178. G. L. Dimmick, "Galvanometers for variable-area recording," *Jour. SMPE, 15:* 428–438, Oct. 1930.
179. G. L. Dimmick and H. Belar, "Extension of the frequency range of film recording and reproduction," *Jour. SMPE, 19:* 401–406, Nov. 1932.
180. M. C. Batsel and E. W. Kellogg, "RCA sound recording system," *Jour. SMPE, 28:* 507–533, May 1937.
181. G. L. Dimmick, "A newly developed light modulator for sound recording," *Jour. SMPE, 49:* 48–56, July 1947.
182. E. W. Kellogg, U.S. Pat. 1,740,406.
182a. L. T. Sachtleben, "Characteristics of the Photophone light-modulating system," *Jour. SMPE, 25:* 175–191, Aug. 1935.
183. G. L. Dimmick, "RCA recording system and its adaptation to various types of sound-track," *Jour. SMPE, 29:* 258–273, Sept. 1937.
184. E. W. Kellogg, "ABC of photographic sound recording," *Jour. SMPE, 44:* 151–194, Mar. 1945.
185. E. W. Kellogg, "Ground-noise reduction systems," *Jour. SMPE, 36:* 137–171, Feb. 1941.
186. Siemens and Halske, British Pat. 288,225, Convention Apr. 9, 1927.
187. L. T. Robinson, U.S. Pat. 1,854,159 and 1,935,417; C. W. Hewlett, U.S. Pat. 1,853,812; and C. R. Hanna, U.S. Pat. 1,888,724.
188. H. C. Silent and J. G. Frayne, "Western Electric noiseless recording, *Jour. SMPE, 18:* 551–570, May 1932.
189. H. McDowell, Jr., U.S. Pat. 1,855,197.
190. R. H. Townsend, H. McDowell, Jr. and L. E. Clark, "Ground-noise reduction RCA Photophone system," Reprint 26, *Tech. Bul. Acad. Mot. Pict. Arts & Sci.*, Feb. 1931.
191. E. W. Kellogg and C. N. Batsel, "A shutter for use in reduction of ground-noise," *Jour. SMPE, 17:* 203–215, Aug. 1931.
192. H. J. Hasbrouck, J. O. Baker and C. N. Batsel, "Improved noise-reduction system for high-fidelity recording," *Jour. SMPE, 29:* 310–316, Sept. 1937.
193. R. R. Scoville and W. L. Bell, "Design and use of noise-reduction bias systems," *Jour. SMPE, 38:* 125–147, Feb. 1942.

194. B. Kreuzer, "Noise reduction with variable-area recording," *Jour. SMPE, 16:* 671–683, June 1931.
195. J. L. Hathaway, "Microphone polarity and overmodulation," *Electronics*, Oct. 1939, p. 28.
196. R. O. Drew and E. W. Kellogg, "Starting characteristics of speech sounds," *Jour. SMPE, 34:* 43–58, Jan. 1940; and *J. Acoust. Soc. Am., 12:* 95–103, July 1940.
197. J. G. Frayne, "Noise-reduction anticipation circuits," *Jour. SMPE, 43:* 313–320, Nov. 1944.
198. J. R. Whitney and J. W. Thatcher, "Increased noise reduction by delay networks," *Jour. SMPTE, 54:* 295–302, Mar. 1950.
199. G. L. Dimmick and A. C. Blaney, "Direct positive system of sound recording," *Jour. SMPE, 33:* 479–487, Nov. 1939. See also Progress Committee Report, *Jour. SMPTE, 58:* 401, May 1952.
200. W. A. Mueller and G. R. Groves, "Magnetic recording in the motion picture studio," *Jour. SMPE, 52:* 605–612, June 1949.
201. Academy Research Council, "Report of basic sound committee on pre-and-post-equalization," *Jour. SMPE, 42:* 187–192, Mar. 1944.
202. J. K. Hilliard, "Projects of the committee on standardization of theater sound projection equipment characteristics," *Jour. SMPE, 30:* 81–95, Jan. 1938.
203. J. K. Hilliard, "Variable-density film-recording system used at MGM studios," *Jour. SMPE, 40:* 143–175, Mar. 1943.
204. K. F. Morgan and D. P. Loye, "Sound picture recording and reproduction characteristics," *Jour. SMPE, 32:* 631–647, June 1939; *33:* 107–108, July 1939.
205. F. E. Carlson, "Properties of lamps and optical systems for sound reproduction," *Jour. SMPE, 33:* 80–96, July 1939.
205a. L. T. Sachtleben, U.S. Pat. 2,158,308.
206. O. O. Ceccarini, "Recent contributions to light-valve Technic," *Jour. SMPE, 17:* 305–325, Sept. 1931.
207. W. Herriott and L. V. Foster, "Recent optical improvements in sound-film recording equipment," *Jour. SMPE, 23:* 167–174, Sept. 1934.
208. E. C. Wente and R. Biddulph, "Light-valve for the stereophonic sound-film system," *Jour. SMPE, 37:* 397–405, Oct. 1941.
209. R. R. Scoville, "Overload limiters for the protection of modulating devices," *Jour. SMPE, 31:* 93–98, July 1938.
210. W. C. Jones and L. W. Giles, "A moving-coil microphone for high-quality sound reproduction," *Jour. SMPE, 17:* 977–993, Dec. 1931.
211. E. C. Wente and A. L. Thuras, "Moving coil telephone receivers and microphones," *J. Acoust. Soc. Am., 3:* 44–55, Jan. 1931.
212. H. F. Olson, "The ribbon microphone," *Jour. SMPE, 16:* 695–708, June 1931; and *J. Acoust. Soc. Am., 3:* 56, July 1931.
213. H. F. Olson, "Collection of sound in reverberant rooms, with special reference to the application of the ribbon microphone," *Proc. IRE, 21:* 655, May 1933.
214. H. F. Olson and Frank Massa, *Applied Acoustics*, P. Blakiston's Sons & Co., Philadelphia, 1934.
215. H. F. Olson, "Unidirectional microphone," *Jour. SMPE, 27:* 284–301, Sept. 1936.
216. R. N. Marshall and W. R. Harry, "Cardioid directional microphone," *Jour. SMPE, 33:* 254–277, Sept. 1939.
217. A. G. Webster, "Acoustical impedance and theory of horns and phonograph," *Proc. Nat. Acad. of Sci., 6:* 275, 1919.
218. C. R. Hanna and J. Slepian, "The Function and design of horns for loud speakers," *Trans. AIEE, 43:* 393, Feb. 1924.
219. C. R. Hanna, "Loudspeakers of high efficiency and load capacity," *Trans. AIEE, 47:* 607, Apr. 1928.
220. H. A. Frederick, "Vertical sound records: recent fundamental advances in mechanical records on wax," *Jour. SMPE, 18:* 141–163, Feb. 1932.
221. L. G. Bostwick, "A loud speaker good to twelve thousand cycles," *Jour. SMPE, 16:* 529–534, May 1931.
222. C. Flannagan, R. Wolf and W. C. Jones, "Modern theater loud speakers and their development," *Jour. SMPE, 28:* 246–264, Mar. 1937.
222a. J. P. Maxfield and P. Flannagan, "Wide range reproduction in theaters," *Jour. SMPE, 26:* 67–78, Jan. 1936.
223. Symposium on Auditory Perspective, pub. in *Elec. Eng.*, (AIEE), *53:* 9, Jan. 1934; and *Bell Sys. Tech. J., 13:* 239, Apr. 1934. This equipment is again described in *Jour. SMPE, 37:* 331–417, Oct. 1941. See also refs. 329–332.
224. H. Fletcher, "Transmission and reproduction of speech and music in auditory perspective," *Jour. SMPE, 22:* 314–329, May 1934.
225. J. Frank, Jr., "RCA photophone high-fidelity sound reproducing equipment," *Jour. SMPE, 27:* 99–104, July 1936.
226. J. K. Hilliard, "Study of theater loud speakers and the resultant development of the Shearer two-way horn system," *Jour. SMPE, 27:* 45–60, July 1936.
227. J. K. Hilliard and H. R. Kimball, "Dividing networks for loud speaker systems," *Jour. SMPE, 27:* 61–73, July 1936.
228. H. F. Hopkins and C. R. Keith, "New theater loud speaker systems," *Jour. SMPE, 51:* 385–398, Oct. 1948.
229. J. G. Frayne and B. N. Locanthi, "Theater loudspeaker system incorporating an acoustic-lens radiator," *Jour. SMPTE, 63:* 82–85, Sept. 1954.
230. H. F. Olson and John Preston, "Wide-range loudspeaker developments," *Jour. SMPE, 47:* 327–352, Oct. 1946.
231. James B. Lansing, "New permanent magnet public address loudspeaker," *Jour. SMPE, 46:* 212–219, Mar. 1946.
232. J. Crabtree, "Sound film printing," *Jour. SMPE, 21:* 294–322, Oct. 1933, and "Sound film printing—II," *Jour. SMPE, 22:* 98–114, Feb. 1934. (Repetition of ref. 120.)
233. C. N. Batsel, "A non-slip sound printer," *Jour. SMPE, 23:* 100–107, Aug. 1934.
234. A. V. Bedford, U.S. Pat. 1,754,187.
235. F. W. Roberts and E. Taenzer, "Photographic duping of variable-area sound," *Jour. SMPE, 34:* 26–37, Jan. 1940.
236. R. V. Wood, "A shrinkage-compensating sound printer," *Jour. SMPE, 18:* 788–791, June 1932.
237. G. L. Dimmick, C. N. Batsel and L. T. Sachtleben, "Optical reduction sound printing," *Jour SMPE, 23:* 108–116, Aug. 1934.
238. M. E. Collins, "Optical reduction sound printer," *Jour. SMPE, 27:* 105–106, July 1936.
239. R. O. Drew and L. T. Sachtleben, "Recent laboratory studies of optical reduction printing," *Jour. SMPE, 41:* 505–513, Dec. 1943.
240. John A. Maurer, "Optical sound-track printing," *Jour. SMPE, 50:* 458–473, May 1948.
241. O. Sandvik and J. G. Streiffert, "A continuous optical reduction sound printer," *Jour. SMPE, 25:* 117–126, Aug. 1935.
242. O. Sandvik and G. Silberstein, "The dependence of the resolving power of a photographic material on the wavelength of light," *J. Opt. Soc. Am., 17:* 107, Aug. 1928.
243. Carl L. Oswald and Warren D. Foster, U.S. Pat. 1,928,392, 2,030,760, 2,055,261, 2,213,531.
244. G. L. Dimmick, "Improved resolution in sound recording and printing by the use of ultraviolet light," *Jour. SMPE, 27:* 168–178, Aug. 1936.
245. G. L. Dimmick and L. T. Sachtleben, "Ultraviolet push-pull recording optical system for newsreel cameras," *Jour. SMPE, 31:* 87–92, July 1938.
246. J. O. Baker, "Processing of ultraviolet recordings on panchromatic films," *Jour. SMPE, 31:* 28–35, July 1938.
247. J. G. Frayne and V. Pagliarulo, "Effects of ultraviolet light on variable-density recording and printing," *Jour. SMPE, 34:* 614–631, June 1940.
248. C. R. Daily and I. M. Chambers, "Production and release applications of fine-grain films for variable-density sound recording," *Jour. SMPE, 38:* 45–55, Jan. 1942.
249. J. Strong, "On a method of decreasing reflections from non-metallic substances," *J. Opt. Soc. Am.*, vol. 26, 73, Jan. 1936.
250. W. P. Strickland, "An analysis of low-reflection coatings as applied to glass," *Jour. SMPE, 49:* 27–36, July 1947.
251. C. W. Faulkner and C. N. Batsel, "Operation of the variable-intensity recording system," *Jour. SMPE, 36:* 125–136, Feb. 1941.
252. W. C. Miller, "Speed up your lens systems," *Jour. SMPE, 35:* 3–16, July 1940; and "Recent improvements in nonreflective lens coating," *Jour. SMPE, 37:* 265–273, Sept. 1941.
253. C. H. Cartwright and A. F. Turner, "Multiple films of high reflecting power," *Phys. Rev., 55:* 595, June 1939.
254. G. L. Dimmick, "A new dichroic reflector and its application to photocell monitoring systems," *Jour. SMPE, 38:* 36–44, Jan. 1942.
255. Mary Ellen Widdop, "Review of work on dichroic mirrors and their light-dividing characteristics," *Jour. SMPTE, 60:* 357–366, Apr. 1953.
256. G. L. Dimmick and Mary Ellen Widdop, "Heat-transmitting mirror," *Jour. SMPTE, 58:* 36–42, Jan. 1952.
257. Progress Reports, *Jour. SMPTE, 56:* 575 May 1951; and *Jour. SMPTE, 62:* 336, May 1954.
258. G. J. Koch, "Interference mirrors for arc projectors," *Jour. SMPTE, 55:* 439–442, Oct. 1950.
259. See reference 146 and accompanying classified bibliography. (J. I. Crabtree, "The motion-picture laboratory," *Jour. SMPTE, 64:* 13–34, Jan. 1955.)
260. E. D. Cook, "The aperture alignment effect," *Jour. SMPE, 21:* 390–402, Nov. 1933.
261. D. Foster, "Effect of orientation of the scanning image on the quality of sound reproduced from variable-width records," *Jour. SMPE, 33:* 502–516, Nov. 1939.
262. N. R. Stryker, "Scanning losses in reproduction," *Jour. SMPE, 15:* 610–623, Nov. 1930.
263. E. D. Cook, "The aperture effect," *Jour. SMPE, 14:* 650–662, June 1930.
264. D. Foster, "The effect of exposure and development on the quality of variable-width photographic sound recording," *Jour. SMPE, 17:* 749–765, Nov. 1931.
265. C. N. Batsel and C. H. Cartwright, "Effect of uneven slit illumination upon distortion in several types of variable-width records," *Jour. SMPE, 29:* 476–483, Nov. 1937.

266. American Standard, Scanning Beam Uniformity Test Film for 35mm, *Jour. SMPE, 51:* 542–544, Nov. 1948; and American Standard, *Jour. SMPTE, 59:* 427–431, Nov. 1952.
267. J. O. Baker and R. O. Drew, "New and old aspects of the origins of 96-cycle distortion," *Jour. SMPE, 37:* 227–255, Sept. 1941.
268. J. G. Frayne and V. Pagliarulo, "Influence of sprocket holes upon the development of adjacent sound track areas," *Jour. SMPE, 28:* 235–245, Mar. 1937.
269. J. Crabtree and W. Herriott, "Film perforation and 96-cycle frequency modulation in sound-film records," *Jour. SMPE, 30:* 25–29, Jan. 1938.
270. J. Crabtree, "Directional effects in continuous film processing," *Jour. SMPE, 18:* 207–231, Feb. 1932.
271. J. Crabtree and J. H. Waddell, "Directional effects in sound-film processing-II," *Jour. SMPE, 21:* 351–373, Nov. 1933.
272. C. E. Ives and E. W. Jensen, "Effect of developer agitation on density uniformity and rate of development," *Jour. SMPE, 40:* 107–136, Feb. 1943. (Repetition of ref. 158.)
273. M. S. Leshing, T. M. Ingman and K. Pier, "Reduction of development sprocket-hole modulation" *Jour. SMPE, 36:* 475–487, May 1941.
274. O. Sandvik and V. C. Hall, "Wave form analysis of variable-density sound recording," *Jour. SMPE, 19:* 346–360, Oct. 1932.
275. O. Sandvik, V. C. Hall and J. G. Streiffert, "Wave-form analysis of variable-width sound records," *Jour. SMPE, 21:* 323–336, Oct. 1933.
276. C. E. K. Mees, "Some photographic aspects of sound recording," *Jour. SMPE, 24:* 285–326, Apr. 1935. See page 324.
277. J. G. Frayne and R. R. Scoville, "Analysis and measurement of distortion in variable-density recording," *Jour. SMPE, 32:* 648–673, June 1939.
278. Kurt Singer and Robert V. McKie, "Cross-modulation compensator," *Jour. SMPTE, 63:* 77–81, Sept. 1954.
279. F. G. Albin, "A dynamic check on the processing of film for sound records," *Jour. SMPE, 25:* 161–170, Aug. 1935.
280. C. E. Kenneth Mees, "The history of professional black and white motion picture films," *Jour. SMPTE, 63:* 134–137, Oct. 1954.
281. R. M. Corbin, N. L. Simmons and D. E. Hyndman, "Two new Eastman fine-grain sound recording films," *Jour. SMPE, 45:* 265–285, Oct. 1945.
282. Hollis W. Moyse, "DuPont fine-grain sound films—Types *232* and *236*," *Jour. SMPE, 45:* 285–293, Oct. 1945.
283. L. A. Jones and Otto Sandvik, "Photographic Characteristics of sound recording film," *Jour. SMPE, 14:* 180–203, Feb. 1930. (Repetition of ref. 173.)
284. C. E. Ives and J. I. Crabtree, "Two new films for duplicating work," *Jour. SMPE, 29:* 317–325, Sept. 1937.
285. Progress Report, *Jour. SMPE, 48:* 305, Apr. 1947.
286. C. R. Daily, "Improvement in sound and picture release through the use of fine-grain film," *Jour. SMPE, 34:* 12–25, Jan. 1940.
287. W. H. Offenhauser, Jr., "Notes on the application of fine grain film to 16mm motion pictures," *Jour. SMPE, 41:* 374–388, Nov. 1943.
288. J. R. Wilkinson and F. L. Eich, "Laboratory modification and procedure in connection with fine-grain release printing," *Jour. SMPE, 38:* 56–65, Jan. 1942.
289. Report on the adaptation of fine-grain films to variable-density sound technics, *Jour. SMPE, 34:* 3–11, Jan. 1940.
290. V. C. Shaner, "Note on the processing of Eastman 1302 fine-grain release positive in Hollywood," *Jour. SMPE, 38:* 66–73, Jan. 1942.
291. E. W. Kellogg, "A comparison of variable-density and variable-width systems," *Jour. SMPE, 25:* 203–226, Sept. 1935.
292. John G. Frayne, "Comparison of recording processes," (reprint) *Jour. SMPTE, 59:* 313–318, Oct. 1952.
293. N. Levinson, "New method of increasing the volume range of talking motion pictures," *Jour. SMPE, 26:* 111–116, Feb. 1936.
294. G. L. Dimmick, "Optical control of wave-shape and amplitude characteristics in variable-density recording," *Jour. SMPE, 33:* 650–663, Dec. 1939.
295. John G. Frayne, "Variable-area recording with the light valve," *Jour. SMPE, 51:* 501–520, Nov. 1948.
296. Lewis B. Browder, "Variable-area light-valve modulator," *Jour. SMPE, 51:* 521–533, Nov. 1948.
297. E. M. Honan and C. R. Keith, "Recent developments in sound tracks," *Jour. SMPE, 41:* 127–135, Aug. 1943.
298. J. P. Livadary and S. J. Twining, "Variable-area release from variable-density original sound tracks," *Jour. SMPE, 45:* 380–388, Nov. 1945.
299. W. K. Grimwood, "Volume compressors for sound recording," *Jour. SMPE, 52:* 49–76, Jan. 1949.
300. J. O. Aalberg and J. G. Stewart, "Application of non-linear volume characteristics to dialog recording," *Jour. SMPE, 31:* 248–255, Sept. 1938.
301. B. F. Miller, "Elimination of relative spectral energy distortion in electronic compressors," *Jour. SMPE, 39:* 317–323, Nov. 1942.
302. M. Rettinger and K. Singer, "Factors governing the frequency response of a variable-area film recording channel," *Jour. SMPE, 47:* 299–326, Oct. 1946.
303. Kurt Singer, "High-quality recording electronic mixer," *Jour. SMPE, 52:* 676–683, June 1949.
304. A. C. Norwine, "Devices for controlling amplitude characteristics of telephonic signals," *Bell Sys. Tech. Jour., 17:* 539–554, Oct. 1938.
304a. D. E. Maxwell, "Dynamic performance of peak-limiting amplifiers," *Proc. I.R.E., 35:* 1349–1356, Nov. 1947.
305. Kurt Singer, "Preselection of variable-gain tubes for compressors," *Jour. SMPE, 52:* 684–689, June 1949.
306. Harold E. Haynes, "New principle for electronic volume compression," *Jour. SMPTE, 58:* 137–144, Feb. 1952.
307. W. C. Miller, "Volume control by the squeeze track," *Jour. SMPE, 15:* 53–59, July 1930.
308. E. H. Hansen and C. W. Faulkner, "Mechanical reversed-bias light-valve recording," *Jour. SMPE, 26:* 117–127, Feb. 1936.
309. J. G. Frayne and H. C. Silent, "Push-pull recording with the light-valve," *Jour. SMPE, 31:* 46–64, July 1938.
310. John G. Frayne, T. B. Cunningham and V. Pagliarulo, "An improved 200-mil push-pull density modulator," *Jour. SMPE, 47:* 494–518, Dec. 1946.
311. Lawrence T. Sachtleben, "Wide-track optics for variable-area recorders," *Jour. SMPE, 52:* 89–96, Jan. 1949.
312. J. E. Jenkins and S. E. Adair, "The control-frequency principle," *Jour. SMPE, 22:* 193, Mar. 1934.
313. J. G. Frayne and F. P. Herrnfeld, "Frequency-modulated control-track for movietone prints," *Jour. SMPE, 38:* 111–124, Feb. 1942.
314. G. L. Dimmick and H. Belar, "An improved system for noiseless recording," *Jour. SMPE, 23:* 48–54, July 1934.
315. D. J. Bloomberg and C. L. Lootens, "Class B push-pull recording for original negatives," *Jour. SMPE, 33:* 664–669, Dec. 1939.
316. A. C. Blaney, "Notes on operating experience using the direct positive push-pull method of recording," *Jour. SMPE, 42:* 279–282, May 1944.
317. A. C. Blaney and G. M. Best, "Latest developments in variable-area processing," *Jour. SMPE, 32:* 237–245, Mar. 1939.
318. W. J. Albersheim, "Device for direct reproduction from variable-density sound negatives," *Jour. SMPE, 29:* 274–280, Sept. 1937.
319. O. L. Dupy, "A method of direct-positive variable-density recording with the light valve," *Jour. SMPTE, 59:* 101–106, Aug. 1952.
320. C. R. Keith and V. Pagliarulo, "Direct-positive variable-density recording with the light valve," *Jour. SMPE, 52:* 690–698, June 1949.
321. Lewis B. Browder, "Direct-positive variable-area recording with the light valve," *Jour. SMPE, 53:* 149–158, Aug. 1949.
322. L. I. Carey and Frank Moran, "Push-pull direct-positive recording, an auxiliary to magnetic recording," *Jour. SMPTE, 58:* 67–70, Jan. 1952.
323. John G. Frayne, "Electrical printing," *Jour. SMPTE, 55:* 590–604, Dec. 1950.
324. C. N. Batsel and L. T. Sachtleben, "Some characteristics of 16mm sound by optical reduction and re-recording," *Jour. SMPE, 24:* 95–101, Feb. 1935.
325. W. A. Mueller, "Audience noise as a limitation to the permissible volume range of dialog in sound motion pictures," *Jour. SMPE, 35:* 48–58, July 1940.
326. H. I. Reiskind, "Multiple-speaker reproducing systems for motion pictures," *Jour. SMPE, 37:* 154–163, Aug. 1941.
327. N. Levinson and L. T. Goldsmith, "Vitasound," *Jour. SMPE, 37:* 147–153, Aug. 1941.
328. J. P. Maxfield, "Demonstration of stereophonic recording with motion pictures," *Jour. SMPE, 30:* 131–135, Feb. 1938.
329. H. Fletcher, "Stereophonic reproduction from film," *Jour. SMPE, 34:* 606–613, June 1940; ibid., "Stereophonic sound-film system—general theory," *Jour. SMPE, 37:* 339–352, Oct. 1941. (*See p. 352.*)
330. J. C. Steinberg, "The stereophonic sound-film system—pre- and post-equalization of compandor systems," *Jour. SMPE, 37:* 366–379, Oct. 1941.
331. W. B. Snow and A. R. Soffel, "Electrical equipment for the stereophonic sound-film system," *Jour. SMPE, 37:* 380–396, Oct. 1941.
332. E. C. Wente, R. Biddulph, L. A. Elmer and A. B. Anderson, "Mechanical and optical equipment for the stereophonic sound-film system," *Jour. SMPE, 37:* 353–365, Oct. 1941.
333. W. E. Garity and J. N. A. Hawkins, "Fantasound," *Jour. SMPE, 37:* 127–146, Aug. 1941.
334. W. E. Garity and W. Jones, "Experiences in road-showing Walt Disney's Fantasia," *Jour. SMPE, 39:* 6–15, July 1942.
335. E. H. Plumb, "Future of Fantasound," *Jour. SMPE, 39:* 16–21, July 1942.
336. See group of papers on stereophonic sound, *Jour. SMPTE, 61:* No. 3, Sept. 1953, Pt. II.
337. Panel Discussion, J. K. Hilliard, moderator "Equipment for stereophonic sound reproduction," *Jour. SMPTE, 62:* 228–237, Mar. 1954.
338. L. D. Grignon, "Experiment in stereophonic sound," *Jour. SMPTE, 62:* 64–70, Sept. 1953, Pt. II.
339. Progress Committee Report, *Jour. SMPTE, 62:* 338, May 1954.
340. Progress Committee Report, *Jour. SMPTE, 60:* 535, May 1953. See also Progress Committee Report, *Jour. SMPTE, 64:* 226–228, May 1955.
341. Russell J. Tinkham, "Stereophonic tape recording equipment," *Jour. SMPTE, 62:* 71–74, Jan. 1954.
342. C. Robert Fine, "Perspecta Stereophonic Sound," paper presented at 75th SMPTE Convention, Washington, D. C., May 1954
343. R. Görisch and P. Görlich, "Reproduction

of color film sound records," *Jour. SMPE, 43:* 206–213, Sept. 1944.
344. Alan M. Glover and Arnold R. Moore, "Phototube for dye image sound track," *Jour. SMPE, 46:* 379–386, May 1946.
345. Lloyd T. Goldsmith, "Preliminary report of research council photocell subcommittee," *Jour. SMPE, 48:* 145–147, Feb. 1947.
346. J. L. Forrest, "Metallic-salt track on Ansco 16-mm color film," *Jour. SMPE, 53:* 40–49, July 1949.
347. Report of the SMPTE Color Committee, L. T. Goldsmith, chairman, "Characteristics of color film sound tracks," *Jour. SMPTE, 54:* 377, Mar. 1950.
348. Robert C. Lovick, "Optimum exposure of sound tracks on Kodachrome films," *Jour. SMPTE, 59:* 81–88, Aug. 1952.
349. Robert C. Lovick, "Densitometry of silver sulfide sound tracks," *Jour. SMPTE, 59:* 89–93, Aug. 1952.
350. C. H. Evans and J. F. Finkle, "Sound track on Eastman color print film," *Jour. SMPTE, 57:* 131–139, Aug. 1951.
351. J. A. Maurer, "Present technical status of 16-mm sound-film, *Jour. SMPE, 33:* 315–325, Sept. 1939.
352. W. C. Miller, "Preliminary report of Academy Research Council Committee on rerecording methods for 16-mm release of 35-mm features," *Jour. SMPE, 45:* 135–142, Aug. 1945.
353. E. E. Masterson, F. L. Putzrath and H E. Roys, "Magnetic sound on 16-mm edge-coated film," *Jour. SMPTE, 57:* 559–566, Dec. 1951.
354. Geo. M. Peterson, *The Drive-In Theater*, Associated Publ. Kansas City, Mo. Book Rev., *Jour. SMPTE, 62:* 401, May 1954.
355. R. H. Heacock, "Improved equipment for drive-in theaters," *Jour. SMPTE, 60:* 716–720, June 1953.
356. C. R. Underhill, Jr., "The trend in drive-in theaters," *Jour. SMPTE, 54:* 161–170, Feb. 1950.
357. S. J. Begun, "Recent developments in the field of magnetic recording," *Jour. SMPE, 48:* 1–13, Jan. 1947.
358. G. W. Carpenter and W. L. Carlson, U.S. Pat. 1,640,881, filed Mar. 26, 1921; issued Aug. 30, 1927.
359. M. Camras, "Theoretical response from a magnetic wire record," *Proc. IRE and Waves and Electrons, 34:* No. 8, 597, Aug. 1946.
360. M. Camras, "Magnetic sound for motion pictures," *Jour. SMPE, 48:* 14–28, Jan. 1947.
361. R. Herr, B. F. Murphy and W. W. Wetzel, "Some distinctive properties of magnetic-recording media," *Jour. SMPE, 52:* 77–88, Jan. 1949.
362. H. A. Howell, "Magnetic sound recording on coated paper tape," *Jour. SMPE, 48:* 36–49, Jan. 1947.
363. Hershel Toomin and David Wildfeuer, "The mechanism of supersonic frequencies as applied to magnetic recording," *Proc. IRE, 32:* 664–668, Nov. 1944.
364. Lynn C. Holmes and Donald L. Clark, "Supersonic bias for magnetic recording," *Electronics, 18:* 126, July 1945.
365. G. L. Dimmick and S. W. Johnson, "Optimum high-frequency bias in magnetic recording," *Jour. SMPE, 51:* 489–500, Nov. 1948.
366. Research Council Basic Sound Committee, (Acad. Mot. Pict. Arts and Sciences), Discussion of Magnetic Recording, *Jour. SMPE, 48:* 50–56, Jan. 1947.
367. Earl Masterson, "35-mm magnetic recording system," *Jour. SMPE, 51:* 481–488, Nov. 1948.
368. Dorothy O'Dea, "Magnetic recording for the technician," *Jour. SMPE, 51:* 468–480, Nov. 1948.
369. O. B. Gunby, "Portable magnetic recording system," *Jour. SMPE, 52:* 613–618, June 1949.
370. Progress Committee Report, *Jour. SMPTE, 56:* 572–573, May 1951.
371. C. C. Davis, J. G. Frayne and E. W. Templin, "Multichannel magnetic film recording and reproducing unit," *Jour. SMPTE, 58:* 105–118, Feb. 1952.
372. Loren L. Ryder, "Motion picture studio use of magnetic recording," *Jour. SMPTE, 55:* 605–612, Dec. 1950.
373. Progress Report, *Jour. SMPTE, 58:* 400–401, May 1952.
374. John G. Frayne and John P. Livadary, "Dual photomagnetic intermediate studio recording," *Jour. SMPTE, 59:* 388–397, Nov. 1952.
375. R. H. Ranger, "Sprocketless synchronous magnetic tape," *Jour. SMPTE, 54:* 328–336, Mar. 1950.
376. D. G. C. Hare and W. D. Fling, "Picture-synchronous magnetic tape recording," *Jour. SMPTE, 54:* 554–566, May 1950.
377. George Lewin, "Synchronous 1/4-in. magnetic tape for motion picture production," *Jour. SMPTE, 56:* 664–671, June 1951.
378. Walter T. Selsted, "Synchronous recording on 1/4- in. magnetic tape," *Jour. SMPTE, 55:* 279–284, Sept. 1950.
379. Warren R. Isom, "Synchronized recordings on perforated tape," *Jour. SMPTE, 63:* 26–28, July 1954.
380. Edward Schmidt, "Commercial experiences with magnastripe," *Jour. SMPTE, 60:* 463–469, Apr. 1953. Part II.
381. Andrew H. Persoon, "Magnetic striping of photographic film by laminating process," *Jour. SMPTE, 60:* 485–490, Apr. 1953, Pt. II.
382. Thomas R. Dedell, "Magnetic sound tracks for processed 16-mm motion picture film," *Jour. SMPTE, 60:* 491–500, Apr. 1953, Pt. II.
383. B. L. Kaspin, A. Roberts, Jr., Harry Robbins, and R. L. Powers, "Magnetic striping techniques and characteristics," *Jour. SMPTE, 60:* 470–484, Apr. 1953, Pt. II.
384. Edward Schmidt and Ernest W. Franck, "Manufacture of magnetic recording materials," *Jour. SMPTE, 60:* 453–462, Apr. 1953, Pt. II.
385. Ernest W. Franck, "A study of dropouts in magnetic film," *Jour. SMPTE, 60:* 507–515, Apr. 1953, Pt. II.
386. G. A. del Valle and L. W. Ferber, "Notes on wear of magnetic heads," *Jour. SMPTE, 60:* 501–506, Apr. 1953, Pt. II.
387. J. D. Bick, "Methods of measuring surface induction of magnetic tape," *Jour. SMPTE, 60:* 516–525, Apr. 1953, Pt. II.
388. Ellis W. D'Arcy, "Standardization needs for 16 mm magnetic sound," *Jour. SMPTE, 60:* 526–534, Apr. 1953, Pt. II.
389. W. E. Theisen, "The history of nitrocellulose as a film base," *Jour. SMPE, 20:* 259–262, Mar. 1933.
390. J. M. Calhoun, "The physical properties and dimensional behavior of motion picture film," *Jour. SMPE, 43:* 227–267, Oct. 1944.
391. Charles R. Fordyce, "Improved safety motion picture film support," *Jour. SMPE, 51:* 331–350, Oct. 1948.
392. See Appendix.
393. Report of Historical Committee. *Jour, SMPE, 17:* 655–663, Oct. 1931. Address by Robert A. Millikan and communication by Wm. R. Ballard.
394. Paul Schubert, *The Electric Word, The Rise of Radio;* The MacMillan Co., New York. 1928, pp. 108 to 129.
395. W. Rupert Maclaurin, *Invention and Innovation in the Radio Industry*, The Macmillan Co., New York, 1949, pp. 95–98, 129.
396. Gleason L. Archer, *History of Radio*, publ. by the Am. Hist. Soc. Inc., New York, 1938, pp. 106–108, 114.
397. Irving Langmuir, "The effect of space charge and residual gases on thermionic currents in high vacuum," *Phys. Rev., N.S.* Vol. II: Nol 6, Dec. 1913.
398. Irving Langmuir, "The pure electron discharge and its applications in radio telegraphy and telephony," *Proc. IRE, 3:* 261, 1915.
400. Just and Hanaman, U.S. Pat. 1,108,502, "Incandescent bodies for electric lamps," filed 1905—issued 1912.
401. W. D. Coolidge, "Ductile tungsten," *Trans. AIEE, 29:* Pt. 2, 961–965.
402. C. G. Fink, "Ductile tungsten and molybdenum," *Trans. Electrochem. Soc., 17:* 229, 1910.
403. Irving Langmuir, "Thermal conduction and convection in gases at extremely high temperatures," *Trans. Electrochem. Soc., 20:* 225, 1911.
404. Irving Langmuir, "Convection and conduction of heat in gases," *Phys. Rev., 34:* 401, 1912.
405. Irving Langmuir, "Nitrogen filled lamps," *Gen. Elec. Rev., 16:* 688, 1913.
406. Irving Langmuir and J. A. Orange, "Tungsten lamps of high efficiency," *Trans. AIEE, Pt. II, 32:* p. 1935, 1913.

MONOGRAPH B-1598

BELL TELEPHONE SYSTEM

TECHNICAL PUBLICATIONS

A mathematical theory of communication

by

C. E. Shannon

Published in

BELL SYSTEM TECHNICAL JOURNAL

Vol. 27, pp. 379-423, 623-656, July, October, 1948

Printed in the UNITED STATES OF AMERICA

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.

2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measure entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\log_2 M = \log_{10} M / \log_{10} 2$$

$$= 3.32 \log_{10} M,$$

a decimal digit is about $3\frac{1}{3}$ bits. A digit wheel on a desk computing machine has ten stable positions and therefore has a storage capacity of one decimal digit. In analytical work where integration and differentiation are involved the base e is sometimes useful. The resulting units of information will be called natural units. Change from the base a to base b merely requires multiplication by $\log_b a$.

By a communication system we will mean a system of the type indicated schematically in Fig. 1. It consists of essentially five parts:

1. An *information source* which produces a message or sequence of messages to be communicated to the receiving terminal. The message may be of various types: e.g. (a) A sequence of letters as in a telegraph or teletype system; (b) A single function of time $f(t)$ as in radio or telephony; (c) A function of time and other variables as in black and white television—here the message may be thought of as a function $f(x, y, t)$ of two space coordinates and time, the light intensity at point (x, y) and time t on a pickup tube plate; (d) Two or more functions of time, say $f(t)$, $g(t)$, $h(t)$—this is the case in "three dimensional" sound transmission or if the system is intended to service several individual channels in multiplex; (e) Several functions of

several variables—in color television the message consists of three functions $f(x, y, t)$, $g(x, y, t)$, $h(x, y, t)$ defined in a three-dimensional continuum—we may also think of these three functions as components of a vector field defined in the region—similarly, several black and white television sources would produce "messages" consisting of a number of functions of three variables; (f) Various combinations also occur, for example in television with an associated audio channel.

2. A *transmitter* which operates on the message in some way to produce a signal suitable for transmission over the channel. In telephony this operation consists merely of changing sound pressure into a proportional electrical current. In telegraphy we have an encoding operation which produces a sequence of dots, dashes and spaces on the channel corresponding to the message. In a multiplex PCM system the different speech functions must be sampled, compressed, quantized and encoded, and finally interleaved

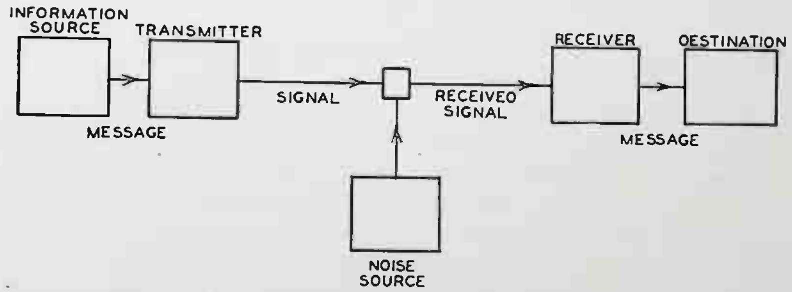


Fig. 1—Schematic diagram of a general communication system.

properly to construct the signal. Vocoder systems, television, and frequency modulation are other examples of complex operations applied to the message to obtain the signal.

3. The *channel* is merely the medium used to transmit the signal from transmitter to receiver. It may be a pair of wires, a coaxial cable, a band of radio frequencies, a beam of light, etc.

4. The *receiver* ordinarily performs the inverse operation of that done by the transmitter, reconstructing the message from the signal.

5. The *destination* is the person (or thing) for whom the message is intended.

We wish to consider certain general problems involving communication systems. To do this it is first necessary to represent the various elements involved as mathematical entities, suitably idealized from their physical counterparts. We may roughly classify communication systems into three main categories: discrete, continuous and mixed. By a discrete system we will mean one in which both the message and the signal are a sequence of discrete symbols. A typical case is telegraphy where the message is a sequence of letters and the signal a sequence of dots, dashes and spaces. A continuous system is one in which the message and signal are both treated as continuous functions, e.g. radio or television. A mixed system is one in which both discrete and continuous variables appear, e.g., PCM transmission of speech.

We first consider the discrete case. This case has applications not only in communication theory, but also in the theory of computing machines, the design of telephone exchanges and other fields. In addition the discrete case forms a foundation for the continuous and mixed cases which will be treated in the second half of the paper.

PART I: DISCRETE NOISELESS SYSTEMS

1. The Discrete Noiseless Channel

Teletype and telegraphy are two simple examples of a discrete channel for transmitting information. Generally, a discrete channel will mean a system whereby a sequence of choices from a finite set of elementary symbols $S_1 \cdots S_n$ can be transmitted from one point to another. Each of the symbols S_i is assumed to have a certain duration in time t_i seconds (not necessarily the same for different S_i, for example the dots and dashes in telegraphy). It is not required that all possible sequences of the S_i be capable of transmission on the system; certain sequences only may be allowed. These will be possible signals for the channel. Thus in telegraphy suppose the symbols are: (1) A dot, consisting of line closure for a unit of time and then line open for a unit of time; (2) A dash, consisting of three time units of closure and one unit open; (3) A letter space consisting of, say, three units of line open; (4) A word space of six units of line open. We might place the restriction on allowable sequences that no spaces follow each other (for if two letter spaces are adjacent, it is identical with a word space). The question we now consider is how one can measure the capacity of such a channel to transmit information.

In the teletype case where all symbols are of the same duration, and any sequence of the 32 symbols is allowed the answer is easy. Each symbol represents five bits of information. If the system transmits n symbols per second it is natural to say that the channel has a capacity of $5n$ bits per second. This does not mean that the teletype channel will always be transmitting information at this rate—this is the maximum possible rate and whether or not the actual rate reaches this maximum depends on the source of information which feeds the channel, as will appear later.

In the more general case with different lengths of symbols and constraints on the allowed sequences, we make the following definition:

Definition: The capacity C of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration T.

It is easily seen that in the teletype case this reduces to the previous result. It can be shown that the limit in question will exist as a finite number in most cases of interest. Suppose all sequences of the symbols $S_1, \cdots, S_n$ are allowed and these symbols have durations $t_1, \cdots, t_n$. What is the channel capacity? If $N(t)$ represents the number of sequences of duration t we have

$$N(t) = N(t - t_1) + N(t - t_2) + \cdots + N(t - t_n)$$

The total number is equal to the sum of the numbers of sequences ending in $S_1, S_2, \cdots, S_n$ and these are $N(t - t_1), N(t - t_2), \cdots, N(t - t_n)$, respectively. According to a well known result in finite differences, $N(t)$ is then asymptotic for large t to X_0^t where X_0 is the largest real solution of the characteristic equation:

$$X^{-t_1} + X^{-t_2} + \cdots + X^{-t_n} = 1$$

and therefore

$$C = \log X_0$$

In case there are restrictions on allowed sequences we may still often obtain a difference equation of this type and find C from the characteristic equation. In the telegraphy case mentioned above

$$N(t) = N(t - 2) + N(t - 4) + N(t - 5) + N(t - 7) + N(t - 8) + N(t - 10)$$

as we see by counting sequences of symbols according to the last or next to the last symbol occurring. Hence C is $-\log \mu_0$ where μ_0 is the positive root of $1 = \mu^2 + \mu^4 + \mu^5 + \mu^7 + \mu^8 + \mu^{10}$. Solving this we find $C = 0.539$.

A very general type of restriction which may be placed on allowed sequences is the following: We imagine a number of possible states $a_1, a_2, \cdots, a_m$. For each state only certain symbols from the set $S_1, \cdots, S_n$ can be transmitted (different subsets for the different states). When one of these has been transmitted the state changes to a new state depending both on the old state and the particular symbol transmitted. The telegraph case is a simple example of this. There are two states depending on whether or not a space was the last symbol transmitted. If so then only a dot or a dash can be sent next and the state always changes. If not, any symbol can be transmitted and the state changes if a space is sent, otherwise it remains the same. The conditions can be indicated in a linear graph as shown in Fig. 2. The junction points correspond to the states and the lines indicate the symbols possible in a state and the resulting state. In Appendix I it is shown that if the conditions on allowed sequences can be described in this form C will exist and can be calculated in accordance with the following result:

Theorem 1: Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to log W where W is the largest real root of the determinant equation:

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0.$$

where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise.

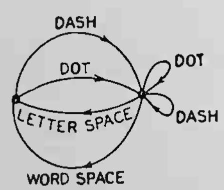


Fig. 2—Graphical representation of the constraints on telegraph symbols.

For example, in the telegraph case (Fig. 2) the determinant is:

$$\begin{vmatrix} -1 & (W^{-2} + W^{-4}) \\ (W^{-3} + W^{-6}) & (W^{-2} + W^{-4} - 1) \end{vmatrix} = 0$$

On expansion this leads to the equation given above for this case.

2. The Discrete Source of Information

We have seen that under very general conditions the logarithm of the number of possible signals in a discrete channel increases linearly with time. The capacity to transmit information can be specified by giving this rate of increase, the number of bits per second required to specify the particular signal used.

We now consider the information source. How is an information source to be described mathematically, and how much information in bits per second is produced in a given source? The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity

of the channel, by the use of proper encoding of the information. In telegraphy, for example, the messages to be transmitted consist of sequences of letters. These sequences, however, are not completely random. In general, they form sentences and have the statistical structure of, say, English. The letter E occurs more frequently than Q, the sequence TH more frequently than XP, etc. The existence of this structure allows one to make a saving in time (or channel capacity) by properly encoding the message sequences into signal sequences. This is already done to a limited extent in telegraphy by using the shortest channel symbol, a dot, for the most common English letter E; while the infrequent letters, Q, X, Z are represented by longer sequences of dots and dashes. This idea is carried still further in certain commercial codes where common words and phrases are represented by four- or five-letter code groups with a considerable saving in average time. The standardized greeting and anniversary telegrams now in use extend this to the point of encoding a sentence or two into a relatively short sequence of numbers.

We can think of a discrete source as generating the message, symbol by symbol. It will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities is known as a stochastic process.[3] We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as:

1. Natural written languages such as English, German, Chinese.
2. Continuous information sources that have been rendered discrete by some quantizing process. For example, the quantized speech from a PCM transmitter, or a quantized television signal.
3. Mathematical cases where we merely define abstractly a stochastic process which generates a sequence of symbols. The following are examples of this last type of source.
   (A) Suppose we have five letters A, B, C, D, E which are chosen each with probability .2, successive choices being independent. This would lead to a sequence of which the following is a typical example.
   B D C B C E C C C A D C B D D A A E C E E A
   A B B D A E E C A C E E B A E E C B C E A D
   This was constructed with the use of a table of random numbers.[4]

[3] See, for example, S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Reviews of Modern Physics*, v. 15, No. 1, January 1943, p. 1.
[4] Kendall and Smith, "Tables of Random Sampling Numbers," Cambridge, 1939.

a space was the last symbol transmitted. If so then only a dot or a dash can be sent next and the state always changes. If not, any symbol can be transmitted and the state changes if a space is sent, otherwise it remains the same. The conditions can be indicated in a linear graph as shown in Fig. 2. The junction points correspond to the states and the lines indicate the symbols possible in a state and the resulting state. In Appendix I it is shown that if the conditions on allowed sequences can be described in this form C will exist and can be calculated in accordance with the following result:

Theorem 1: Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to log W where W is the largest real root of the determinant equation:

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0.$$

where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise.

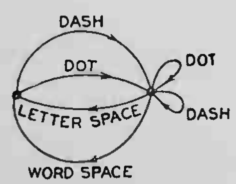


Fig. 2—Graphical representation of the constraints on telegraph symbols.

For example, in the telegraph case (Fig. 2) the determinant is:

$$\begin{vmatrix} -1 & (W^{-2} + W^{-4}) \\ (W^{-3} + W^{-6}) & (W^{-2} + W^{-4} - 1) \end{vmatrix} = 0$$

On expansion this leads to the equation given above for this case.

2. The Discrete Source of Information

We have seen that under very general conditions the logarithm of the number of possible signals in a discrete channel increases linearly with time. The capacity to transmit information can be specified by giving this rate of increase, the number of bits per second required to specify the particular signal used.

We now consider the information source. How is an information source to be described mathematically, and how much information in bits per second is produced in a given source? The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity

of the channel, by the use of proper encoding of the information. In telegraphy, for example, the messages to be transmitted consist of sequences of letters. These sequences, however, are not completely random. In general, they form sentences and have the statistical structure of, say, English. The letter E occurs more frequently than Q, the sequence TH more frequently than XP, etc. The existence of this structure allows one to make a saving in time (or channel capacity) by properly encoding the message sequences into signal sequences. This is already done to a limited extent in telegraphy by using the shortest channel symbol, a dot, for the most common English letter E; while the infrequent letters, Q, X, Z are represented by longer sequences of dots and dashes. This idea is carried still further in certain commercial codes where common words and phrases are represented by four- or five-letter code groups with a considerable saving in average time. The standardized greeting and anniversary telegrams now in use extend this to the point of encoding a sentence or two into a relatively short sequence of numbers.

We can think of a discrete source as generating the message, symbol by symbol. It will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities is known as a stochastic process.[3] We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as:

1. Natural written languages such as English, German, Chinese.
2. Continuous information sources that have been rendered discrete by some quantizing process. For example, the quantized speech from a PCM transmitter, or a quantized television signal.
3. Mathematical cases where we merely define abstractly a stochastic process which generates a sequence of symbols. The following are examples of this last type of source.
   (A) Suppose we have five letters A, B, C, D, E which are chosen each with probability .2, successive choices being independent. This would lead to a sequence of which the following is a typical example.
   B D C B C E C C C A D C B D D A A E C E E A
   A B B D A E E C A C E E B A E E C B C E A D
   This was constructed with the use of a table of random numbers.[4]

[3] See, for example, S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Reviews of Modern Physics*, v. 15, No. 1, January 1943, p. 1.
[4] Kendall and Smith, "Tables of Random Sampling Numbers," Cambridge, 1939.

(B) Using the same five letters let the probabilities be .4, .1, .2, .2, .1 respectively, with successive choices independent. A typical message from this source is then:

A A A C D C B D C E A A D A D A C E D A
E A D C A B E D A D D C E C A A A A A D

(C) A more complicated structure is obtained if successive symbols are not chosen independently but their probabilities depend on preceding letters. In the simplest case of this type a choice depends only on the preceding letter and not on ones before that. The statistical structure can then be described by a set of transition probabilities $p_i(j)$, the probability that letter i is followed by letter j. The indices i and j range over all the possible symbols. A second equivalent way of specifying the structure is to give the "digram" probabilities $p(i, j)$, i.e., the relative frequency of the digram $i\,j$. The letter frequencies $p(i)$, (the probability of letter i), the transition probabilities $p_i(j)$ and the digram probabilities $p(i, j)$ are related by the following formulas.

$$p(i) = \sum_j p(i, j) = \sum_j p(j, i) = \sum_j p(j) p_j(i)$$

$$p(i, j) = p(i) p_i(j)$$

$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i,j} p(i, j) = 1.$$

As a specific example suppose there are three letters A, B, C with the probability tables:

$p_i(j)$		j: A	B	C
	A	0	$\frac{4}{5}$	$\frac{1}{5}$
i	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

i	$p(i)$
A	$\frac{9}{27}$
B	$\frac{16}{27}$
C	$\frac{2}{27}$

$p(i, j)$		j: A	B	C
	A	0	$\frac{4}{15}$	$\frac{1}{15}$
i	B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

A typical message from this source is the following:

A B B A B A B A B A B A B A B B B A B B B B B A B
A B A B A B A B B B A C A C A B B A B B B B A B B
A B A C B B B A B A

The next increase in complexity would involve trigram frequencies but no more. The choice of a letter would depend on the preceding two letters but not on the message before that point. A set of trigram frequencies $p(i, j, k)$ or equivalently a set of transition probabilities $p_{ij}(k)$ would be required. Continuing in this way one obtains successively more complicated stochastic processes. In the general n-gram case a set of n-gram probabilities $p(i_1, i_2, \cdots, i_n)$ or of transition probabilities $p_{i_1, i_2, \cdots, i_{n-1}}(i_n)$ is required to specify the statistical structure.

(D) Stochastic processes can also be defined which produce a text consisting of a sequence of "words." Suppose there are five letters A, B, C, D, E and 16 "words" in the language with associated probabilities:

.10 A	.16 BEBE	.11 CABED	.04 DEB
.04 ADEB	.04 BED	.05 CEED	.15 DEED
.05 ADEE	.02 BEED	.08 DAB	.01 EAB
.01 BADD	.05 CA	.04 DAD	.05 EE

Suppose successive "words" are chosen independently and are separated by a space. A typical message might be:

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE
BEBE BEBE ADEE BED DEED DEED CEED ADEE A DEED
DEED BEBE CABED BEBE BED DAB DEED ADEB

If all the words are of finite length this process is equivalent to one of the preceding type, but the description may be simpler in terms of the word structure and probabilities. We may also generalize here and introduce transition probabilities between words, etc.

These artificial languages are useful in constructing simple problems and examples to illustrate various possibilities. We can also approximate to a natural language by means of a series of simple artificial languages. The zero-order approximation is obtained by choosing all letters with the same probability and independently. The first-order approximation is obtained by choosing successive letters independently but each letter having the same probability that it does in the natural language.[5] Thus, in the first-order approximation to English, E is chosen with probability .12 (its frequency in normal English) and W with probability .02, but there is no influence between adjacent letters and no tendency to form the preferred digrams such as *TH*, *ED*, etc. In the second-order approximation, digram structure is introduced. After a letter is chosen, the next one is chosen in accordance with the frequencies with which the various letters follow the first one. This requires a table of digram frequencies $p_i(j)$. In the third-order approximation, trigram structure is introduced. Each letter is chosen with probabilities which depend on the preceding two letters.

[5] Letter, digram and trigram frequencies are given in "Secret and Urgent" by Fletcher Pratt, Blue Ribbon Books 1939. Word frequencies are tabulated in "Relative Frequency of English Speech Sounds," G. Dewey, Harvard University Press, 1923.

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equi-probable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
   FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
   ALHENHTTPA OOBTTVA NAH BRL

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
   ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
   TIZIN ANDY TOBE SEACE CTISBE

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
   PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
   REGOACTIONA OF CRE

5. First-Order Word Approximation. Rather than continue with tetragram, $\cdots$, n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-Order Word Approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction. Thus in (3) the statistical process insures reasonable text for two-letter sequence, but four-letter sequences from the sample can usually be fitted into good sentences. In (6) sequences of four or more

words can easily be placed in sentences without unusual or strained constructions. The particular sequence of ten words "attack on an English writer that the character of this" is not at all unreasonable. It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source.

The first two samples were constructed by the use of a book of random numbers in conjunction with (for example 2) a table of letter frequencies. This method might have been continued for (3), (4), and (5), since digram, trigram, and word frequency tables are available, but a simpler equivalent method was used. To construct (3) for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5), and (6). It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

4. Graphical Representation of a Markoff Process

Stochastic processes of the type described above are known mathematically as discrete Markoff processes and have been extensively studied in the literature.[6] The general case can be described as follows: There exist a finite number of possible "states" of a system; $S_1, S_2, \cdots, S_n$. In addition there is a set of transition probabilities; $p_i(j)$ the probability that if the system is in state S_i it will next go to state S_j. To make this Markoff process into an information source we need only assume that a letter is produced for each transition from one state to another. The states will correspond to the "residue of influence" from preceding letters.

The situation can be represented graphically as shown in Figs. 3, 4 and 5. The "states" are the junction points in the graph and the probabilities and letters produced for a transition are given beside the corresponding line. Figure 3 is for the example B in Section 2, while Fig. 4 corresponds to the example C. In Fig. 3 there is only one state since successive letters are independent. In Fig. 4 there are as many states as letters. If a trigram example were constructed there would be at most n^2 states corresponding to the possible pairs of letters preceding the one being chosen. Figure 5 is a graph for the case of word structure in example D. Here S corresponds to the "space" symbol.

[6] For a detailed treatment see M. Frechet, "Methods des fonctions arbitraires. Theorie des énénements en chaine dans le cas d'un nombre fini d'états possibles." Paris, Gauthier-Villars, 1938.

5. Ergodic and Mixed Sources

As we have indicated above a discrete source for our purposes can be considered to be represented by a Markoff process. Among the possible discrete Markoff processes there is a group with special properties of significance in

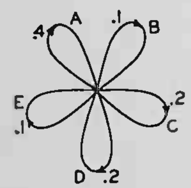


Fig. 3—A graph corresponding to the source in example B.

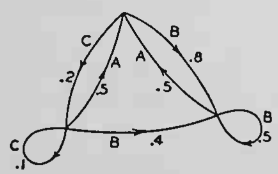


Fig. 4—A graph corresponding to the source in example C.

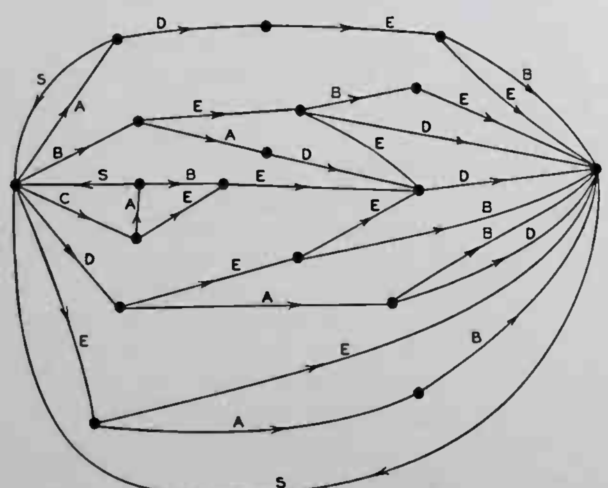


Fig. 5—A graph corresponding to the source in example D.

communication theory. This special class consists of the "ergodic" processes and we shall call the corresponding sources ergodic sources. Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple. In an ergodic process every sequence produced by the process is the same in statistical properties. Thus the letter frequencies, digram frequencies, etc., obtained from particular sequences will, as the lengths of the sequences increase, approach definite limits independent of the particular sequence. Actually this is not true of every sequence but the set for which it is false has probability zero. Roughly the ergodic property means statistical homogeneity.

All the examples of artificial languages given above are ergodic. This property is related to the structure of the corresponding graph. If the graph has the following two properties[7] the corresponding process will be ergodic:

1. The graph does not consist of two isolated parts A and B such that it is impossible to go from junction points in part A to junction points in part B along lines of the graph in the direction of arrows and also impossible to go from junctions in part B to junctions in part A.
2. A closed series of lines in the graph with all arrows on the lines pointing in the same orientation will be called a "circuit." The "length" of a circuit is the number of lines in it. Thus in Fig. 5 the series BEBES is a circuit of length 5. The second property required is that the greatest common divisor of the lengths of all circuits in the graph be one.

If the first condition is satisfied but the second one violated by having the greatest common divisor equal to $d > 1$, the sequences have a certain type of periodic structure. The various sequences fall into d different classes which are statistically the same apart from a shift of the origin (i.e., which letter in the sequence is called letter 1). By a shift of from 0 up to $d - 1$ any sequence can be made statistically equivalent to any other. A simple example with $d = 2$ is the following: There are three possible letters a, b, c. Letter a is followed with either b or c with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Either b or c is always followed by letter a. Thus a typical sequence is

$$a\,b\,a\,c\,a\,c\,a\,c\,a\,b\,a\,c\,a\,b\,a\,b\,a\,c\,a\,c$$

This type of situation is not of much importance for our work.

If the first condition is violated the graph may be separated into a set of subgraphs each of which satisfies the first condition. We will assume that the second condition is also satisfied for each subgraph. We have in this case what may be called a "mixed" source made up of a number of pure components. The components correspond to the various subgraphs. If L_1, L_2, L_3, $\cdots$ are the component sources we may write

$$L = p_1L_1 + p_2L_2 + p_3L_3 + \cdots$$

where p_i is the probability of the component source L_i.

[7] These are restatements in terms of the graph of conditions given in Frechet.

Physically the situation represented is this: There are several different sources $L_1, L_2, L_3, \cdots$ which are each of homogeneous statistical structure (i.e., they are ergodic). We do not know *a priori* which is to be used, but once the sequence starts in a given pure component L_i it continues indefinitely according to the statistical structure of that component.

As an example one may take two of the processes defined above and assume $p_1 = .2$ and $p_2 = .8$. A sequence from the mixed source

$$L = .2\,L_1 + .8\,L_2$$

would be obtained by choosing first L_1 or L_2 with probabilities .2 and .8 and after this choice generating a sequence from whichever was chosen.

Except when the contrary is stated we shall assume a source to be ergodic. This assumption enables one to identify averages along a sequence with averages over the ensemble of possible sequences (the probability of a discrepancy being zero). For example the relative frequency of the letter A in a particular infinite sequence will be, with probability one, equal to its relative frequency in the ensemble of sequences.

If P_i is the probability of state i and $p_i(j)$ the transition probability to state j, then for the process to be stationary it is clear that the P_i must satisfy equilibrium conditions:

$$P_j = \sum_i P_i p_i(j).$$

In the ergodic case it can be shown that with any starting conditions the probabilities $P_j(N)$ of being in state j after N symbols, approach the equilibrium values as $N \to \infty$.

6. Choice, Uncertainty and Entropy

We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \cdots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \cdots, p_n)$, it is reasonable to require of it the following properties:

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}H(\tfrac{2}{3}, \tfrac{1}{3})$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

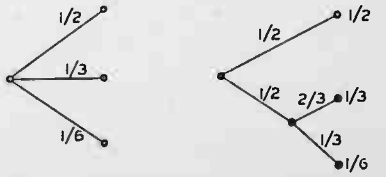


Fig. 6—Decomposition of a choice from three possibilities.

In Appendix II, the following result is established:

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

Quantities of the form $H = -\Sigma\, p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[8] where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\Sigma\, p_i \log p_i$ the entropy of the set of probabilities

[8] See, for example, R. C. Tolman, "Principles of Statistical Mechanics," Oxford, Clarendon, 1938.

Physically the situation represented is this: There are several different sources $L_1, L_2, L_3, \cdots$ which are each of homogeneous statistical structure (i.e., they are ergodic). We do not know *a priori* which is to be used, but once the sequence starts in a given pure component L_i it continues indefinitely according to the statistical structure of that component.

As an example one may take two of the processes defined above and assume $p_1 = .2$ and $p_2 = .8$. A sequence from the mixed source

$$L = .2\,L_1 + .8\,L_2$$

would be obtained by choosing first L_1 or L_2 with probabilities .2 and .8 and after this choice generating a sequence from whichever was chosen.

Except when the contrary is stated we shall assume a source to be ergodic. This assumption enables one to identify averages along a sequence with averages over the ensemble of possible sequences (the probability of a discrepancy being zero). For example the relative frequency of the letter A in a particular infinite sequence will be, with probability one, equal to its relative frequency in the ensemble of sequences.

If P_i is the probability of state i and $p_i(j)$ the transition probability to state j, then for the process to be stationary it is clear that the P_i must satisfy equilibrium conditions:

$$P_j = \sum_i P_i p_i(j).$$

In the ergodic case it can be shown that with any starting conditions the probabilities $P_j(N)$ of being in state j after N symbols, approach the equilibrium values as $N \to \infty$.

6. Choice, Uncertainty and Entropy

We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \cdots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \cdots, p_n)$, it is reasonable to require of it the following properties:

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing

function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}, \frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}H(\tfrac{2}{3}, \tfrac{1}{3})$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

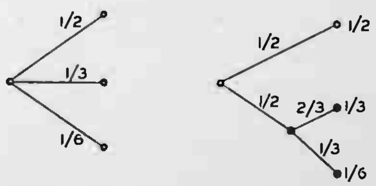


Fig. 6—Decomposition of a choice from three possibilities.

In Appendix II, the following result is established:

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

Quantities of the form $H = -\Sigma\, p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[8] where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\Sigma\, p_i \log p_i$ the entropy of the set of probabilities

[8] See, for example, R. C. Tolman, "Principles of Statistical Mechanics," Oxford, Clarendon, 1938.

$p_1, \cdots, p_n$. If x is a chance variable we will write $H(x)$ for its entropy; thus x is not an argument of a function but a label for a number, to differentiate it from $H(y)$ say, the entropy of the chance variable y.

The entropy in the case of two possibilities with probabilities p and $q = 1 - p$, namely

$$H = -(p \log p + q \log q)$$

is plotted in Fig. 7 as a function of p.

The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.

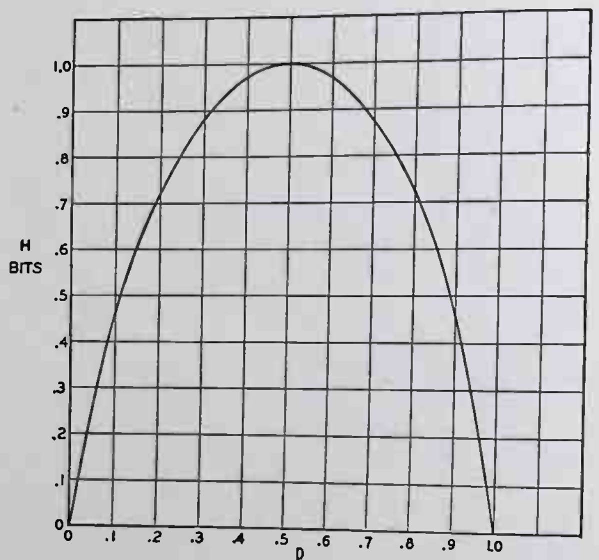


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1 - p)$.

1. $H = 0$ if and only if all the p_i but one are zero, this one having the value unity. Thus only when we are certain of the outcome does H vanish. Otherwise H is positive.

2. For a given n, H is a maximum and equal to $\log n$ when all the p_i are equal $\left(\text{i.e., } \frac{1}{n}\right)$. This is also intuitively the most uncertain situation.

3. Suppose there are two events, x and y, in question with m possibilities for the first and n for the second. Let $p(i, j)$ be the probability of the joint occurrence of i for the first and j for the second. The entropy of the joint event is

$$H(x, y) = -\sum_{i,j} p(i, j) \log p(i, j)$$

while

$$H(x) = -\sum_{i,j} p(i, j) \log \sum_j p(i, j)$$

$$H(y) = -\sum_{i,j} p(i, j) \log \sum_i p(i, j).$$

It is easily shown that

$$H(x, y) \le H(x) + H(y)$$

with equality only if the events are independent (i.e., $p(i, j) = p(i)\, p(j)$). The uncertainty of a joint event is less than or equal to the sum of the individual uncertainties.

4. Any change toward equalization of the probabilities $p_1, p_2, \cdots, p_n$ increases H. Thus if $p_1 < p_2$ and we increase p_1, decreasing p_2 an equal amount so that p_1 and p_2 are more nearly equal, then H increases. More generally, if we perform any "averaging" operation on the p_i of the form

$$p'_i = \sum_j a_{ij} p_j$$

where $\sum_i a_{ij} = \sum_j a_{ij} = 1$, and all $a_{ij} \ge 0$, then H increases (except in the special case where this transformation amounts to no more than a permutation of the p_j with H of course remaining the same).

5. Suppose there are two chance events x and y as in 3, not necessarily independent. For any particular value i that x can assume there is a conditional probability $p_i(j)$ that y has the value j. This is given by

$$p_i(j) = \frac{p(i, j)}{\sum_j p(i, j)}.$$

We define the *conditional entropy* of y, $H_x(y)$ as the average of the entropy of y for each value of x, weighted according to the probability of getting that particular x. That is

$$H_x(y) = -\sum_{i,j} p(i, j) \log p_i(j).$$

This quantity measures how uncertain we are of y on the average when we know x. Substituting the value of $p_i(j)$ we obtain

$$H_x(y) = -\sum_{ij} p(i, j) \log p(i, j) + \sum_{ij} p(i, j) \log \sum_j p(i, j)$$

$$= H(x, y) - H(x)$$

or

$$H(x, y) = H(x) + H_x(y)$$

The uncertainty (or entropy) of the joint event x, y is the uncertainty of x plus the uncertainty of y when x is known.

6. From 3 and 5 we have

$$H(x) + H(y) \geq H(x, y) = H(x) + H_x(y)$$

Hence

$$H(y) \geq H_x(y)$$

The uncertainty of y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed.

7. The Entropy of an Information Source

Consider a discrete source of the finite state type considered above. For each possible state i there will be a set of probabilities $p_i(j)$ of producing the various possible symbols j. Thus there is an entropy H_i for each state. The entropy of the source will be defined as the average of these H_i weighted in accordance with the probability of occurrence of the states in question:

$$H = \sum_i P_i H_i$$

$$= -\sum_{i,j} P_i p_i(j) \log p_i(j)$$

This is the entropy of the source per symbol of text. If the Markoff process is proceeding at a definite time rate there is also an entropy per second

$$H' = \sum_i f_i H_i$$

where f_i is the average frequency (occurrences per second) of state i. Clearly

$$H' = mH$$

where m is the average number of symbols produced per second. H or H' measures the amount of information generated by the source per symbol or per second. If the logarithmic base is 2, they will represent bits per symbol or per second.

If successive symbols are independent then H is simply $-\Sigma\, p_i \log p_i$ where p_i is the probability of symbol i. Suppose in this case we consider a long message of N symbols. It will contain with high probability about p_1N occurrences of the first symbol, p_2N occurrences of the second, etc. Hence the probability of this particular message will be roughly

$$p = p_1^{p_1N} p_2^{p_2N} \cdots p_n^{p_nN}$$

or

$$\log p \doteq N \sum_i p_i \log p_i$$

$$\log p \doteq -NH$$

$$H \doteq \frac{\log 1/p}{N}.$$

H is thus approximately the logarithm of the reciprocal probability of a typical long sequence divided by the number of symbols in the sequence. The same result holds for any source. Stated more precisely we have (see Appendix III):

Theorem 3: Given any $\epsilon > 0$ and $\delta > 0$, we can find an N_0 such that the sequences of any length $N \geq N_0$ fall into two classes:

1. A set whose total probability is less than ϵ.
2. The remainder, all of whose members have probabilities satisfying the inequality

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta$$

In other words we are almost certain to have $\frac{\log p^{-1}}{N}$ very close to H when N is large.

A closely related result deals with the number of sequences of various probabilities. Consider again the sequences of length N and let them be arranged in order of decreasing probability. We define $n(q)$ to be the number we must take from this set starting with the most probable one in order to accumulate a total probability q for those taken.

Theorem 4:

$$\lim_{N\to\infty} \frac{\log n(q)}{N} = H$$

when q does not equal 0 or 1.

We may interpret $\log n(q)$ as the number of bits required to specify the sequence when we consider only the most probable sequences with a total probability q. Then $\frac{\log n(q)}{N}$ is the number of bits per symbol for the specification. The theorem says that for large N this will be independent of q and equal to H. The rate of growth of the logarithm of the number of reasonably probable sequences is given by H, regardless of our interpretation of "reasonably probable." Due to these results, which are proved in appendix III, it is possible for most purposes to treat the long sequences as though there were just 2^{HN} of them, each with a probability 2^{-HN}.

The next two theorems show that H and H' can be determined by limiting operations directly from the statistics of the message sequences, without reference to the states and transition probabilities between states.

Theorem 5: Let $p(B_i)$ be the probability of a sequence B_i of symbols from the source. Let

$$G_N = -\frac{1}{N}\sum_i p(B_i)\log p(B_i)$$

where the sum is over all sequences B_i containing N symbols. Then G_N is a monotonic decreasing function of N and

$$\lim_{N\to\infty} G_N = H.$$

Theorem 6: Let $p(B_i, S_j)$ be the probability of sequence B_i followed by symbol S_j and $p_{B_i}(S_j) = p(B_i, S_j)/p(B_i)$ be the conditional probability of S_j after B_i. Let

$$F_N = -\sum_{i,j} p(B_i, S_j)\log p_{B_i}(S_j)$$

where the sum is over all blocks B_i of $N - 1$ symbols and over all symbols S_j. Then F_N is a monotonic decreasing function of N,

$$F_N = NG_N - (N-1)\,G_{N-1},$$

$$G_N = \frac{1}{N}\sum_1^n F_N,$$

$$F_N \leq G_N,$$

and $\lim_{N\to\infty} F_N = H.$

These results are derived in appendix III. They show that a series of approximations to H can be obtained by considering only the statistical structure of the sequences extending over 1, 2, $\cdots$ N symbols. F_N is the better approximation. In fact F_N is the entropy of the N^{th} order approximation to the source of the type discussed above. If there are no statistical influences extending over more than N symbols, that is if the conditional probability of the next symbol knowing the preceding $(N - 1)$ is not changed by a knowledge of any before that, then $F_N = H$. F_N of course is the conditional entropy of the next symbol when the $(N - 1)$ preceding ones are known, while G_N is the entropy per symbol of blocks of N symbols.

The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its *relative entropy*. This is the maximum compression possible when we encode into the same alphabet. One minus the relative entropy is the *redundancy*. The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters is roughly 50%. This means that when we write English half of what we write is determined by the structure of the language and half is chosen freely. The figure 50% was found by several independent methods which all gave results in this neighborhood. One is by calculation of the entropy of the approximations to English. A second method is to delete a certain fraction of the letters from a sample of English text and then let someone attempt to restore them. If they can be restored when 50% are deleted the redundancy must be greater than 50%. A third method depends on certain known results in cryptography.

Two extremes of redundancy in English prose are represented by Basic English and by James Joyces' book "Finigans Wake." The Basic English vocabulary is limited to 850 words and the redundancy is very high. This is reflected in the expansion that occurs when a passage is translated into Basic English. Joyce on the other hand enlarges the vocabulary and is alleged to achieve a compression of semantic content.

The redundancy of a language is related to the existence of crossword puzzles. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just possible when the redundancy is 50%. If the redundancy is 33%, three dimensional crossword puzzles should be possible, etc.

8. Representation of the Encoding and Decoding Operations

We have yet to represent mathematically the operations performed by the transmitter and receiver in encoding and decoding the information. Either of these will be called a discrete transducer. The input to the transducer is a sequence of input symbols and its output a sequence of output symbols. The transducer may have an internal memory so that its output depends not only on the present input symbol but also on the past history. We assume that the internal memory is finite, i.e. there exists a finite number m of possible states of the transducer and that its output is a function of the present state and the present input symbol. The next state will be a second function of these two quantities. Thus a transducer can be described by two functions:

$$y_n = f(x_n, \alpha_n)$$

$$\alpha_{n+1} = g(x_n, \alpha_n)$$

where: x_n is the n^{th} input symbol,

α_n is the state of the transducer when the n^{th} input symbol is introduced,

y_n is the output symbol (or sequence of output symbols) produced when x_n is introduced if the state is α_n.

If the output symbols of one transducer can be identified with the input symbols of a second, they can be connected in tandem and the result is also a transducer. If there exists a second transducer which operates on the output of the first and recovers the original input, the first transducer will be called non-singular and the second will be called its inverse.

Theorem 7: The output of a finite state transducer driven by a finite state statistical source is a finite state statistical source, with entropy (per unit time) less than or equal to that of the input. If the transducer is non-singular they are equal.

Let α represent the state of the source, which produces a sequence of symbols x_i; and let β be the state of the transducer, which produces, in its output, blocks of symbols y_j. The combined system can be represented by the "product state space" of pairs (α, β). Two points in the space, (α_1, β_1) and $(\alpha_2 \beta_2)$, are connected by a line if α_1 can produce an x which changes β_1 to β_2, and this line is given the probability of that x in this case. The line is labeled with the block of y_j symbols produced by the transducer. The entropy of the output can be calculated as the weighted sum over the states. If we sum first on β each resulting term is less than or equal to the corresponding term for α, hence the entropy is not increased. If the transducer is non-singular let its output be connected to the inverse transducer. If H_1', H_2' and H_3' are the output entropies of the source, the first and second transducers respectively, then $H_1' \geq H_2' \geq H_3' = H_1'$ and therefore $H_1' = H_2'$.

Suppose we have a system of constraints on possible sequences of the type which can be represented by a linear graph as in Fig. 2. If probabilities $p_{ij}^{(s)}$ were assigned to the various lines connecting state i to state j this would become a source. There is one particular assignment which maximizes the resulting entropy (see Appendix IV).

Theorem 8: Let the system of constraints considered as a channel have a capacity C. If we assign

$$p_{ij}^{(s)} = \frac{B_j}{B_i} C^{-\ell_{ij}^{(s)}}$$

where $\ell_{ij}^{(s)}$ is the duration of the s^{th} symbol leading from state i to state j and the B_i satisfy

$$B_i = \sum_{s,j} B_j C^{-\ell_{ij}^{(s)}}$$

then H is maximized and equal to C.

By proper assignment of the transition probabilities the entropy of symbols on a channel can be maximized at the channel capacity.

9. The Fundamental Theorem for a Noiseless Channel

We will now justify our interpretation of H as the rate of generating information by proving that H determines the channel capacity required with most efficient coding.

Theorem 9: Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{C}{H} - \epsilon$ symbols per second over the channel where ϵ is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

The converse part of the theorem, that $\frac{C}{H}$ cannot be exceeded, may be proved by noting that the entropy of the channel input per second is equal to that of the source, since the transmitter must be non-singular, and also this entropy cannot exceed the channel capacity. Hence $H' \leq C$ and the number of symbols per second $= H'/H \leq C/H$.

The first part of the theorem will be proved in two different ways. The first method is to consider the set of all sequences of N symbols produced by the source. For N large we can divide these into two groups, one containing less than $2^{(H+\eta)N}$ members and the second containing less than 2^{RN} members (where R is the logarithm of the number of different symbols) and having a total probability less than μ. As N increases η and μ approach zero. The number of signals of duration T in the channel is greater than $2^{(C-\theta)T}$ with θ small when T is large. If we choose

$$T = \left(\frac{H}{C} + \lambda\right) N$$

then there will be a sufficient number of sequences of channel symbols for the high probability group when N and T are sufficiently large (however small λ) and also some additional ones. The high probability group is coded in an arbitrary one to one way into this set. The remaining sequences are represented by larger sequences, starting and ending with one of the sequences not used for the high probability group. This special sequence acts as a start and stop signal for a different code. In between a sufficient time is allowed to give enough different sequences for all the low probability messages. This will require

$$T_1 = \left(\frac{R}{C} + \varphi\right) N$$

where φ is small. The mean rate of transmission in message symbols per second will then be greater than

$$\left[(1-\delta)\frac{T}{N}+\delta\frac{T_1}{N}\right]^{-1}=\left[(1-\delta)\left(\frac{H}{C}+\lambda\right)+\delta\left(\frac{R}{C}+\varphi\right)\right]^{-1}$$

As N increases δ, λ and φ approach zero and the rate approaches $\frac{C}{H}$.

Another method of performing this coding and proving the theorem can be described as follows: Arrange the messages of length N in order of decreasing probability and suppose their probabilities are $p_1 \geq p_2 \geq p_3 \ldots \geq p_n$. Let $P_s = \sum_1^{s-1} p_i$; that is P_s is the cumulative probability up to, but not including, p_s. We first encode into a binary system. The binary code for message s is obtained by expanding P_s as a binary number. The expansion is carried out to m_s places, where m_s is the integer satisfying:

$$\log_2\frac{1}{p_s} \leq m_s < 1+\log_2\frac{1}{p_s}$$

Thus the messages of high probability are represented by short codes and those of low probability by long codes. From these inequalities we have

$$\frac{1}{2^{m_s}} \leq p_s < \frac{1}{2^{m_s-1}}.$$

The code for P_s will differ from all succeeding ones in one or more of its m_s places, since all the remaining P_i are at least $\frac{1}{2^{m_s}}$ larger and their binary expansions therefore differ in the first m_s places. Consequently all the codes are different and it is possible to recover the message from its code. If the channel sequences are not already sequences of binary digits, they can be ascribed binary numbers in an arbitrary fashion and the binary code thus translated into signals suitable for the channel.

The average number H' of binary digits used per symbol of original message is easily estimated. We have

$$H' = \frac{1}{N}\Sigma m_s p_s$$

But,

$$\frac{1}{N}\Sigma\left(\log_2\frac{1}{p_s}\right)p_s \leq \frac{1}{N}\Sigma m_s p_s < \frac{1}{N}\Sigma\left(1+\log_2\frac{1}{p_s}\right)p_s$$

and therefore,

$$G_N \leq H' < G_N + \frac{1}{N}$$

As N increases G_N approaches H, the entropy of the source and H' approaches H.

We see from this that the inefficiency in coding, when only a finite delay of N symbols is used, need not be greater than $\frac{1}{N}$ plus the difference between the true entropy H and the entropy G_N calculated for sequences of length N. The per cent excess time needed over the ideal is therefore less than

$$\frac{G_N}{H}+\frac{1}{HN}-1.$$

This method of encoding is substantially the same as one found independently by R. M. Fano.[9] His method is to arrange the messages of length N in order of decreasing probability. Divide this series into two groups of as nearly equal probability as possible. If the message is in the first group its first binary digit will be 0, otherwise 1. The groups are similarly divided into subsets of nearly equal probability and the particular subset determines the second binary digit. This process is continued until each subset contains only one message. It is easily seen that apart from minor differences (generally in the last digit) this amounts to the same thing as the arithmetic process described above.

10. Discussion and Examples

In order to obtain the maximum power transfer from a generator to a load a transformer must in general be introduced so that the generator as seen from the load has the load resistance. The situation here is roughly analogous. The transducer which does the encoding should match the source to the channel in a statistical sense. The source as seen from the channel through the transducer should have the same statistical structure as the source which maximizes the entropy in the channel. The content of Theorem 9 is that, although an exact match is not in general possible, we can approximate it as closely as desired. The ratio of the actual rate of transmission to the capacity C may be called the efficiency of the coding system. This is of course equal to the ratio of the actual entropy of the channel symbols to the maximum possible entropy.

In general, ideal or nearly ideal encoding requires a long delay in the transmitter and receiver. In the noiseless case which we have been considering, the main function of this delay is to allow reasonably good

[9] Technical Report No. 65, The Research Laboratory of Electronics, M. I. T.

matching of probabilities to corresponding lengths of sequences. With a good code the logarithm of the reciprocal probability of a long message must be proportional to the duration of the corresponding signal, in fact

$$\left| \frac{\log p^{-1}}{T} - C \right|$$

must be small for all but a small fraction of the long messages.

If a source can produce only one particular message its entropy is zero, and no channel is required. For example, a computing machine set up to calculate the successive digits of π produces a definite sequence with no chance element. No channel is required to "transmit" this to another point. One could construct a second machine to compute the same sequence at the point. However, this may be impractical. In such a case we can choose to ignore some or all of the statistical knowledge we have of the source. We might consider the digits of π to be a random sequence in that we construct a system capable of sending any sequence of digits. In a similar way we may choose to use some of our statistical knowledge of English in constructing a code, but not all of it. In such a case we consider the source with the maximum entropy subject to the statistical conditions we wish to retain. The entropy of this source determines the channel capacity which is necessary and sufficient. In the π example the only information retained is that all the digits are chosen from the set 0, 1, ..., 9. In the case of English one might wish to use the statistical saving possible due to letter frequencies, but nothing else. The maximum entropy source is then the first approximation to English and its entropy determines the required channel capacity.

As a simple example of some of these results consider a source which produces a sequence of letters chosen from among A, B, C, D with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$, successive symbols being chosen independently. We have

$$H = -\left(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{2}{8}\log\tfrac{1}{8}\right)$$

$$= \tfrac{7}{4} \text{ bits per symbol.}$$

Thus we can approximate a coding system to encode messages from this source into binary digits with an average of $\frac{7}{4}$ binary digit per symbol. In this case we can actually achieve the limiting value by the following code (obtained by the method of the second proof of Theorem 9):

A	0
B	10
C	110
D	111

The average number of binary digits used in encoding a sequence of N symbols will be

$$N\left(\tfrac{1}{2}\times 1 + \tfrac{1}{4}\times 2 + \tfrac{2}{8}\times 3\right) = \tfrac{7}{4}N$$

It is easily seen that the binary digits 0, 1 have probabilities $\frac{1}{2}$, $\frac{1}{2}$ so the H for the coded sequences is one bit per symbol. Since, on the average, we have $\frac{7}{4}$ binary symbols per original letter, the entropies on a time basis are the same. The maximum possible entropy for the original set is $\log 4 = 2$, occurring when A, B, C, D have probabilities $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$. Hence the relative entropy is $\frac{7}{8}$. We can translate the binary sequences into the original set of symbols on a two-to-one basis by the following table:

00	A'
01	B'
10	C'
11	D'

This double process then encodes the original message into the same symbols but with an average compression ratio $\frac{7}{8}$.

As a second example consider a source which produces a sequence of A's and B's with probability p for A and q for B. If $p << q$ we have

$$H = -\log p^p(1-p)^{1-p}$$

$$= -p \log p\,(1-p)^{(1-p)/p}$$

$$\doteq p \log \frac{e}{p}$$

In such a case one can construct a fairly good coding of the message on a 0, 1 channel by sending a special sequence, say 0000, for the infrequent symbol A and then a sequence indicating the *number* of B's following it. This could be indicated by the binary representation with all numbers containing the special sequence deleted. All numbers up to 16 are represented as usual; 16 is represented by the next binary number after 16 which does not contain four zeros, namely $17 = 10001$, etc.

It can be shown that as $p \rightarrow 0$ the coding approaches ideal provided the length of the special sequence is properly adjusted.

PART II: THE DISCRETE CHANNEL WITH NOISE

11. Representation of a Noisy Discrete Channel

We now consider the case where the signal is perturbed by noise during transmission or at one or the other of the terminals. This means that the received signal is not necessarily the same as that sent out by the transmitter. Two cases may be distinguished. If a particular transmitted signal always produces the same received signal, i.e. the received signal is a definite function of the transmitted signal, then the effect may be called distortion. If this function has an inverse—no two transmitted signals producing the same received signal—distortion may be corrected, at least in principle, by merely performing the inverse functional operation on the received signal.

The case of interest here is that in which the signal does not always undergo the same change in transmission. In this case we may assume the received signal E to be a function of the transmitted signal S and a second variable, the noise N.

$$E = f(S, N)$$

The noise is considered to be a chance variable just as the message was above. In general it may be represented by a suitable stochastic process. The most general type of noisy discrete channel we shall consider is a generalization of the finite state noise free channel described previously. We assume a finite number of states and a set of probabilities

$$p_{\alpha, i}(\beta, j).$$

This is the probability, if the channel is in state α and symbol i is transmitted, that symbol j will be received and the channel left in state β. Thus α and β range over the possible states, i over the possible transmitted signals and j over the possible received signals. In the case where successive symbols are independently perturbed by the noise there is only one state, and the channel is described by the set of transition probabilities $p_i(j)$, the probability of transmitted symbol i being received as j.

If a noisy channel is fed by a source there are two statistical processes at work: the source and the noise. Thus there are a number of entropies that can be calculated. First there is the entropy $H(x)$ of the source or of the input to the channel (these will be equal if the transmitter is non-singular). The entropy of the output of the channel, i.e. the received signal, will be denoted by $H(y)$. In the noiseless case $H(y) = H(x)$. The joint entropy of input and output will be $H(xy)$. Finally there are two conditional entropies $H_x(y)$ and $H_y(x)$, the entropy of the output when the input is known and conversely. Among these quantities we have the relations

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

All of these entropies can be measured on a per-second or a per-symbol basis.

12. Equivocation and Channel Capacity

If the channel is noisy it is not in general possible to reconstruct the original message or the transmitted signal with *certainty* by any operation on the received signal E. There are, however, ways of transmitting the information which are optimal in combating noise. This is the problem which we now consider.

Suppose there are two possible symbols 0 and 1, and we are transmitting at a rate of 1000 symbols per second with probabilities $p_0 = p_1 = \frac{1}{2}$. Thus our source is producing information at the rate of 1000 bits per second. During transmission the noise introduces errors so that, on the average, 1 in 100 is received incorrectly (a 0 as 1, or 1 as 0). What is the rate of transmission of information? Certainly less than 1000 bits per second since about 1% of the received symbols are incorrect. Our first impulse might be to say the rate is 990 bits per second, merely subtracting the expected number of errors. This is not satisfactory since it fails to take into account the recipient's lack of knowledge of where the errors occur. We may carry it to an extreme case and suppose the noise so great that the received symbols are entirely independent of the transmitted symbols. The probability of receiving 1 is $\frac{1}{2}$ whatever was transmitted and similarly for 0. Then about half of the received symbols are correct due to chance alone, and we would be giving the system credit for transmitting 500 bits per second while actually no information is being transmitted at all. Equally "good" transmission would be obtained by dispensing with the channel entirely and flipping a coin at the receiving point.

Evidently the proper correction to apply to the amount of information transmitted is the amount of this information which is missing in the received signal, or alternatively the uncertainty when we have received a signal of what was actually sent. From our previous discussion of entropy as a measure of uncertainty it seems reasonable to use the conditional entropy of the message, knowing the received signal, as a measure of this missing information. This is indeed the proper definition, as we shall see later. Following this idea the rate of actual transmission, R, would be obtained by subtracting from the rate of production (i.e., the entropy of the source) the average rate of conditional entropy.

$$R = H(x) - H_y(x)$$

The conditional entropy $H_y(x)$ will, for convenience, be called the equivocation. It measures the average ambiguity of the received signal.

In the example considered above, if a 0 is received the *a posteriori* probability that a 0 was transmitted is .99, and that a 1 was transmitted is .01. These figures are reversed if a 1 is received. Hence

$$H_y(x) = -[.99 \log .99 + 0.01 \log 0.01]$$
$$= .081 \text{ bits/symbol}$$

or 81 bits per second. We may say that the system is transmitting at a rate 1000 − 81 = 919 bits per second. In the extreme case where a 0 is equally likely to be received as a 0 or 1 and similarly for 1, the a posteriori probabilities are $\frac{1}{2}$, $\frac{1}{2}$ and

$$H_y(x) = -[\tfrac{1}{2} \log \tfrac{1}{2} + \tfrac{1}{2} \log \tfrac{1}{2}]$$
$$= 1 \text{ bit per symbol}$$

or 1000 bits per second. The rate of transmission is then 0 as it should be.

The following theorem gives a direct intuitive interpretation of the equivocation and also serves to justify it as the unique appropriate measure. We consider a communication system and an observer (or auxiliary device) who can see both what is sent and what is recovered (with errors due to noise). This observer notes the errors in the recovered message and transmits data to the receiving point over a "correction channel" to enable the receiver to correct the errors. The situation is indicated schematically in Fig. 8.

Theorem 10: If the correction channel has a capacity equal to $H_y(x)$ it is possible to so encode the correction data as to send it over this channel and correct all but an arbitrarily small fraction ϵ of the errors. This is not possible if the channel capacity is less than $H_y(x)$.

Roughly then, $H_y(x)$ is the amount of additional information that must be supplied per second at the receiving point to correct the received message.

To prove the first part, consider long sequences of received message M' and corresponding original message M. There will be logarithmically $TH_y(x)$ of the M's which could reasonably have produced each M'. Thus we have $TH_y(x)$ binary digits to send each T seconds. This can be done with ϵ frequency of errors on a channel of capacity $H_y(x)$.

The second part can be proved by noting, first, that for any discrete chance variables x, y, z

$$H_y(x, z) \geq H_y(x)$$

The left-hand side can be expanded to give

$$H_y(z) + H_{yz}(x) \geq H_y(x)$$
$$H_{yz}(x) \geq H_y(x) - H_y(z) \geq H_y(x) - H(z)$$

If we identify x as the output of the source, y as the received signal and z as the signal sent over the correction channel, then the right-hand side is the equivocation less the rate of transmission over the correction channel. If the capacity of this channel is less than the equivocation the right-hand side will be greater than zero and $H_{yz}(x) \geq 0$. But this is the uncertainty of what was sent, knowing both the received signal and the correction signal. If this is greater than zero the frequency of errors cannot be arbitrarily small.

Example:

Suppose the errors occur at random in a sequence of binary digits: probability p that a digit is wrong and $q = 1 - p$ that it is right. These errors can be corrected if their position is known. Thus the correction channel need only send information as to these positions. This amounts to trans-

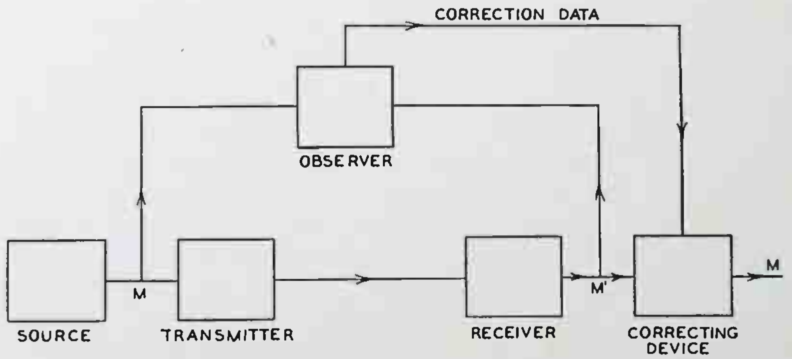


Fig. 8—Schematic diagram of a correction system.

mitting from a source which produces binary digits with probability p for 1 (correct) and q for 0 (incorrect). This requires a channel of capacity

$$-[p \log p + q \log q]$$

which is the equivocation of the original system.

The rate of transmission R can be written in two other forms due to the identities noted above. We have

$$R = H(x) - H_y(x)$$
$$= H(y) - H_x(y)$$
$$= H(x) + H(y) - H(x, y).$$

The first defining expression has already been interpreted as the amount of information sent less the uncertainty of what was sent. The second measures the amount received less the part of this which is due to noise. The third is the sum of the two amounts less the joint entropy and therefore in a sense is the number of bits per second common to the two. Thus all three expressions have a certain intuitive significance.

The capacity C of a noisy channel should be the maximum possible rate of transmission, i.e., the rate when the source is properly matched to the channel. We therefore define the channel capacity by

$$C = \text{Max}\ (H(x) - H_y(x))$$

where the maximum is with respect to all possible information sources used as input to the channel. If the channel is noiseless, $H_y(x) = 0$. The definition is then equivalent to that already given for a noiseless channel since the maximum entropy for the channel is its capacity.

13. The Fundamental Theorem for a Discrete Channel with Noise

It may seem surprising that we should define a definite capacity C for a noisy channel since we can never send certain information in such a case. It is clear, however, that by sending the information in a redundant form the probability of errors can be reduced. For example, by repeating the message many times and by a statistical study of the different received versions of the message the probability of errors could be made very small. One would expect, however, that to make this probability of errors approach zero, the redundancy of the encoding must increase indefinitely, and the rate of transmission therefore approach zero. This is by no means true. If it were, there would not be a very well defined capacity, but only a capacity for a given frequency of errors, or a given equivocation; the capacity going down as the error requirements are made more stringent. Actually the capacity C defined above has a very definite significance. It is possible to send information at the rate C through the channel *with as small a frequency of errors or equivocation as desired* by proper encoding. This statement is not true for any rate greater than C. If an attempt is made to transmit at a higher rate than C, say $C + R_1$, then there will necessarily be an equivocation equal to a greater than the excess R_1. Nature takes payment by requiring just that much uncertainty, so that we are not actually getting any more than C through correctly.

The situation is indicated in Fig. 9. The rate of information into the channel is plotted horizontally and the equivocation vertically. Any point above the heavy line in the shaded region can be attained and those below cannot. The points on the line cannot in general be attained, but there will usually be two points on the line that can.

These results are the main justification for the definition of C and will now be proved.

Theorem 11. Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \epsilon$ where ϵ is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$.

The method of proving the first part of this theorem is not by exhibiting a coding method having the desired properties, but by showing that such a code must exist in a certain group of codes. In fact we will average the frequency of errors over this group and show that this average can be made less than ϵ. If the average of a set of numbers is less than ϵ there must exist at least one in the set which is less than ϵ. This will establish the desired result.

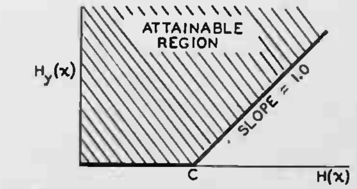


Fig. 9—The equivocation possible for a given input entropy to a channel.

The capacity C of a noisy channel has been defined as

$$C = \text{Max}\ (H(x) - H_y(x))$$

where x is the input and y the output. The maximization is over all sources which might be used as input to the channel.

Let S_0 be a source which achieves the maximum capacity C. If this maximum is not actually achieved by any source let S_0 be a source which approximates to giving the maximum rate. Suppose S_0 is used as input to the channel. We consider the possible transmitted and received sequences of a long duration T. The following will be true:

1. The transmitted sequences fall into two classes, a high probability group with about $2^{TH(x)}$ members and the remaining sequences of small total probability.
2. Similarly the received sequences have a high probability set of about $2^{TH(y)}$ members and a low probability set of remaining sequences.
3. Each high probability output could be produced by about $2^{TH_y(x)}$ inputs. The probability of all other cases has a small total probability.

All the ϵ's and δ's implied by the words "small" and "about" in these statements approach zero as we allow T to increase and S_0 to approach the maximizing source.

The situation is summarized in Fig. 10 where the input sequences are points on the left and output sequences points on the right. The fan of cross lines represents the range of possible causes for a typical output.

Now suppose we have another source producing information at rate R with $R < C$. In the period T this source will have 2^{TR} high probability outputs. We wish to associate these with a selection of the possible channe

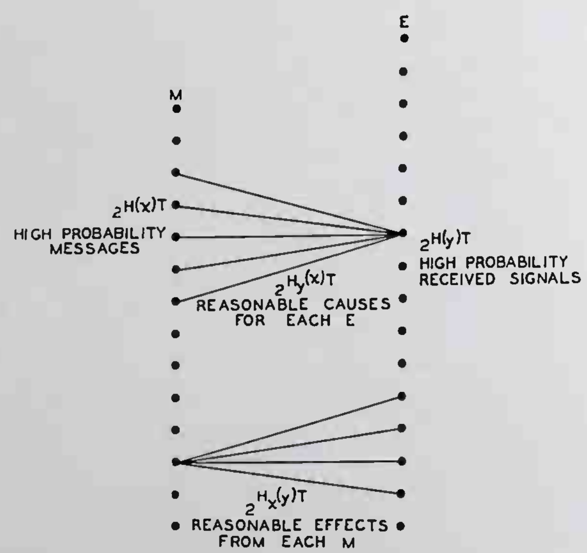


Fig. 10—Schematic representation of the relations between inputs and outputs in a channel.

inputs in such a way as to get a small frequency of errors. We will set up this association in all possible ways (using, however, only the high probability group of inputs as determined by the source S_0) and average the frequency of errors for this large class of possible coding systems. This is the same as calculating the frequency of errors for a random association of the messages and channel inputs of duration T. Suppose a particular output y_1 is observed. What is the probability of more than one message in the set of possible causes of y_1? There are 2^{TR} messages distributed at random in $2^{TH(x)}$ points. The probability of a particular point being a message is thus

$$2^{T(R-H(x))}$$

The probability that none of the points in the fan is a message (apart from the actual originating message) is

$$P = [1 - 2^{T(R-H(x))}]^{2^{TH_y(x)}}$$

Now $R < H(x) - H_y(x)$ so $R - H(x) = -H_y(x) - \eta$ with η positive. Consequently

$$P = [1 - 2^{-TH_y(x)-T\eta}]^{2^{TH_y(x)}}$$

approaches (as $T \to \infty$)

$$1 - 2^{-T\eta}.$$

Hence the probability of an error approaches zero and the first part of the theorem is proved.

The second part of the theorem is easily shown by noting that we could merely send C bits per second from the source, completely neglecting the remainder of the information generated. At the receiver the neglected part gives an equivocation $H(x) - C$ and the part transmitted need only add ϵ. This limit can also be attained in many other ways, as will be shown when we consider the continuous case.

The last statement of the theorem is a simple consequence of our definition of C. Suppose we can encode a source with $R = C + a$ in such a way as to obtain an equivocation $H_y(x) = a - \epsilon$ with ϵ positive. Then $R = H(x) = C + a$ and

$$H(x) - H_y(x) = C + \epsilon$$

with ϵ positive. This contradicts the definition of C as the maximum of $H(x) - H_y(x)$.

Actually more has been proved than was stated in the theorem. If the average of a set of numbers is within ϵ of their maximum, a fraction of at most $\sqrt{\epsilon}$ can be more than $\sqrt{\epsilon}$ below the maximum. Since ϵ is arbitrarily small we can say that almost all the systems are arbitrarily close to the ideal.

14. Discussion

The demonstration of theorem 11, while not a pure existence proof, has some of the deficiencies of such proofs. An attempt to obtain a good approximation to ideal coding by following the method of the proof is generally impractical. In fact, apart from some rather trivial cases and certain limiting situations, no explicit description of a series of approximation to the ideal has been found. Probably this is no accident but is related to the difficulty of giving an explicit construction for a good approximation to a random sequence.

An approximation to the ideal would have the property that if the signal is altered in a reasonable way by the noise, the original can still be recovered. In other words the alteration will not in general bring it closer to another reasonable signal than the original. This is accomplished at the cost of a certain amount of redundancy in the coding. The redundancy must be introduced in the proper way to combat the particular noise structure involved. However, any redundancy in the source will usually help if it is utilized at the receiving point. In particular, if the source already has a certain redundancy and no attempt is made to eliminate it in matching to the channel, this redundancy will help combat noise. For example, in a noiseless telegraph channel one could save about 50% in time by proper encoding of the messages. This is not done and most of the redundancy of English remains in the channel symbols. This has the advantage, however, of allowing considerable noise in the channel. A sizable fraction of the letters can be received incorrectly and still reconstructed by the context. In fact this is probably not a bad approximation to the ideal in many cases, since the statistical structure of English is rather involved and the reasonable English sequences are not too far (in the sense required for theorem) from a random selection.

As in the noiseless case a delay is generally required to approach the ideal encoding. It now has the additional function of allowing a large sample of noise to affect the signal before any judgment is made at the receiving point as to the original message. Increasing the sample size always sharpens the possible statistical assertions.

The content of theorem 11 and its proof can be formulated in a somewhat different way which exhibits the connection with the noiseless case more clearly. Consider the possible signals of duration T and suppose a subset of them is selected to be used. Let those in the subset all be used with equal probability, and suppose the receiver is constructed to select, as the original signal, the most probable cause from the subset, when a perturbed signal is received. We define $N(T, q)$ to be the maximum number of signals we can choose for the subset such that the probability of an incorrect interpretation is less than or equal to q.

Theorem 12: $\lim_{T \to \infty} \frac{\log N(T, q)}{T} = C$, where C is the channel capacity, provided that q does not equal 0 or 1.

In other words, no matter how we set our limits of reliability, we can distinguish reliably in time T enough messages to correspond to about CT bits, when T is sufficiently large. Theorem 12 can be compared with the definition of the capacity of a noiseless channel given in section 1.

15. Example of a Discrete Channel and Its Capacity

A simple example of a discrete channel is indicated in Fig. 11. There are three possible symbols. The first is never affected by noise. The second and third each have probability p of coming through undisturbed, and q of being changed into the other of the pair. We have (letting $\alpha = -[p \log p + q \log q]$ and P and Q be the probabilities of using the first or second symbols)

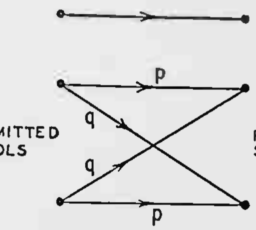


Fig. 11—Example of a discrete channel.

$$H(x) = -P \log P - 2Q \log Q$$

$$H_y(x) = 2Q\alpha$$

We wish to choose P and Q in such a way as to maximize $H(x) - H_y(x)$, subject to the constraint $P + 2Q = 1$. Hence we consider

$$U = -P \log P - 2Q \log Q - 2Q\alpha + \lambda(P + 2Q)$$

$$\frac{\partial U}{\partial P} = -1 - \log P + \lambda = 0$$

$$\frac{\partial U}{\partial Q} = -2 - 2 \log Q - 2\alpha + 2\lambda = 0.$$

Eliminating λ

$$\log P = \log Q + \alpha$$

$$P = Qe^{\alpha} = Q\beta$$

$$P = \frac{\beta}{\beta + 2} \qquad Q = \frac{1}{\beta + 2}.$$

The channel capacity is then

$$C = \log \frac{\beta + 2}{\beta}.$$

Note how this checks the obvious values in the cases $p = 1$ and $p = \frac{1}{2}$. In the first, $\beta = 1$ and $C = \log 3$, which is correct since the channel is then noiseless with three possible symbols. If $p = \frac{1}{2}$, $\beta = 2$ and $C = \log 2$. Here the second and third symbols cannot be distinguished at all and act together like one symbol. The first symbol is used with probability $P = \frac{1}{2}$ and the second and third together with probability $\frac{1}{2}$. This may be distributed in any desired way and still achieve the maximum capacity.

For intermediate values of p the channel capacity will lie between log 2 and log 3. The distinction between the second and third symbols conveys some information but not as much as in the noiseless case. The first symbol is used somewhat more frequently than the other two because of its freedom from noise.

16. The Channel Capacity in Certain Special Cases

If the noise affects successive channel symbols independently it can be described by a set of transition probabilities p_{ij}. This is the probability, if symbol i is sent, that j will be received. The maximum channel rate is then given by the maximum of

$$\sum_{i,j} P_i p_{ij} \log \sum_i P_i p_{ij} - \sum_{i,j} P_i p_{ij} \log p_{ij}$$

where we vary the P_i subject to $\Sigma P_i = 1$. This leads by the method of Lagrange to the equations,

$$\sum_j p_{sj} \log \frac{p_{sj}}{\sum_i P_i p_{ij}} = \mu \qquad s = 1, 2, \cdots .$$

Multiplying by P_s and summing on s shows that $\mu = -C$. Let the inverse of p_{sj} (if it exists) be h_{st} so that $\sum_s h_{st} p_{sj} = \delta_{tj}$. Then:

$$\sum_{s,j} h_{st} p_{sj} \log p_{sj} - \log \sum_i P_i p_{it} = -C \sum_s h_{st} .$$

Hence:

$$\sum_i P_i p_{it} = \exp \left[C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj}\right]$$

or,

$$P_i = \sum_t h_{it} \exp \left[C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj}\right].$$

This is the system of equations for determining the maximizing values of P_i, with C to be determined so that $\Sigma P_i = 1$. When this is done C will be the channel capacity, and the P_i the proper probabilities for the channel symbols to achieve this capacity.

If each input symbol has the same set of probabilities on the lines emerging from it, and the same is true of each output symbol, the capacity can be easily calculated. Examples are shown in Fig. 12. In such a case $H_x(y)$ is independent of the distribution of probabilities on the input symbols, and is given by $-\Sigma\, p_i \log p_i$ where the p_i are the values of the transition probabilities from any input symbol. The channel capacity is

$$\text{Max} \quad [H(y) - H_x(y)]$$
$$= \text{Max}\, H(y) + \Sigma\, p_i \log p_i .$$

The maximum of $H(y)$ is clearly log m where m is the number of output

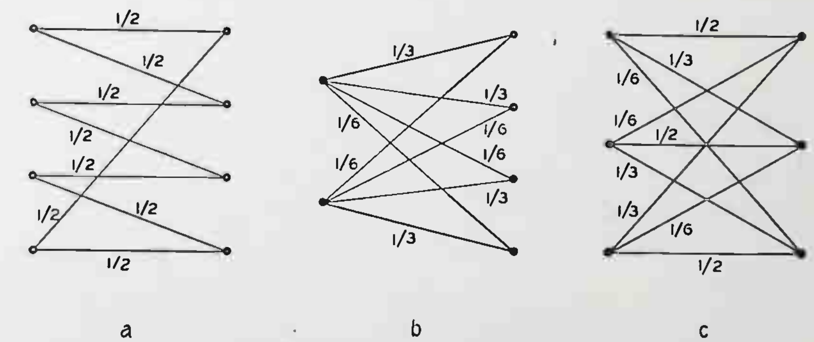


Fig. 12—Examples of discrete channels with the same transition probabilities for each input and for each output.

symbols, since it is possible to make them all equally probable by making the input symbols equally probable. The channel capacity is therefore

$$C = \log m + \Sigma\, p_i \log p_i .$$

In Fig. 12a it would be

$$C = \log 4 - \log 2 = \log 2 .$$

This could be achieved by using only the 1st and 3d symbols. In Fig. 12b

$$C = \log 4 - \tfrac{2}{3} \log 3 - \tfrac{1}{3} \log 6$$
$$= \log 4 - \log 3 - \tfrac{1}{3} \log 2$$
$$= \log \tfrac{1}{3}\, 2^{\frac{5}{3}} .$$

In Fig. 12c we have

$$C = \log 3 - \tfrac{1}{2} \log 2 - \tfrac{1}{3} \log 3 - \tfrac{1}{6} \log 6$$
$$= \log \frac{3}{2^{\frac{1}{2}}\, 3^{\frac{1}{3}}\, 6^{\frac{1}{6}}} .$$

Suppose the symbols fall into several groups such that the noise never causes a symbol in one group to be mistaken for a symbol in another group. Let the capacity for the nth group be C_n when we use only the symbols in this group. Then it is easily shown that, for best use of the entire set, the total probability P_n of all symbols in the nth group should be

$$P_n = \frac{2^{C_n}}{\Sigma 2^{C_n}}.$$

Within a group the probability is distributed just as it would be if these were the only symbols being used. The channel capacity is

$$C = \log \Sigma 2^{C_n}.$$

17. An Example of Efficient Coding

The following example, although somewhat unrealistic, is a case in which exact matching to a noisy channel is possible. There are two channel symbols, 0 and 1, and the noise affects them in blocks of seven symbols. A block of seven is either transmitted without error, or exactly one symbol of the seven is incorrect. These eight possibilities are equally likely. We have

$$\begin{aligned} C &= \text{Max}\,[H(y) - H_x(y)] \\ &= \tfrac{1}{7}\left[7 + \tfrac{8}{8}\log \tfrac{1}{8}\right] \\ &= \tfrac{4}{7}\text{ bits/symbol}. \end{aligned}$$

An efficient code, allowing complete correction of errors and transmitting at the rate C, is the following (found by a method due to R. Hamming):

Let a block of seven symbols be $X_1, X_2, \ldots X_7$. Of these X_3, X_5, X_6 and X_7 are message symbols and chosen arbitrarily by the source. The other three are redundant and calculated as follows:

X_4 is chosen to make $\alpha = X_4 + X_5 + X_6 + X_7$ even
X_2 " " " $\beta = X_2 + X_3 + X_6 + X_7$ "
X_1 " " " $\gamma = X_1 + X_3 + X_5 + X_7$ "

When a block of seven is received α, β and γ are calculated and if even called zero, if odd called one. The binary number $\alpha\,\beta\,\gamma$ then gives the subscript of the X_i that is incorrect (if 0 there was no error).

APPENDIX 1

The Growth of the Number of Blocks of Symbols With a Finite State Condition

Let $N_i(L)$ be the number of blocks of symbols of length L ending in state i. Then we have

$$N_j(L) = \sum_{is} N_i(L - b_{ij}^{(s)})$$

where $b_{ij}^1, b_{ij}^2, \ldots b_{ij}^m$ are the length of the symbols which may be chosen in state i and lead to state j. These are linear difference equations and the behavior as $L \to \infty$ must be of the type

$$N_j = A_j W^L$$

Substituting in the difference equation

$$A_j W^L = \sum_{i,s} A_i W^{L - b_{ij}^{(s)}}$$

or

$$A_j = \sum_{is} A_i W^{-b_{ij}^{(s)}}$$

$$\sum_i \left(\sum_s W^{-b_{ij}^{(s)}} - \delta_{ij}\right) A_i = 0.$$

For this to be possible the determinant

$$D(W) = |a_{ij}| = \left|\sum_s W^{-b_{ij}^{(s)}} - \delta_{ij}\right|$$

must vanish and this determines W, which is, of course, the largest real root of $D = 0$.

The quantity C is then given by

$$C = \lim_{L \to \infty} \frac{\log \Sigma A_j W^L}{L} = \log W$$

and we also note that the same growth properties result if we require that all blocks start in the same (arbitrarily chosen) state.

APPENDIX 2

Derivation of $H = -\Sigma\, p_i \log p_i$

Let $H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) = A(n)$. From condition (3) we can decompose a choice from s^m equally likely possibilities into a series of m choices each from s equally likely possibilities and obtain

$$A(s^m) = m\,A(s)$$

Similarly

$$A(t^n) = n\,A(t)$$

We can choose n arbitrarily large and find an m to satisfy

$$s^m \le t^n < s^{(m+1)}$$

Thus, taking logarithms and dividing by $n \log s$,

$$\frac{m}{n} \le \frac{\log t}{\log s} \le \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon$$

where ϵ is arbitrarily small.

Now from the monotonic property of $A(n)$

$$A(s^m) \le A(t^n) \le A(s^{m+1})$$

$$m A(s) \le nA(t) \le (m + 1) A(s)$$

Hence, dividing by $nA(s)$,

$$\frac{m}{n} \le \frac{A(t)}{A(s)} \le \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \epsilon$$

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \le 2\epsilon \qquad A(t) = -K \log t$$

where K must be positive to satisfy (2).

Now suppose we have a choice from n possibilities with commeasurable probabilities $p_i = \frac{n_i}{\Sigma n_i}$ where the n_i are integers. We can break down a choice from Σn_i possibilities into a choice from n possibilities with probabilities $p_i \ldots p_n$ and then, if the ith was chosen, a choice from n_i with equal probabilities. Using condition 3 again, we equate the total choice from Σn_i as computed by two methods

$$K \log \Sigma n_i = H(p_1, \ldots, p_n) + K\Sigma\, p_i \log n_i$$

Hence

$$H = K\,[\Sigma\, p_i \log \Sigma\, n_i - \Sigma\, p_i \log n_i]$$

$$= -K\, \Sigma\, p_i \log \frac{n_i}{\Sigma n_i} = -K\, \Sigma\, p_i \log p_i\,.$$

If the p_i are incommeasurable, they may be approximated by rationals and the same expression must hold by our continuity assumption. Thus the expression holds in general. The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure.

APPENDIX 3

Theorems on Ergodic Sources

If it is possible to go from any state with $P > 0$ to any other along a path of probability $p > 0$, the system is ergodic and the strong law of large numbers can be applied. Thus the number of times a given path p_{ij} in the net-

work is traversed in a long sequence of length N is about proportional to the probability of being at i and then choosing this path, $P_i p_{ij} N$. If N is large enough the probability of percentage error $\pm\,\delta$ in this is less than ϵ so that for all but a set of small probability the actual numbers lie within the limits

$$(P_i p_{ij} \pm \delta)N$$

Hence nearly all sequences have a probability p given by

$$p = \Pi\, p_{ij}^{(P_i p_{ij} \pm \delta)N}$$

and $\frac{\log p}{N}$ is limited by

$$\frac{\log p}{N} = \Sigma (P_i\, p_{ij} \pm \delta) \log p_{ij}$$

or

$$\left| \frac{\log p}{N} - \Sigma P_i\, p_{ij} \log p_{ij} \right| < \eta.$$

This proves theorem 3.

Theorem 4 follows immediately from this on calculating upper and lower bounds for $n(q)$ based on the possible range of values of p in Theorem 3.

In the mixed (not ergodic) case if

$$L = \Sigma\, p_i\, L_i$$

and the entropies of the components are $H_1 \ge H_2 \ge \ldots \ge H_n$ we have the

Theorem: $\lim\limits_{N\to\infty} \frac{\log n(q)}{N} = \varphi(q)$ is a decreasing step function,

$$\varphi(q) = H_s \quad \text{in the interval} \quad \sum_1^{s-1} \alpha_i < q < \sum_1^{s} \alpha_i\,.$$

To prove theorems 5 and 6 first note that F_N is monotonic decreasing because increasing N adds a subscript to a conditional entropy. A simple substitution for $p_{B_i}(S_j)$ in the definition of F_N shows that

$$F_N = N\, G_N - (N - 1)\, G_{N-1}$$

and summing this for all N gives $G_N = \frac{1}{N} \Sigma\, F_N$. Hence $G_N \ge F_N$ and G_N monotonic decreasing. Also they must approach the same limit. By using theorem 3 we see that $\lim\limits_{N\to\infty} G_N = H$.

APPENDIX 4

Maximizing the Rate for a System of Constraints

Suppose we have a set of constraints on sequences of symbols that is of the finite state type and can be represented therefore by a linear graph.

Let $\ell_{ij}^{(s)}$ be the lengths of the various symbols that can occur in passing from state i to state j. What distribution of probabilities P_i for the different states and $p_{ij}^{(s)}$ for choosing symbol s in state i and going to state j maximizes the rate of generating information under these constraints? The constraints define a discrete channel and the maximum rate must be less than or equal to the capacity C of this channel, since if all blocks of large length were equally likely, this rate would result, and if possible this would be best. We will show that this rate can be achieved by proper choice of the P_i and $p_{ij}^{(s)}$.

The rate in question is

$$\frac{-\Sigma P_i p_{ij}^{(s)} \log p_{ij}^{(s)}}{\Sigma P_{(i)} p_{ij}^{(s)} \ell_{ij}^{(s)}} = \frac{N}{M}.$$

Let $\ell_{ij} = \sum_s \ell_{ij}^{(s)}$. Evidently for a maximum $p_{ij}^{(s)} = k \exp \ell_{ij}^{(s)}$. The constraints on maximization are $\Sigma P_i = 1$, $\sum_j p_{ij} = 1$, $\Sigma P_i(p_{ij} - \delta_{ij}) = 0$. Hence we maximize

$$U = \frac{-\Sigma P_i p_{ij} \log p_{ij}}{\Sigma P_i p_{ij} \ell_{ij}} + \lambda \sum_i P_i + \Sigma \mu_i p_{ij} + \Sigma \eta_j P_i (p_{ij} - \delta_{ij})$$

$$\frac{\partial U}{\partial p_{ij}} = -\frac{MP_i(1 + \log p_{ij}) + NP_i \ell_{ij}}{M^2} + \lambda + \mu_i + \eta_i P_i = 0.$$

Solving for p_{ij}

$$p_{ij} = A_i B_j D^{-\ell_{ij}}.$$

Since

$$\sum_j p_{ij} = 1, \qquad A_i^{-1} = \sum_j B_j D^{-\ell_{ij}}$$

$$p_{ij} = \frac{B_j D^{-\ell_{ij}}}{\sum_s B_s D^{-\ell_{is}}}.$$

The correct value of D is the capacity C and the B_j are solutions of

$$B_i = \Sigma B_j C^{-\ell_{ij}}$$

for then

$$p_{ij} = \frac{B_j}{B_i} C^{-\ell_{ij}}$$

$$\Sigma P_i \frac{B_j}{B_i} C^{-\ell_{ij}} = P_j$$

or

$$\Sigma \frac{P_i}{B_i} C^{-\ell_{ij}} = \frac{P_j}{B_j}$$

So that if λ_i satisfy

$$\Sigma \gamma_i C^{-\ell_{ij}} = \gamma_j$$

$$P_i = B_i \gamma_i$$

Both of the sets of equations for B_i and γ_i can be satisfied since C is such that

$$|C^{-\ell_{ij}} - \delta_{ij}| = 0$$

In this case the rate is

$$-\frac{\Sigma P_i p_{ij} \log \frac{B_j}{B_i} C^{-\ell_{ij}}}{\Sigma P_i p_{ij} \ell_{ij}}$$

$$= C - \frac{\Sigma P_i p_{ij} \log \frac{B_j}{B_i}}{\Sigma P_i p_{ij} \ell_{ij}}$$

but

$$\Sigma P_i p_{ij} (\log B_j - \log B_i) = \sum_j P_j \log B_j - \Sigma P_i \log B_i = 0$$

Hence the rate is C and as this could never be exceeded this is the maximum' justifying the assumed solution.

PART III: MATHEMATICAL PRELIMINARIES

In this final installment of the paper we consider the case where the signals or the messages or both are continuously variable, in contrast with the discrete nature assumed until now. To a considerable extent the continuous case can be obtained through a limiting process from the discrete case by dividing the continuum of messages and signals into a large but finite number of small regions and calculating the various parameters involved on a discrete basis. As the size of the regions is decreased these parameters in general approach as limits the proper values for the continuous case. There are, however, a few new effects that appear and also a general change of emphasis in the direction of specialization of the general results to particular cases.

We will not attempt, in the continuous case, to obtain our results with the greatest generality, or with the extreme rigor of pure mathematics, since this would involve a great deal of abstract measure theory and would obscure the main thread of the analysis. A preliminary study, however, indicates that the theory can be formulated in a completely axiomatic and rigorous manner which includes both the continuous and discrete cases and many others. The occasional liberties taken with limiting processes in the present analysis can be justified in all cases of practical interest.

18. Sets and Ensembles of Functions

We shall have to deal in the continuous case with sets of functions and ensembles of functions. A set of functions, as the name implies, is merely a class or collection of functions, generally of one variable, time. It can be specified by giving an explicit representation of the various functions in the set, or implicitly by giving a property which functions in the set possess and others do not. Some examples are:

1. The set of functions:

$$f_\theta(t) = \sin (t + \theta).$$

Each particular value of θ determines a particular function in the set.

2. The set of all functions of time containing no frequencies over W cycles per second.
3. The set of all functions limited in band to W and in amplitude to A.
4. The set of all English speech signals as functions of time.

An *ensemble* of functions is a set of functions together with a probability measure whereby we may determine the probability of a function in the set having certain properties.[1] For example with the set,

$$f_\theta(t) = \sin (t + \theta),$$

we may give a probability distribution for θ, $P(\theta)$. The set then becomes an ensemble.

Some further examples of ensembles of functions are:

1. A finite set of functions $f_k(t)$ $(k = 1, 2, \cdots, n)$ with the probability of f_k being p_k.
2. A finite dimensional family of functions

$$f(\alpha_1, \alpha_2, \cdots, \alpha_n; t)$$

with a probability distribution for the parameters α_i:

$$p(\alpha_1, \cdots, \alpha_n)$$

For example we could consider the ensemble defined by

$$f(a_1, \cdots, a_n, \theta_1, \cdots, \theta_n; t) = \sum_{n=1}^{n} a_n \sin n(\omega t + \theta_n)$$

with the amplitudes a_i distributed normally and independently, and the phrases θ_i distributed uniformly (from 0 to 2π) and independently.

3. The ensemble

$$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

with the a_i normal and independent all with the same standard deviation $\sqrt{N}$. This is a representation of "white" noise, band-limited to the band from 0 to W cycles per second and with average power N.[2]

[1] In mathematical terminology the functions belong to a measure space whose total measure is unity.

[2] This representation can be used as a definition of band limited white noise. It has certain advantages in that it involves fewer limiting operations than do definitions that have been used in the past. The name "white noise," already firmly intrenched in the literature, is perhaps somewhat unfortunate. In optics white light means either any continuous spectrum as contrasted with a point spectrum, or a spectrum which is flat with wavelength (which is not the same as a spectrum flat with frequency).

4. Let points be distributed on the t axis according to a Poisson distribution. At each selected point the function $f(t)$ is placed and the different functions added, giving the ensemble

$$\sum_{k=-\infty}^{\infty} f(t + t_k)$$

where the t_k are the points of the Poisson distribution. This ensemble can be considered as a type of impulse or shot noise where all the impulses are identical.
5. The set of English speech functions with the probability measure given by the frequency of occurrence in ordinary use.

An ensemble of functions $f_\alpha(t)$ is *stationary* if the same ensemble results when all functions are shifted any fixed amount in time. The ensemble

$$f_\theta(t) = \sin (t + \theta)$$

is stationary if θ distributed uniformly from 0 to 2π. If we shift each function by t_1 we obtain

$$f_\theta(t + t_1) = \sin (t + t_1 + \theta)$$
$$= \sin (t + \varphi)$$

with φ distributed uniformly from 0 to 2π. Each function has changed but the ensemble as a whole is invariant under the translation. The other examples given above are also stationary.

An ensemble is *ergodic* if it is stationary, and there is no subset of the functions in the set with a probability different from 0 and 1 which is stationary. The ensemble

$$\sin (t + \theta)$$

is ergodic. No subset of these functions of probability $\neq 0, 1$ is transformed into itself under all time translations. On the other hand the ensemble

$$a \sin (t + \theta)$$

with a distributed normally and θ uniform is stationary but not ergodic. The subset of these functions with a between 0 and 1 for example is stationary.

Of the examples given, 3 and 4 are ergodic, and 5 may perhaps be considered so. If an ensemble is ergodic we may say roughly that each function in the set is typical of the ensemble. More precisely it is known that with an ergodic ensemble an average of any statistic over the ensemble is equal (with probability 1) to an average over all the time translations of a

2. The set of all functions of time containing no frequencies over W cycles per second.
3. The set of all functions limited in band to W and in amplitude to A.
4. The set of all English speech signals as functions of time.

An *ensemble* of functions is a set of functions together with a probability measure whereby we may determine the probability of a function in the set having certain properties.[1] For example with the set,

$$f_\theta(t) = \sin (t + \theta),$$

we may give a probability distribution for θ, $P(\theta)$. The set then becomes an ensemble.

Some further examples of ensembles of functions are:

1. A finite set of functions $f_k(t)$ $(k = 1, 2, \cdots, n)$ with the probability of f_k being p_k.
2. A finite dimensional family of functions

$$f(\alpha_1, \alpha_2, \cdots, \alpha_n; t)$$

with a probability distribution for the parameters α_i:

$$p(\alpha_1, \cdots, \alpha_n)$$

For example we could consider the ensemble defined by

$$f(a_1, \cdots, a_n, \theta_1, \cdots, \theta_n; t) = \sum_{n=1}^{n} a_n \sin n(\omega t + \theta_n)$$

with the amplitudes a_i distributed normally and independently, and the phrases θ_i distributed uniformly (from 0 to 2π) and independently.

3. The ensemble

$$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

with the a_i normal and independent all with the same standard deviation $\sqrt{N}$. This is a representation of "white" noise, band-limited to the band from 0 to W cycles per second and with average power N.[2]

[1] In mathematical terminology the functions belong to a measure space whose total measure is unity.

[2] This representation can be used as a definition of band limited white noise. It has certain advantages in that it involves fewer limiting operations than do definitions that have been used in the past. The name "white noise," already firmly intrenched in the literature, is perhaps somewhat unfortunate. In optics white light means either any continuous spectrum as contrasted with a point spectrum, or a spectrum which is flat with *wavelength* (which is not the same as a spectrum flat with frequency).

4. Let points be distributed on the t axis according to a Poisson distribution. At each selected point the function $f(t)$ is placed and the different functions added, giving the ensemble

$$\sum_{k=-\infty}^{\infty} f(t + t_k)$$

where the t_k are the points of the Poisson distribution. This ensemble can be considered as a type of impulse or shot noise where all the impulses are identical.

5. The set of English speech functions with the probability measure given by the frequency of occurrence in ordinary use.

An ensemble of functions $f_\alpha(t)$ is *stationary* if the same ensemble results when all functions are shifted any fixed amount in time. The ensemble

$$f_\theta(t) = \sin (t + \theta)$$

is stationary if θ distributed uniformly from 0 to 2π. If we shift each function by t_1 we obtain

$$f_\theta(t + t_1) = \sin (t + t_1 + \theta)$$
$$= \sin (t + \varphi)$$

with φ distributed uniformly from 0 to 2π. Each function has changed but the ensemble as a whole is invariant under the translation. The other examples given above are also stationary.

An ensemble is *ergodic* if it is stationary, and there is no subset of the functions in the set with a probability different from 0 and 1 which is stationary. The ensemble

$$\sin (t + \theta)$$

is ergodic. No subset of these functions of probability $\neq 0, 1$ is transformed into itself under all time translations. On the other hand the ensemble

$$a \sin (t + \theta)$$

with a distributed normally and θ uniform is stationary but not ergodic. The subset of these functions with a between 0 and 1 for example is stationary.

Of the examples given, 3 and 4 are ergodic, and 5 may perhaps be considered so. If an ensemble is ergodic we may say roughly that each function in the set is typical of the ensemble. More precisely it is known that with an ergodic ensemble an average of any statistic over the ensemble is equal (with probability 1) to an average over all the time translations of a particular function in the set.[3] Roughly speaking, each function can be expected, as time progresses, to go through, with the proper frequency, all the convolutions of any of the functions in the set.

Just as we may perform various operations on numbers or functions to obtain new numbers or functions, we can perform operations on ensembles to obtain new ensembles. Suppose, for example, we have an ensemble of functions $f_\alpha(t)$ and an operator T which gives for each function $f_\alpha(t)$ a result $g_\alpha(t)$:

$$g_\alpha(t) = Tf_\alpha(t)$$

Probability measure is defined for the set $g_\alpha(t)$ by means of that for the set $f_\alpha(t)$. The probability of a certain subset of the $g_\alpha(t)$ functions is equal to that of the subset of the $f_\alpha(t)$ functions which produce members of the given subset of g functions under the operation T. Physically this corresponds to passing the ensemble through some device, for example, a filter, a rectifier or a modulator. The output functions of the device form the ensemble $g_\alpha(t)$.

A device or operator T will be called invariant if shifting the input merely shifts the output, i.e., if

$$g_\alpha(t) = Tf_\alpha(t)$$

implies

$$g_\alpha(t + t_1) = Tf_\alpha(t + t_1)$$

for all $f_\alpha(t)$ and all t_1. It is easily shown (see appendix 1) that if T is invariant and the input ensemble is stationary then the output ensemble is stationary. Likewise if the input is ergodic the output will also be ergodic.

A filter or a rectifier is invariant under all time translations. The operation of modulation is not since the carrier phase gives a certain time structure. However, modulation is invariant under all translations which are multiples of the period of the carrier.

Wiener has pointed out the intimate relation between the invariance of physical devices under time translations and Fourier theory.[4] He has shown, in fact, that if a device is linear as well as invariant Fourier analysis is then the appropriate mathematical tool for dealing with the problem.

An ensemble of functions is the appropriate mathematical representation of the messages produced by a continuous source (for example speech), of the signals produced by a transmitter, and of the perturbing noise. Communication theory is properly concerned, as has been emphasized by Wiener, not with operations on particular functions, but with operations on ensembles of functions. A communication system is designed not for a particular speech function and still less for a sine wave, but for the ensemble of speech functions.

19. Band Limited Ensembles of Functions

If a function of time $f(t)$ is limited to the band from 0 to W cycles per second it is completely determined by giving its ordinates at a series of discrete points spaced $\frac{1}{2W}$ seconds apart in the manner indicated by the following result.[5]

Theorem 13: Let $f(t)$ contain no frequencies over W.

Then

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

where

$$X_n = f\left(\frac{n}{2W}\right).$$

In this expansion $f(t)$ is represented as a sum of orthogonal functions. The coefficients X_n of the various terms can be considered as coordinates in an infinite dimensional "function space." In this space each function corresponds to precisely one point and each point to one function.

A function can be considered to be substantially limited to a time T if all the ordinates X_n outside this interval of time are zero. In this case all but $2TW$ of the coordinates will be zero. Thus functions limited to a band W and duration T correspond to points in a space of $2TW$ dimensions.

A subset of the functions of band W and duration T corresponds to a region in this space. For example, the functions whose total energy is less

[3] This is the famous ergodic theorem or rather one aspect of this theorem which was proved is somewhat different formulations by Birkhoff, von Neumann, and Koopman, and subsequently generalized by Wiener, Hopf, Hurewicz and others. The literature on ergodic theory is quite extensive and the reader is referred to the papers of these writers for precise and general formulations; e.g., E. Hopf "Ergodentheorie" Ergebnisse der Mathematic und ihrer Grenzgebiete, Vol. 5, "On Causality Statistics and Probability" Journal of Mathematics and Physics, Vol. XIII, No. 1, 1934; N. Weiner "The Ergodic Theorem" Duke Mathematical Journal, Vol. 5, 1939.

[4] Communication theory is heavily indebted to Wiener for much of its basic philosophy and theory. His classic NDRC report "The Interpolation, Extrapolation, and Smoothing of Stationary Time Series," to appear soon in book form, contains the first clear-cut formulation of communication theory as a statistical problem, the study of operations on time series. This work, although chiefly concerned with the linear prediction and filtering problem, is an important collateral reference in connection with the present paper. We may also refer here to Wiener's forthcoming book "Cybernetics" dealing with the general problems of communication and control.

[5] For a proof of this theorem and further discussion see the author's paper "Communication in the Presence of Noise" to be published in the *Proceedings of the Institute of Radio Engineers.*

than or equal to E correspond to points in a $2TW$ dimensional sphere with radius $r = \sqrt{2WE}$.

An *ensemble* of functions of limited duration and band will be represented by a probability distribution $p(x_1 \cdots x_n)$ in the corresponding n dimensional space. If the ensemble is not limited in time we can consider the $2TW$ coordinates in a given interval T to represent substantially the part of the function in the interval T and the probability distribution $p(x_1, \cdots, x_n)$ to give the statistical structure of the ensemble for intervals of that duration.

20. Entropy of a Continuous Distribution

The entropy of a discrete set of probabilities $p_1, \cdots p_n$ has been defined as:

$$H = -\sum p_i \log p_i .$$

In an analogous manner we define the entropy of a continuous distribution with the density distribution function $p(x)$ by:

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$

With an n dimensional distribution $p(x_1, \cdots, x_n)$ we have

$$H = -\int \cdots \int p(x_1 \cdots x_n) \log p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

If we have two arguments x and y (which may themselves be multi-dimensional) the joint and conditional entropies of $p(x, y)$ are given by

$$H(x, y) = -\iint p(x, y) \log p(x, y)\, dx\, dy$$

and

$$H_x(y) = -\iint p(x, y) \log \frac{p(x, y)}{p(x)}\, dx\, dy$$

$$H_y(x) = -\iint p(x, y) \log \frac{p(x, y)}{p(y)}\, dx\, dy$$

where

$$p(x) = \int p(x, y)\, dy$$

$$p(y) = \int p(x, y)\, dx.$$

The entropies of continuous distributions have most (but not all) of the properties of the discrete case. In particular we have the following:

1. If x is limited to a certain volume v in its space, then $H(x)$ is a maximum and equal to $\log v$ when $p(x)$ is constant $\left(\frac{1}{v}\right)$ in the volume.
2. With any two variables x, y we have

$$H(x, y) \leq H(x) + H(y)$$

with equality if (and only if) x and y are independent, i.e., $p(x, y) = p(x)\, p(y)$ (apart possibly from a set of points of probability zero).

3. Consider a generalized averaging operation of the following type:

$$p'(y) = \int a(x, y) p(x)\, dx$$

with

$$\int a(x, y)\, dx = \int a(x, y)\, dy = 1, \qquad a(x, y) \geq 0.$$

Then the entropy of the averaged distribution $p'(y)$ is equal to or greater than that of the original distribution $p(x)$.

4. We have

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

and

$$H_x(y) \leq H(y).$$

5. Let $p(x)$ be a one-dimensional distribution. The form of $p(x)$ giving a maximum entropy subject to the condition that the standard deviation of x be fixed at σ is gaussian. To show this we must maximize

$$H(x) = -\int p(x) \log p(x)\, dx$$

with

$$\sigma^2 = \int p(x) x^2\, dx \quad \text{and} \quad 1 = \int p(x)\, dx$$

as constraints. This requires, by the calculus of variations, maximizing

$$\int [-p(x) \log p(x) + \lambda p(x) x^2 + \mu p(x)]\, dx.$$

The condition for this is

$$-1 - \log p(x) + \lambda x^2 + \mu = 0$$

and consequently (adjusting the constants to satisfy the constraints)

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x^2/2\sigma^2)}.$$

Similarly in n dimensions, suppose the second order moments of $p(x_1, \cdots, x_n)$ are fixed at A_{ij}:

$$A_{ij} = \int \cdots \int x_i x_j p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

Then the maximum entropy occurs (by a similar calculation) when $p(x_1, \cdots, x_n)$ is the n dimensional gaussian distribution with the second order moments A_{ij}.

6. The entropy of a one-dimensional gaussian distribution whose standard deviation is σ is given by

$$H(x) = \log \sqrt{2\pi e}\,\sigma.$$

This is calculated as follows:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x^2/2\sigma^2)}$$

$$-\log p(x) = \log \sqrt{2\pi}\,\sigma + \frac{x^2}{2\sigma^2}$$

$$H(x) = -\int p(x) \log p(x)\, dx$$

$$= \int p(x) \log \sqrt{2\pi}\,\sigma\, dx + \int p(x) \frac{x^2}{2\sigma^2}\, dx$$

$$= \log \sqrt{2\pi}\,\sigma + \frac{\sigma^2}{2\sigma^2}$$

$$= \log \sqrt{2\pi}\,\sigma + \log \sqrt{e}$$

$$= \log \sqrt{2\pi e}\,\sigma.$$

Similarly the n dimensional gaussian distribution with associated quadratic form a_{ij} is given by

$$p(x_1, \cdots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\Sigma a_{ij} X_i X_j\right)$$

and the entropy can be calculated as

$$H = \log (2\pi e)^{n/2} |a_{ij}|^{\frac{1}{2}}$$

where $|a_{ij}|$ is the determinant whose elements are a_{ij}.

7. If x is limited to a half line ($p(x) = 0$ for $x \le 0$) and the first moment of x is fixed at a:

$$a = \int_0^\infty p(x) x\, dx,$$

then the maximum entropy occurs when

$$p(x) = \frac{1}{a} e^{-(x/a)}$$

and is equal to $\log ea$.

8. There is one important difference between the continuous and discrete entropies. In the discrete case the entropy measures in an *absolute* way the randomness of the chance variable. In the continuous case the measurement is *relative to the coordinate system*. If we change coordinates the entropy will in general change. In fact if we change to coordinates $y_1 \cdots y_n$ the new entropy is given by

$$H(y) = \int \cdots \int p(x_1 \cdots x_n) J\left(\frac{x}{y}\right) \log p(x_1 \cdots x_n) J\left(\frac{x}{y}\right) dy_1 \cdots dy_n$$

where $J\left(\frac{x}{y}\right)$ is the Jacobian of the coordinate transformation. On expanding the logarithm and changing variables to $x_1 \cdots x_n$, we obtain:

$$H(y) = H(x) - \int \cdots \int p(x_1, \cdots, x_n) \log J\left(\frac{x}{y}\right) dx_1 \cdots dx_n .$$

Thus the new entropy is the old entropy less the expected logarithm of the Jacobian. In the continuous case the entropy can be considered a measure of randomness *relative to an assumed standard*, namely the coordinate system chosen with each small volume element $dx_1 \cdots dx_n$ given equal weight. When we change the coordinate system the entropy in the new system measures the randomness when equal volume elements $dy_1 \cdots dy_n$ in the new system are given equal weight.

In spite of this dependence on the coordinate system the entropy concept is as important in the continuous case as the discrete case. This is due to the fact that the derived concepts of information rate and channel capacity depend on the *difference* of two entropies and this difference *does not* depend on the coordinate frame, each of the two terms being changed by the same amount.

The entropy of a continuous distribution can be negative. The scale of measurements sets an arbitrary zero corresponding to a uniform distribution over a unit volume. A distribution which is more confined than this has less entropy and will be negative. The rates and capacities will, however, always be non-negative.

9. A particular case of changing coordinates is the linear transformation

$$y_j = \sum_i a_{ij} x_i .$$

In this case the Jacobian is simply the determinant $|a_{ij}|^{-1}$ and

$$H(y) = H(x) + \log |a_{ij}|.$$

In the case of a rotation of coordinates (or any measure preserving transformation) $J = 1$ and $H(y) = H(x)$.

21. Entropy of an Ensemble of Functions

Consider an ergodic ensemble of functions limited to a certain band of width W cycles per second. Let

$$p(x_1 \cdots x_n)$$

be the density distribution function for amplitudes $x_1 \cdots x_n$ at n successive sample points. We define the entropy of the ensemble per degree of freedom by

$$H' = -\operatorname*{Lim}_{n \to \infty} \frac{1}{n} \int \cdots \int p(x_1 \cdots x_n) \log p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

We may also define an entropy H per second by dividing, not by n, but by the time T in seconds for n samples. Since $n = 2TW$, $H' = 2WH$.

With white thermal noise p is gaussian and we have

$$H' = \log \sqrt{2\pi e N},$$

$$H = W \log 2\pi e N.$$

For a given average power N, white noise has the maximum possible entropy. This follows from the maximizing properties of the Gaussian distribution noted above.

The entropy for a continuous stochastic process has many properties analogous to that for discrete processes. In the discrete case the entropy was related to the logarithm of the *probability* of long sequences, and to the *number* of reasonably probable sequences of long length. In the continuous case it is related in a similar fashion to the logarithm of the *probability density* for a long series of samples, and the *volume* of reasonably high probability in the function space.

More precisely, if we assume $p(x_1 \cdots x_n)$ continuous in all the x_i for all n, then for sufficiently large n

$$\left| \frac{\log p}{n} - H' \right| < \epsilon$$

for all choices of $(x_1, \cdots, x_n)$ apart from a set whose total probability is less than δ, with δ and ϵ arbitrarily small. This follows from the ergodic property if we divide the space into a large number of small cells.

The relation of H to volume can be stated as follows: Under the same assumptions consider the n dimensional space corresponding to $p(x_1, \cdots, x_n)$. Let $V_n(q)$ be the smallest volume in this space which includes in its interior a total probability q. Then

$$\operatorname*{Lim}_{n \to \infty} \frac{\log V_n(q)}{n} = H'$$

provided q does not equal 0 or 1.

These results show that for large n there is a rather well-defined volume (at least in the logarithmic sense) of high probability, and that within this volume the probability density is relatively uniform (again in the logarithmic sense).

In the white noise case the distribution function is given by

$$p(x_1 \cdots x_n) = \frac{1}{(2\pi N)^{n/2}} \exp - \frac{1}{2N} \Sigma x_i^2 .$$

Since this depends only on Σx_i^2 the surfaces of equal probability density are spheres and the entire distribution has spherical symmetry. The region of high probability is a sphere of radius $\sqrt{nN}$. As $n \to \infty$ the probability of being outside a sphere of radius $\sqrt{n(N + \epsilon)}$ approaches zero and $\frac{1}{n}$ times the logarithm of the volume of the sphere approaches $\log \sqrt{2\pi e N}$.

In the continuous case it is convenient to work not with the entropy H of an ensemble but with a derived quantity which we will call the entropy power. This is defined as the power in a white noise limited to the same band as the original ensemble and having the same entropy. In other words if H' is the entropy of an ensemble its entropy power is

$$N_1 = \frac{1}{2\pi e} \exp 2H'.$$

In the geometrical picture this amounts to measuring the high probability volume by the squared radius of a sphere having the same volume. Since white noise has the maximum entropy for a given power, the entropy power of any noise is less than or equal to its actual power.

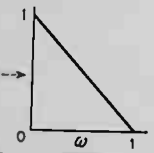
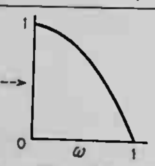
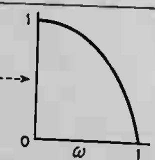
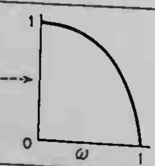
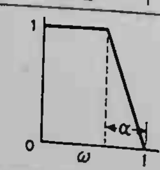
22. Entropy Loss in Linear Filters

Theorem 14: If an ensemble having an entropy H_1 per degree of freedom in band W is passed through a filter with characteristic $Y(f)$ the output ensemble has an entropy

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 \, df.$$

The operation of the filter is essentially a linear transformation of coordinates. If we think of the different frequency components as the original coordinate system, the new frequency components are merely the old ones multiplied by factors. The coordinate transformation matrix is thus es-

TABLE I

GAIN	ENTROPY POWER FACTOR	ENTROPY POWER GAIN IN DECIBELS	IMPULSE RESPONSE
$1-\omega$ 	$\frac{1}{e^2}$	-8.68	$\frac{\sin^2 \pi t}{(\pi t)^2}$
$1-\omega^2$ 	$\left(\frac{2}{e}\right)^4$	-5.32	$2\left[\frac{\sin t}{t^3} - \frac{\cos t}{t^2}\right]$
$1-\omega^3$ 	0.384	-4.15	$6\left[\frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3}\right]$
$\sqrt{1-\omega^2}$ 	$\left(\frac{2}{e}\right)^2$	-2.66	$\frac{\pi}{2} \frac{J_1(t)}{t}$
	$\frac{1}{e^{2\alpha}}$	$-8.68\,\alpha$	$\frac{1}{\alpha t^2}\left[\cos (1-\alpha)t - \cos t\right]$

sentially diagonalized in terms of these coordinates. The Jacobian of the transformation is (for n sine and n cosine components)

$$J = \prod_{i=1}^{n} | Y(f_i) |^2$$

where the f_i are equally spaced through the band W. This becomes in the limit

$$\exp \frac{1}{W} \int_W \log | Y(f) |^2 \, df.$$

Since J is constant its average value is this same quantity and applying the theorem on the change of entropy with a change of coordinates, the result follows. We may also phrase it in terms of the entropy power. Thus if the entropy power of the first ensemble is N_1 that of the second is

$$N_1 \exp \frac{1}{W} \int_W \log | Y(f) |^2 \, df.$$

The final entropy power is the initial entropy power multiplied by the geometric mean gain of the filter. If the gain is measured in *db*, then the output entropy power will be increased by the arithmetic mean *db* gain over W.

In Table I the entropy power loss has been calculated (and also expressed in *db*) for a number of ideal gain characteristics. The impulsive responses of these filters are also given for $W = 2\pi$, with phase assumed to be 0.

The entropy loss for many other cases can be obtained from these results. For example the entropy power factor $\frac{1}{e^2}$ for the first case also applies to any gain characteristic obtained from $1 - \omega$ by a measure preserving transformation of the ω axis. In particular a linearly increasing gain $G(\omega) = \omega$, or a "saw tooth" characteristic between 0 and 1 have the same entropy loss. The reciprocal gain has the reciprocal factor. Thus $\frac{1}{\omega}$ has the factor e^2. Raising the gain to any power raises the factor to this power.

23. Entropy of the Sum of Two Ensembles

If we have two ensembles of functions $f_\alpha(t)$ and $g_\beta(t)$ we can form a new ensemble by "addition." Suppose the first ensemble has the probability density function $p(x_1, \cdots, x_n)$ and the second $q(x_1, \cdots, x_n)$. Then the density function for the sum is given by the convolution:

$$r(x_1, \cdots, x_n) = \int \cdots \int p(y_1, \cdots, y_n) \cdot q(x_1 - y_1, \cdots, x_n - y_n) \, dy_1, dy_2, \cdots, dy_n .$$

Physically this corresponds to adding the noises or signals represented by the original ensembles of functions.

The following result is derived in Appendix 6.

Theorem 15: Let the average power of two ensembles be N_1 and N_2 and let their entropy powers be $\bar{N}_1$ and $\bar{N}_2$. Then the entropy power of the sum, $\bar{N}_3$, is bounded by

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2 .$$

White Gaussian noise has the peculiar property that it can absorb any other noise or signal ensemble which may be added to it with a resultant entropy power approximately equal to the sum of the white noise power and the signal power (measured from the average signal value, which is normally zero), provided the signal power is small, in a certain sense, compared to the noise.

Consider the function space associated with these ensembles having n dimensions. The white noise corresponds to a spherical Gaussian distribution in this space. The signal ensemble corresponds to another probability distribution, not necessarily Gaussian or spherical. Let the second moments of this distribution about its center of gravity be a_{ij}. That is, if $p(x_1, \cdots, x_n)$ is the density distribution function

$$a_{ij} = \int \cdots \int p(x_i - \alpha_i)(x_j - \alpha_j)\, dx_1, \cdots, dx_n$$

where the α_i are the coordinates of the center of gravity. Now a_{ij} is a positive definite quadratic form, and we can rotate our coordinate system to align it with the principal directions of this form. a_{ij} is then reduced to diagonal form b_{ii}. We require that each b_{ii} be small compared to N, the squared radius of the spherical distribution.

In this case the convolution of the noise and signal produce a Gaussian distribution whose corresponding quadratic form is

$$N + b_{ii} .$$

The entropy power of this distribution is

$$[\Pi(N + b_{ii})]^{1/n}$$

or approximately

$$= [(N)^n + \Sigma b_{ii}(N)^{n-1}]^{1/n}$$

$$\doteq N + \frac{1}{n} \Sigma b_{ii} .$$

The last term is the signal power, while the first is the noise power.

PART IV: THE CONTINUOUS CHANNEL

24. The Capacity of a Continuous Channel

In a continuous channel the input or transmitted signals will be continuous functions of time $f(t)$ belonging to a certain set, and the output or received signals will be perturbed versions of these. We will consider only the case where both transmitted and received signals are limited to a certain band W. They can then be specified, for a time T, by $2TW$ numbers, and their statistical structure by finite dimensional distribution functions. Thus the statistics of the transmitted signal will be determined by

$$P(x_1, \cdots, x_n) = P(x)$$

and those of the noise by the conditional probability distribution

$$P_{x_1, \cdots, x_n}(y_1, \cdots, y_n) = P_x(y).$$

The rate of transmission of information for a continuous channel is defined in a way analogous to that for a discrete channel, namely

$$R = H(x) - H_y(x)$$

where $H(x)$ is the entropy of the input and $H_y(x)$ the equivocation. The channel capacity C is defined as the maximum of R when we vary the input over all possible ensembles. This means that in a finite dimensional approximation we must vary $P(x) = P(x_1, \cdots, x_n)$ and maximize

$$- \int P(x) \log P(x)\, dx + \iint P(x, y) \log \frac{P(x, y)}{P(y)}\, dx\, dy.$$

This can be written

$$\iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

using the fact that $\iint P(x, y) \log P(x)\, dx\, dy = \int P(x) \log P(x)\, dx$. The channel capacity is thus expressed

$$C = \lim_{T \to \infty} \max_{P(x)} \frac{1}{T} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy.$$

It is obvious in this form that R and C are independent of the coordinate system since the numerator and denominator in $\log \frac{P(x, y)}{P(x)P(y)}$ will be multiplied by the same factors when x and y are transformed in any one to one way. This integral expression for C is more general than $H(x) - H_y(x)$. Properly interpreted (see Appendix 7) it will always exist while $H(x) - H_y(x)$

may assume an indeterminate form $\infty - \infty$ in some cases. This occurs, for example, if x is limited to a surface of fewer dimensions than n in its n dimensional approximation.

If the logarithmic base used in computing $H(x)$ and $H_y(x)$ is two then C is the maximum number of binary digits that can be sent per second over the channel with arbitrarily small equivocation, just as in the discrete case. This can be seen physically by dividing the space of signals into a large number of small cells, sufficiently small so that the probability density $P_x(y)$ of signal x being perturbed to point y is substantially constant over a cell (either of x or y). If the cells are considered as distinct points the situation is essentially the same as a discrete channel and the proofs used there will apply. But it is clear physically that this quantizing of the volume into individual points cannot in any practical situation alter the final answer significantly, provided the regions are sufficiently small. Thus the capacity will be the limit of the capacities for the discrete subdivisions and this is just the continuous capacity defined above.

On the mathematical side it can be shown first (see Appendix 7) that if u is the message, x is the signal, y is the received signal (perturbed by noise) and v the recovered message then

$$H(x) - H_y(x) \geq H(u) - H_v(u)$$

regardless of what operations are performed on u to obtain x or on y to obtain v. Thus no matter how we encode the binary digits to obtain the signal, or how we decode the received signal to recover the message, the discrete rate for the binary digits does not exceed the channel capacity we have defined. On the other hand, it is possible under very general conditions to find a coding system for transmitting binary digits at the rate C with as small an equivocation or frequency of errors as desired. This is true, for example, if, when we take a finite dimensional approximating space for the signal functions, $P(x, y)$ is continuous in both x and y except at a set of points of probability zero.

An important special case occurs when the noise is added to the signal and is independent of it (in the probability sense). Then $P_x(y)$ is a function only of the difference $n = (y - x)$,

$$P_x(y) = Q(y - x)$$

and we can assign a definite entropy to the noise (independent of the statistics of the signal), namely the entropy of the distribution $Q(n)$. This entropy will be denoted by $H(n)$.

Theorem 16: If the signal and noise are independent and the received signal is the sum of the transmitted signal and the noise then the rate of

transmission is

$$R = H(y) - H(n)$$

i.e., the entropy of the received signal less the entropy of the noise. The channel capacity is

$$C = \max_{P(x)} H(y) - H(n).$$

We have, since $y = x + n$:

$$H(x, y) = H(x, n).$$

Expanding the left side and using the fact that x and n are independent

$$H(y) + H_y(x) = H(x) + H(n).$$

Hence

$$R = H(x) - H_y(x) = H(y) - H(n).$$

Since $H(n)$ is independent of $P(x)$, maximizing R requires maximizing $H(y)$, the entropy of the received signal. If there are certain constraints on the ensemble of transmitted signals, the entropy of the received signal must be maximized subject to these constraints.

25. Channel Capacity with an Average Power Limitation

A simple application of Theorem 16 is the case where the noise is a white thermal noise and the transmitted signals are limited to a certain average power P. Then the received signals have an average power $P + N$ where N is the average noise power. The maximum entropy for the received signals occurs when they also form a white noise ensemble since this is the greatest possible entropy for a power $P + N$ and can be obtained by a suitable choice of the ensemble of transmitted signals, namely if they form a white noise ensemble of power P. The entropy (per second) of the received ensemble is then

$$H(y) = W \log 2\pi e(P + N),$$

and the noise entropy is

$$H(n) = W \log 2\pi eN.$$

The channel capacity is

$$C = H(y) - H(n) = W \log \frac{P + N}{N}.$$

Summarizing we have the following:

Theorem 17: The capacity of a channel of band W perturbed by white thermal noise of power N when the average transmitter power is P is given by

$$C = W \log \frac{P + N}{N}.$$

This means of course that by sufficiently involved encoding systems we can transmit binary digits at the rate $W \log_2 \frac{P + N}{N}$ bits per second, with arbitrarily small frequency of errors. It is not possible to transmit at a higher rate by any encoding system without a definite positive frequency of errors.

To approximate this limiting rate of transmission the transmitted signals must approximate, in statistical properties, a white noise.[6] A system which approaches the ideal rate may be described as follows: Let $M = 2^s$ samples of white noise be constructed each of duration T. These are assigned binary numbers from 0 to $(M - 1)$. At the transmitter the message sequences are broken up into groups of s and for each group the corresponding noise sample is transmitted as the signal. At the receiver the M samples are known and the actual received signal (perturbed by noise) is compared with each of them. The sample which has the least R.M.S. discrepancy from the received signal is chosen as the transmitted signal and the corresponding binary number reconstructed. This process amounts to choosing the most probable (*a posteriori*) signal. The number M of noise samples used will depend on the tolerable frequency ϵ of errors, but for almost all selections of samples we have

$$\lim_{\epsilon \to 0} \lim_{T \to \infty} \frac{\log M(\epsilon, T)}{T} = W \log \frac{P + N}{N},$$

so that no matter how small ϵ is chosen, we can, by taking T sufficiently large, transmit as near as we wish to $TW \log \frac{P + N}{N}$ binary digits in the time T.

Formulas similar to $C = W \log \frac{P + N}{N}$ for the white noise case have been developed independently by several other writers, although with somewhat different interpretations. We may mention the work of N. Wiener,[7] W. G. Tuller,[8] and H. Sullivan in this connection.

In the case of an arbitrary perturbing noise (not necessarily white thermal noise) it does not appear that the maximizing problem involved in determining the channel capacity C can be solved explicitly. However, upper and lower bounds can be set for C in terms of the average noise power N and the noise entropy power N_1. These bounds are sufficiently close together in most practical cases to furnish a satisfactory solution to the problem.

[6] This and other properties of the white noise case are discussed from the geometrical point of view in "Communication in the Presence of Noise," loc. cit.

[7] "Cybernetics," loc. cit.

[8] Sc. D. thesis, Department of Electrical Engineering, M.I.T., 1948

Theorem 18: The capacity of a channel of band W perturbed by an arbitrary noise is bounded by the inequalities

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

where

P = average transmitter power

N = average noise power

N_1 = entropy power of the noise.

Here again the average power of the perturbed signals will be $P + N$. The maximum entropy for this power would occur if the received signal were white noise and would be $W \log 2\pi e(P + N)$. It may not be possible to achieve this; i.e. there may not be any ensemble of transmitted signals which, added to the perturbing noise, produce a white thermal noise at the receiver, but at least this sets an upper bound to $H(y)$. We have, therefore

$$\begin{aligned} C &= \max H(y) - H(n) \\ &\leq W \log 2\pi e(P + N) - W \log 2\pi e N_1. \end{aligned}$$

This is the upper limit given in the theorem. The lower limit can be obtained by considering the rate if we make the transmitted signal a white noise, of power P. In this case the entropy power of the received signal must be at least as great as that of a white noise of power $P + N_1$ since we have shown in a previous theorem that the entropy power of the sum of two ensembles is greater than or equal to the sum of the individual entropy powers. Hence

$$\max H(y) \geq W \log 2\pi e(P + N_1)$$

and

$$\begin{aligned} C &\geq W \log 2\pi e(P + N_1) - W \log 2\pi e N_1 \\ &= W \log \frac{P + N_1}{N_1}. \end{aligned}$$

As P increases, the upper and lower bounds approach each other, so we have as an asymptotic rate

$$W \log \frac{P + N}{N_1}$$

If the noise is itself white, $N = N_1$ and the result reduces to the formula proved previously:

$$C = W \log \left(1 + \frac{P}{N}\right).$$

If the noise is Gaussian but with a spectrum which is not necessarily flat, N_1 is the geometric mean of the noise power over the various frequencies in the band W. Thus

$$N_1 = \exp \frac{1}{W} \int_W \log N(f)\, df$$

where $N(f)$ is the noise power at frequency f.

Theorem 19: If we set the capacity for a given transmitter power P equal to

$$C = W \log \frac{P + N - \eta}{N_1}$$

then η is monotonic decreasing as P increases and approaches 0 as a limit.

Suppose that for a given power P_1 the channel capacity is

$$W \log \frac{P_1 + N - \eta_1}{N_1}$$

This means that the best signal distribution, say $p(x)$, when added to the noise distribution $q(x)$, gives a received distribution $r(y)$ whose entropy power is $(P_1 + N - \eta_1)$. Let us increase the power to $P_1 + \Delta P$ by adding a white noise of power ΔP to the signal. The entropy of the received signal is now at least

$$H(y) = W \log 2\pi e(P_1 + N - \eta_1 + \Delta P)$$

by application of the theorem on the minimum entropy power of a sum. Hence, since we can attain the H indicated, the entropy of the maximizing distribution must be at least as great and η must be monotonic decreasing. To show that $\eta \to 0$ as $P \to \infty$ consider a signal which is a white noise with a large P. Whatever the perturbing noise, the received signal will be approximately a white noise, if P is sufficiently large, in the sense of having an entropy power approaching $P + N$.

26. The Channel Capacity with a Peak Power Limitation

In some applications the transmitter is limited not by the average power output but by the peak instantaneous power. The problem of calculating the channel capacity is then that of maximizing (by variation of the ensemble of transmitted symbols)

$$H(y) - H(n)$$

subject to the constraint that all the functions $f(t)$ in the ensemble be less than or equal to $\sqrt{S}$, say, for all t. A constraint of this type does not work out as well mathematically as the average power limitation. The most we have obtained for this case is a lower bound valid for all $\frac{S}{N}$, an "asymptotic" upper band $\left(\text{valid for large } \frac{S}{N}\right)$ and an asymptotic value of C for $\frac{S}{N}$ small.

Theorem 20: The channel capacity C for a band W perturbed by white thermal noise of power N is bounded by

$$C \geq W \log \frac{2}{\pi e^3} \frac{S}{N},$$

where S is the peak allowed transmitter power. For sufficiently large $\frac{S}{N}$

$$C \leq W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon)$$

where ϵ is arbitrarily small. As $\frac{S}{N} \to 0$ (and provided the band W starts at 0)

$$C \to W \log \left(1 + \frac{S}{N}\right).$$

We wish to maximize the entropy of the received signal. If $\frac{S}{N}$ is large this will occur very nearly when we maximize the entropy of the transmitted ensemble.

The asymptotic upper bound is obtained by relaxing the conditions on the ensemble. Let us suppose that the power is limited to S not at every instant of time, but only at the sample points. The maximum entropy of the transmitted ensemble under these weakened conditions is certainly greater than or equal to that under the original conditions. This altered problem can be solved easily. The maximum entropy occurs if the different samples are independent and have a distribution function which is constant from $-\sqrt{S}$ to $+\sqrt{S}$. The entropy can be calculated as

$$W \log 4S.$$

The received signal will then have an entropy less than

$$W \log (4S + 2\pi e N)(1 + \epsilon)$$

with $\epsilon \to 0$ as $\frac{S}{N} \to \infty$ and the channel capacity is obtained by subtracting the entropy of the white noise, $W \log 2\pi e N$

$$W \log (4S + 2\pi e N)(1 + \epsilon) - W \log (2\pi e N) = W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon).$$

This is the desired upper bound to the channel capacity.

To obtain a lower bound consider the same ensemble of functions. Let these functions be passed through an ideal filter with a triangular transfer characteristic. The gain is to be unity at frequency 0 and decline linearly down to gain 0 at frequency W. We first show that the output functions of the filter have a peak power limitation S at all times (not just the sample points). First we note that a pulse $\frac{\sin 2\pi W t}{2\pi W t}$ going into the filter produces

$$\frac{1}{2} \frac{\sin^2 \pi W t}{(\pi W t)^2}$$

in the output. This function is never negative. The input function (in the general case) can be thought of as the sum of a series of shifted functions

$$a \frac{\sin 2\pi W t}{2\pi W t}$$

where a, the amplitude of the sample, is not greater than $\sqrt{S}$. Hence the output is the sum of shifted functions of the non-negative form above with the same coefficients. These functions being non-negative, the greatest positive value for any t is obtained when all the coefficients a have their maximum positive values, i.e. $\sqrt{S}$. In this case the input function was a constant of amplitude $\sqrt{S}$ and since the filter has unit gain for D.C., the output is the same. Hence the output ensemble has a peak power S.

The entropy of the output ensemble can be calculated from that of the input ensemble by using the theorem dealing with such a situation. The output entropy is equal to the input entropy plus the geometrical mean gain of the filter;

$$\int_0^W \log G^2 \, df = \int_0^W \log \left(\frac{W - f}{W} \right)^2 df = -2W$$

Hence the output entropy is

$$W \log 4S - 2W = W \log \frac{4S}{e^2}$$

and the channel capacity is greater than

$$W \log \frac{2}{\pi e^3} \frac{S}{N}.$$

We now wish to show that, for small $\frac{S}{N}$ (peak signal power over average white noise power), the channel capacity is approximately

$$C = W \log \left(1 + \frac{S}{N} \right).$$

More precisely $C/W \log \left(1 + \frac{S}{N} \right) \to 1$ as $\frac{S}{N} \to 0$. Since the average signal power P is less than or equal to the peak S, it follows that for all $\frac{S}{N}$

$$C \le W \log \left(1 + \frac{P}{N} \right) \le W \log \left(1 + \frac{S}{N} \right).$$

Therefore, if we can find an ensemble of functions such that they correspond to a rate nearly $W \log \left(1 + \frac{S}{N} \right)$ and are limited to band W and peak S the result will be proved. Consider the ensemble of functions of the following type. A series of t samples have the same value, either $+\sqrt{S}$ or $-\sqrt{S}$, then the next t samples have the same value, etc. The value for a series is chosen at random, probability $\frac{1}{2}$ for $+\sqrt{S}$ and $\frac{1}{2}$ for $-\sqrt{S}$. If this ensemble be passed through a filter with triangular gain characteristic (unit gain at D.C.), the output is peak limited to $\pm S$. Furthermore the average power is nearly S and can be made to approach this by taking t sufficiently large. The entropy of the sum of this and the thermal noise can be found by applying the theorem on the sum of a noise and a small signal. This theorem will apply if

$$\sqrt{t} \frac{S}{N}$$

is sufficiently small. This can be insured by taking $\frac{S}{N}$ small enough (after t is chosen). The entropy power will be $S + N$ to as close an approximation as desired, and hence the rate of transmission as near as we wish to

$$W \log \left(\frac{S + N}{N} \right).$$

PART V: THE RATE FOR A CONTINUOUS SOURCE

27. Fidelity Evaluation Functions

In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with *exact recovery* at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. Of course, as the fidelity requirements are increased the rate will increase. It will be shown that we can, in very general cases, define such a rate, having the property that it is possible, by properly encoding the information, to transmit it over a channel whose capacity is equal to the rate in question, and satisfy the fidelity requirements. A channel of smaller capacity is insufficient.

It is first necessary to give a general mathematical formulation of the idea of fidelity of transmission. Consider the set of messages of a long duration, say T seconds. The source is described by giving the probability density, in the associated space, that the source will select the message in question $P(x)$. A given communication system is described (from the external point of view) by giving the conditional probability $P_x(y)$ that if message x is produced by the source the recovered message at the receiving point will be y. The system as a whole (including source and transmission system) is described by the probability function $P(x, y)$ of having message x and final output y. If this function is known, the complete characteristics of the system from the point of view of fidelity are known. Any evaluation of fidelity must correspond mathematically to an operation applied to $P(x, y)$. This operation must at least have the properties of a simple ordering of systems; i.e. it must be possible to say of two systems represented by $P_1(x, y)$ and $P_2(x, y)$ that, according to our fidelity criterion, either (1) the first has higher fidelity, (2) the second has higher fidelity, or (3) they have

equal fidelity. This means that a criterion of fidelity can be represented by a numerically valued function:

$$v(P(x, y))$$

whose argument ranges over possible probability functions $P(x, y)$.

We will now show that under very general and reasonable assumptions the function $v(P(x, y))$ can be written in a seemingly much more specialized form, namely as an average of a function $\rho(x, y)$ over the set of possible values of x and y:

$$v(P(x, y)) = \iint P(x, y)\, \rho(x, y)\, dx\, dy$$

To obtain this we need only assume (1) that the source and system are ergodic so that a very long sample will be, with probability nearly 1, typical of the ensemble, and (2) that the evaluation is "reasonable" in the sense that it is possible, by observing a typical input and output x_1 and y_1, to form a tentative evaluation on the basis of these samples; and if these samples are increased in duration the tentative evaluation will, with probability 1, approach the exact evaluation based on a full knowledge of $P(x, y)$. Let the tentative evaluation be $\rho(x, y)$. Then the function $\rho(x, y)$ approaches (as $T \to \infty$) a constant for almost all (x, y) which are in the high probability region corresponding to the system:

$$\rho(x, y) \to v(P(x, y))$$

and we may also write

$$\rho(x, y) \to \iint P(x, y)\rho(x, y)\, dx,\, dy$$

since

$$\iint P(x, y)\, dx\, dy = 1$$

This establishes the desired result.

The function $\rho(x, y)$ has the general nature of a "distance" between x and y.[9] It measures how bad it is (according to our fidelity criterion) to receive y when x is transmitted. The general result given above can be restated as follows: Any reasonable evaluation can be represented as an average of a distance function over the set of messages and recovered messages x and y weighted according to the probability $P(x, y)$ of getting the pair in question, provided the duration T of the messages be taken sufficiently large.

[9] It is not a "metric" in the strict sense, however, since in general it does not satisfy either $\rho(x, y) = \rho(y, x)$ or $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

The following are simple examples of evaluation functions:

1. R.M.S. Criterion.

$$v = \overline{(x(t) - y(t))^2}$$

In this very commonly used criterion of fidelity the distance function $\rho(x, y)$ is (apart from a constant factor) the square of the ordinary euclidean distance between the points x and y in the associated function space.

$$\rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 \, dt$$

2. Frequency weighted R.M.S. criterion. More generally one can apply different weights to the different frequency components before using an R.M.S. measure of fidelity. This is equivalent to passing the difference $x(t) - y(t)$ through a shaping filter and then determining the average power in the output. Thus let

$$e(t) = x(t) - y(t)$$

and

$$f(t) = \int_{-\infty}^{\infty} e(\tau)k(t - \tau) \, d\tau$$

then

$$\rho(x, y) = \frac{1}{T} \int_0^T f(t)^2 \, dt.$$

3. Absolute error criterion.

$$\rho(x, y) = \frac{1}{T} \int_0^T | x(t) - y(t) | \, dt$$

4. The structure of the ear and brain determine implicitly an evaluation, or rather a number of evaluations, appropriate in the case of speech or music transmission. There is, for example, an "intelligibility" criterion in which $\rho(x, y)$ is equal to the relative frequency of incorrectly interpreted words when message $x(t)$ is received as $y(t)$. Although we cannot give an explicit representation of $\rho(x, y)$ in these cases it could, in principle, be determined by sufficient experimentation. Some of its properties follow from well-known experimental results in hearing, e.g., the ear is relatively insensitive to phase and the sensitivity to amplitude and frequency is roughly logarithmic.
5. The discrete case can be considered as a specialization in which we have tacitly assumed an evaluation based on the frequency of errors. The function $\rho(x, y)$ is then defined as the number of symbols in the sequence y differing from the corresponding symbols in x divided by the total number of symbols in x.

28. The Rate for a Source Relative to a Fidelity Evaluation

We are now in a position to define a rate of generating information for a continuous source. We are given $P(x)$ for the source and an evaluation v determined by a distance function $\rho(x, y)$ which will be assumed continuous in both x and y. With a particular system $P(x, y)$ the quality is measured by

$$v = \iint \rho(x, y) \, P(x, y) \, dx \, dy$$

Furthermore the rate of flow of binary digits corresponding to $P(x, y)$ is

$$R = \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy.$$

We define the rate R_1 of generating information for a given quality v_1 of reproduction to be the minimum of R when we keep v fixed at v_1 and vary $P_x(y)$. That is:

$$R_1 = \operatorname*{Min}_{P_x(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy$$

subject to the constraint:

$$v_1 = \iint P(x, y)\rho(x, y) \, dx \, dy.$$

This means that we consider, in effect, all the communication systems that might be used and that transmit with the required fidelity. The rate of transmission in bits per second is calculated for each one and we choose that having the least rate. This latter rate is the rate we assign the source for the fidelity in question.

The justification of this definition lies in the following result:

Theorem 21: If a source has a rate R_1 for a valuation v_1 it is possible to encode the output of the source and transmit it over a channel of capacity C with fidelity as near v_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.

The last statement in the theorem follows immediately from the definition of R_1 and previous results. If it were not true we could transmit more than C bits per second over a channel of capacity C. The first part of the theorem is proved by a method analogous to that used for Theorem 11. We may, in the first place, divide the (x, y) space into a large number of small cells and represent the situation as a discrete case. This will not change the evaluation function by more than an arbitrarily small amount (when the cells are very small) because of the continuity assumed for $\rho(x, y)$. Suppose that $P_1(x, y)$ is the particular system which minimizes the rate and gives R_1. We choose from the high probability y's a set at random containing

$$2^{(R_1 + \epsilon)T}$$

members where $\epsilon \to 0$ as $T \to \infty$. With large T each chosen point will be connected by a high probability line (as in Fig. 10) to a set of x's. A calculation similar to that used in proving Theorem 11 shows that with large T almost all x's are covered by the fans from the chosen y points for almost all choices of the y's. The communication system to be used operates as follows: The selected points are assigned binary numbers. When a message x is originated it will (with probability approaching 1 as $T \to \infty$) lie within one at least of the fans. The corresponding binary number is transmitted (or one of them chosen arbitrarily if there are several) over the channel by suitable coding means to give a small probability of error. Since $R_1 \le C$ this is possible. At the receiving point the corresponding y is reconstructed and used as the recovered message.

The evaluation v_1' for this system can be made arbitrarily close to v_1 by taking T sufficiently large. This is due to the fact that for each long sample of message $x(t)$ and recovered message $y(t)$ the evaluation approaches v_1 (with probability 1).

It is interesting to note that, in this system, the noise in the recovered message is actually produced by a kind of general quantizing at the transmitter and is not produced by the noise in the channel. It is more or less analogous to the quantizing noise in P.C.M.

29. The Calculation of Rates

The definition of the rate is similar in many respects to the definition of channel capacity. In the former

$$R = \min_{P_x(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

with $P(x)$ and $v_1 = \iint P(x, y)\rho(x, y)\, dx\, dy$ fixed. In the latter

$$C = \max_{P(x)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

with $P_x(y)$ fixed and possibly one or more other constraints (e.g., an average power limitation) of the form $K = \iint P(x, y)\, \lambda(x, y)\, dx\, dy$.

A partial solution of the general maximizing problem for determining the rate of a source can be given. Using Lagrange's method we consider

$$\iint \left[P(x, y) \log \frac{P(x, y)}{P(x)P(y)} + \mu\, P(x, y)\rho(x, y) + \nu(x)P(x, y) \right] dx\, dy$$

The variational equation (when we take the first variation on $P(x, y)$) leads to

$$P_y(x) = B(x)\, e^{-\lambda\rho(x, y)}$$

where λ is determined to give the required fidelity and $B(x)$ is chosen to satisfy

$$\int B(x) e^{-\lambda\rho(x, y)}\, dx = 1$$

This shows that, with best encoding, the conditional probability of a certain cause for various received y, $P_y(x)$ will decline exponentially with the distance function $\rho(x, y)$ between the x and y is question.

In the special case where the distance function $\rho(x, y)$ depends only on the (vector) difference between x and y,

$$\rho(x, y) = \rho(x - y)$$

we have

$$\int B(x) e^{-\lambda\rho(x - y)}\, dx = 1.$$

Hence $B(x)$ is constant, say α, and

$$P_y(x) = \alpha e^{-\lambda\rho(x - y)}$$

Unfortunately these formal solutions are difficult to evaluate in particular cases and seem to be of little value. In fact, the actual calculation of rates has been carried out in only a few very simple cases.

If the distance function $\rho(x, y)$ is the mean square discrepancy between x and y and the message ensemble is white noise, the rate can be determined. In that case we have

$$R = \text{Min}\, [H(x) - H_y(x)] = H(x) - \text{Max}\, H_y(x)$$

with $N = \overline{(x - y)^2}$. But the Max $H_y(x)$ occurs when $y - x$ is a white noise, and is equal to $W_1 \log 2\pi e\, N$ where W_1 is the bandwidth of the message ensemble. Therefore

$$R = W_1 \log 2\pi e Q - W_1 \log 2\pi e N$$
$$= W_1 \log \frac{Q}{N}$$

where Q is the average message power. This proves the following:

Theorem 22: The rate for a white noise source of power Q and band W_1 relative to an R.M.S. measure of fidelity is

$$R = W_1 \log \frac{Q}{N}$$

where N is the allowed mean square error between original and recovered messages.

More generally with any message source we can obtain inequalities bounding the rate relative to a mean square error criterion.

Theorem 23: The rate for any source of band W_1 is bounded by

$$W_1 \log \frac{Q_1}{N} \leq R \leq W_1 \log \frac{Q}{N}$$

where Q is the average power of the source, Q_1 its entropy power and N the allowed mean square error.

The lower bound follows from the fact that the max $H_y(x)$ for a given $\overline{(x-y)^2} = N$ occurs in the white noise case. The upper bound results if we place the points (used in the proof of Theorem 21) not in the best way but at random in a sphere of radius $\sqrt{Q - N}$.

Acknowledgments

The writer is indebted to his colleagues at the Laboratories, particularly to Dr. H. W. Bode, Dr. J. R. Pierce, Dr. B. McMillan, and Dr. B. M. Oliver for many helpful suggestions and criticisms during the course of this work. Credit should also be given to Professor N. Wiener, whose elegant solution of the problems of filtering and prediction of stationary ensembles has considerably influenced the writer's thinking in this field.

APPENDIX 5

Let S_1 be any measurable subset of the g ensemble, and S_2 the subset of the f ensemble which gives S_1 under the operation T. Then

$$S_1 = TS_2.$$

Let H^λ be the operator which shifts all functions in a set by the time λ. Then

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2$$

since T is invariant and therefore commutes with H^λ. Hence if $m[S]$ is the probability measure of the set S

$$m[H^\lambda S_1] = m[TH^\lambda S_2] = m[H^\lambda S_2] = m[S_2] = m[S_1]$$

where the second equality is by definition of measure in the g space the third since the f ensemble is stationary, and the last by definition of g measure again.

To prove that the ergodic property is preserved under invariant operations, let S_1 be a subset of the g ensemble which is invariant under H^λ, and let S_2 be the set of all functions f which transform into S_1. Then

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2 = S_1$$

so that $H^\lambda S_1$ is included in S_1 for all λ. Now, since

$$m[H^\lambda S_2] = m[S_1]$$

this implies

$$H^\lambda S_2 = S_2$$

for all λ with $m[S_2] \neq 0, 1$. This contradiction shows that S_1 does not exist.

APPENDIX 6

The upper bound, $\bar{N}_3 \leq N_1 + N_2$, is due to the fact that the maximum possible entropy for a power $N_1 + N_2$ occurs when we have a white noise of this power. In this case the entropy power is $N_1 + N_2$.

To obtain the lower bound, suppose we have two distributions in n dimensions $p(x_i)$ and $q(x_i)$ with entropy powers $\bar{N}_1$ and $\bar{N}_2$. What form should p and q have to minimize the entropy power $\bar{N}_3$ of their convolution $r(x_i)$:

$$r(x_i) = \int p(y_i) q(x_i - y_i)\, dy_i .$$

The entropy H_3 of r is given by

$$H_3 = -\int r(x_i) \log r(x_i)\, dx_i .$$

We wish to minimize this subject to the constraints

$$H_1 = -\int p(x_i) \log p(x_i)\, dx_i$$

$$H_2 = -\int q(x_i) \log q(x_i)\, dx_i .$$

We consider then

$$U = -\int [r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)]\, dx$$

$$\delta U = -\int [[1 + \log r(x)]\delta r(x) + \lambda[1 + \log p(x)]\delta p(x) + \mu[1 + \log q(x)\delta q(x)]]\, dx.$$

If $p(x)$ is varied at a particular argument $x_i = s_i$, the variation in $r(x)$ is

$$\delta r(x) = q(x_i - s_i)$$

and

$$\delta U = -\int q(x_i - s_i) \log r(x_i)\, dx_i - \lambda \log p(s_i) = 0$$

and similarly when q is varied. Hence the conditions for a minimum are

$$\int q(x_i - s_i) \log r(x_i) = -\lambda \log p(s_i)$$

$$\int p(x_i - s_i) \log r(x_i) = -\mu \log q(s_i).$$

If we multiply the first by $p(s_i)$ and the second by $q(s_i)$ and integrate with respect to s we obtain

$$H_3 = -\lambda H_1$$

$$H_3 = -\mu H_2$$

or solving for λ and μ and replacing in the equations

$$H_1 \int q(x_i - s_i) \log r(x_i)\, dx_i = -H_3 \log p(s_i)$$

$$H_2 \int p(x_i - s_i) \log r(x_i)\, dx_i = -H_3 \log p(s_i).$$

Now suppose $p(x_i)$ and $q(x_i)$ are normal

$$p(x_i) = \frac{|A_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\tfrac{1}{2}\Sigma A_{ij} x_i x_j$$

$$q(x_i) = \frac{|B_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\tfrac{1}{2}\Sigma B_{ij} x_i x_j .$$

Then $r(x_i)$ will also be normal with quadratic form C_{ij}. If the inverses of these forms are a_{ij}, b_{ij}, c_{ij} then

$$c_{ij} = a_{ij} + b_{ij}.$$

We wish to show that these functions satisfy the minimizing conditions if and only if $a_{ij} = Kb_{ij}$ and thus give the minimum H_3 under the constraints. First we have

$$\log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \tfrac{1}{2}\Sigma C_{ij} x_i x_j$$

$$\int q(x_i - s_i) \log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \tfrac{1}{2}\Sigma C_{ij} s_i s_j - \tfrac{1}{2}\Sigma C_{ij} b_{ij} .$$

This should equal

$$\frac{H_3}{H_1}\left[\frac{n}{2} \log \frac{1}{2\pi} |A_{ij}| - \tfrac{1}{2}\Sigma A_{ij} s_i s_j\right]$$

which requires $A_{ij} = \frac{H_1}{H_3} C_{ij}$.

In this case $A_{ij} = \frac{H_1}{H_2} B_{ij}$ and both equations reduce to identities.

APPENDIX 7

The following will indicate a more general and more rigorous approach to the central definitions of communication theory. Consider a probability measure space whose elements are ordered pairs (x, y). The variables x, y are to be identified as the possible transmitted and received signals of some long duration T. Let us call the set of all points whose x belongs to a subset S_1 of x points the strip over S_1, and similarly the set whose y belongs to S_2 the strip over S_2. We divide x and y into a collection of non-overlapping measurable subsets X_i and Y_i approximate to the rate of transmission R by

$$R_1 = \frac{1}{T} \sum_i P(X_i, Y_i) \log \frac{P(X_i, Y_i)}{P(X_i)P(Y_i)}$$

where

$P(X_i)$ is the probability measure of the strip over X_i

$P(Y_i)$ is the probability measure of the strip over Y_i

$P(X_i, Y_i)$ is the probability measure of the intersection of the strips.

A further subdivision can never decrease R_1. For let X_1 be divided into $X_1 = X_1' + X_1''$ and let

$$P(Y_1) = a \qquad P(X_1) = b + c$$

$$P(X_1') = b \qquad P(X_1', Y_1) = d$$

$$P(X_1'') = c \qquad P(X_1'', Y_1) = e$$

$$P(X_1, Y_1) = d + e$$

Then in the sum we have replaced (for the X_1, Y_1 intersection)

$$(d + e) \log \frac{d + e}{a(b + c)} \quad \text{by} \quad d \log \frac{d}{ab} + e \log \frac{e}{ac}.$$

It is easily shown that with the limitation we have on b, c, d, e,

$$\left[\frac{d + e}{b + c}\right]^{d+e} \le \frac{d^d e^e}{b^d c^e}$$

and consequently the sum is increased. Thus the various possible subdivisions form a directed set, with R monotonic increasing with refinement of the subdivision. We may define R unambiguously as the least upper bound for the R_1 and write it

$$R = \frac{1}{T} \int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy.$$

This integral, understood in the above sense, includes both the continuous and discrete cases and of course many others which cannot be represented in either form. It is trivial in this formulation that if x and u are in one-to-one correspondence, the rate from u to y is equal to that from x to y. If v is any function of y (not necessarily with an inverse) then the rate from x to y is greater than or equal to that from x to v since, in the calculation of the approximations, the subdivisions of y are essentially a finer subdivision of those for v. More generally if y and v are related not functionally but statistically, i.e., we have a probability measure space (y, v), then $R(x, v) \leq R(x, y)$. This means that any operation applied to the received signal, even though it involves statistical elements, does not increase R.

Another notion which should be defined precisely in an abstract formulation of the theory is that of "dimension rate," that is the average number of dimensions required per second to specify a member of an ensemble. In the band limited case $2W$ numbers per second are sufficient. A general definition can be framed as follows. Let $f_\alpha(t)$ be an ensemble of functions and let $\rho_T[f_\alpha(t), f_\beta(t)]$ be a metric measuring the "distance" from f_α to f_β over the time T (for example the R.M.S. discrepancy over this interval.) Let $N(\epsilon, \delta, T)$ be the least number of elements f which can be chosen such that all elements of the ensemble apart from a set of measure δ are within the distance ϵ of at least one of those chosen. Thus we are covering the space to within ϵ apart from a set of small measure δ. We define the dimension rate λ for the ensemble by the triple limit

$$\lambda = \lim_{\delta \to 0} \lim_{\epsilon \to 0} \lim_{T \to \infty} \frac{\log N(\epsilon, \delta, T)}{T \log \epsilon}.$$

This is a generalization of the measure type definitions of dimension in topology, and agrees with the intuitive dimension rate for simple ensembles where the desired result is obvious.

B-1598: C. E. SHANNON

BELL TELEPHONE LABORATORIES, *incorporated*

463 West Street, New York 14, N. Y.

AMERICAN SCIENTIST

SPRING ISSUE

VOLUME 47 MARCH 1959. NUMBER 1

CONTENTS

AMERICAN SCIENTIST

PUBLISHED IN THE INTEREST OF SCIENTIFIC RESEARCH

AT TWENTIETH AND NORTHAMPTON STREETS, EASTON, PENNSYLVANIA

VOLUME 47 — MARCH 1959 — NUMBER 1

THE AMERICAN SCIENTIST is published quarterly (March, June, September, December) by the Society of the Sigma Xi at the Mack Printing Co., 20th and Northampton Sts., Easton, Pa.

BOARD OF EDITORS
P. O. Box 703, Princeton, N. J.

Emily C. Gillispie, *Secretary*

Second-class mail privileges authorized at Easton, Pa. Additional entry at the New Haven, Conn., post office, December 9, 1943. This publication is authorized to be mailed at the special rates of postage prescribed by section 132.122.

EDITORIAL CORRESPONDENCE and publications for review should be addressed to P.O. Box 703, Princeton, N. J. The submission of manuscripts is welcomed but no responsibility for safe keeping or return can be assumed.

ADVERTISING CORRESPONDENCE should be addressed to the Advertising Representative, Mrs. William A. James, 217 Canner Street, New Haven, Conn.

CHANGE OF ADDRESS should be sent to: Sigma Xi, 56 Hillhouse Avenue, New Haven 11, Conn. In order to avoid the loss of an issue, send such notice promptly, giving full name, previous address, local or national affiliation, as well as the new address. Claims for a lost issue must be made within 60 days of publication.

SUBSCRIPTION CORRESPONDENCE should be addressed to: Sigma Xi, 56 Hillhouse Avenue, New Haven 11, Conn. All checks should be made payable to Sigma Xi.

Annual Subscription, $2.50

Single Copy, 75¢. Foreign Subscription, $3.00.

CONTENTS *and* CONTRIBUTORS

The motions of stars and atoms are better known than the internal motions of our own planet. Here is a lucid and scientifically sound attempt to rectify this state of our knowledge. Projecting experiments hundreds of miles into the earth's interior and formulating ideas about this viscous elastic body we live on is the fascinating theme of this article. The author is Professor of Geophysics at the University of Toronto, Toronto, Canada.

A Professor of Physics in Princeton University has permitted the American Scientist to reproduce a Joseph Henry Lecture on a subject of perennial interest to all thoughtful scientists who contemplate what holds the cosmos together.

The experimental and the theoretical aspects of any science have their ups and downs. Professor Lanczos presents, with fine historical perspective, the part played by Albert Einstein and his predecessors in these internecine interludes. This article was delivered as a lecture to the Sigma Xi Chapter of Oregon State College, Corvallis, Oregon, in late 1957. Professor Lanczos, born and raised in Hungary, has, since 1954, been Senior Professor in the School of Theoretical Physics of the Dublin Institute for Advanced Studies, Dublin, Eire.

Page 60

GUY SUITS, *Education and Science*

The Vice-President and Director of Research in the General Electric Company at Schenectady, New York, considers, in his Procter Award Address, some of the problems of education and science as they affect the producers and consumers of the educated scientist.

Page 68

JAMES C. O'FLAHERTY, *A Humanist Looks at Max Planck*

The author is a professor of German who had the opportunity of examining the diaries and letters of the late Max Planck and of talking with his widow. The article touches a major question of our time: how does man the scientist face unavoidable moral issues? Professor O'Flaherty examines the actions of an individual scientist caught in the terror and retrogression of the ugliest episode of our century. This record of will to resist evil is a most important document and a message of hope for the future. Originally it was a Beecher Lecture delivered at Amherst College, Amherst, Massachusetts, in April 1958. Address: Dept. of Modern Languages, Wake Forest College, Winston-Salem, North Carolina.

Page 80

CARL R. EKLUND and FREDERICK E. CHARLTON, *Measuring the Temperatures of Incubating Penguin Eggs*

Dr. Eklund has had much experience as a Wildlife Research Biologist, *inter alia* with Admiral Byrd in Antarctica, 1939–1941, and as IGY Wilkes Station Scientific Leader in 1957–1958. He is now with the Polar Division, Army Research Office, Pentagon, Washington 25, D. C. His associate in this research is Frederick E. Charlton of Route 1, Mt. Vernon, Washington, D. C. We are indebted to the Editor of "Research Reviews," A. T. Drury of the Office of Naval Research, for permission to reprint this interesting report of an unusual scientific experiment in a novel out-of-doors setting.

Page 87

JOHN D. TSILIKIS, *Simplicity and Elegance in Theoretical Physics*

The requirements for a scientific law to be acceptable are formidable: not only must it correlate known facts, but it, also, should predict new ones. Further, there is a more subtle requirement: it should satisfy our esthetic sense. The latter is here discussed by an exchange student from Greece, the birthplace of simplicity and elegance, now studying at the University of Michigan. Address: 439 S. Division St., Ann Arbor, Michigan.

Page 97

JOHN TURKEVICH, *The World of Fine Particles*

The Eugene Higgins Professor of Chemistry in Princeton University is at home both in the exploration of fine particles and, by reason of his intimate knowledge of things Russian, in service to his country, across the Continent, at Geneva, in Moscow. This Sigma Xi-RESA lecture brings to a wider audience an address enthusiastically received by Chapters and Clubs.

BOOKS RECEIVED

The Stars by IRVING ADLER; 126 pages; \$0.35; New American Library, 1958.

Quantitative Chemical Analysis by GILBERT H. AYRES; 726 pages; \$7.50; Harper & Brothers, 1958.

The Science of Photography—Written for the Layman by H. BAINES; 319 pages; \$7.50; John Wiley & Sons, 1958.

Schizophrenia: A Review of the Syndrome, edited by LEOPOLD BELLAK; 1010 pages; \$14.75; Logos Press, 1958.

The Oil Century: From the Drake Well to the Conservation Era by J. STANLEY CLARK; 280 pages; \$3.95; University of Oklahoma Press, 1958.

Physiology of Fungi by VINCENT W. COCHRANE; 524 pages; \$9.75; John Wiley & Sons, 1958.

Serological and Biochemical Comparisons of Proteins, edited by WM. H. COLE, 119 pages; \$2.00; Rutgers University Press, 1958.

Growth and Perfection of Crystals by R. H. DOREMUS, B. W. ROBERTS, and DAVID TURNBULL; 609 pages; \$12.50; John Wiley & Sons, 1958.

(*Continued on page 144*)

NEWS AND VIEWS

By the Board of Editors and the Membership of the Sigma XI-RESA Societies

THE SCIENTIFIC RESEARCH SOCIETY OF AMERICA: PROCEEDINGS OF THE TENTH ANNUAL CONVENTION, WASHINGTON, D.C.

DECEMBER 29, 1958

The tenth Annual Convention of the Scientific Research Society of America was held in the Continental Room of The Sheraton-Park Hotel, Washington, D.C. on December 29, 1958. The Chairman, Dr. Edward R. Weidlein, called the meeting to order at 10:35 A.M.

The roll call, checked by replies from branch officers received in return from the formal notice of the Convention, showed the following delegates present:

Air Force Cambridge Research Center: Edwin Kessler; ALCOA Branch: G. B. Todd; Argonne National Laboratory Branch: Leon M. Dorfman; Army Chemical Center Branch: Dorothy Bergner and Bernard Jandorff; ASARCO Branch: Max Hollander; Bakelite Branch, Union Carbide Company: Mark Stanek; Communicable Disease Center Branch: Robert Ellis; Engineer Research and Development Branch: Z. V. Harvalik, L. W. Ames and Herbert Mueller; Esso Research Club Branch: Robert F. Leary; Fort Detrick Branch: Robert L. Weintraub; Frankford Arsenal Branch: Henry Lipinski and Stanley Dubroff; Guthrie Clinic Branch: Arthur King; Hampton Roads Branch: Samuel Katzoff; Hercules Research Center Branch: H. G. Tennent; Natick Quartermaster Branch: E. F. Degering; Naval Material Laboratory Branch: Solomon Goldspiel; Naval Research Laboratory Branch: Victor J. Linnenbom, George Abraham, J. E. Dinger; Navy Hydrographic Office Branch: Adrian F. Richards, Max. C. McLean and A. Russell Mooney; Roche Research Club Branch: Ernest G. Wollish; Summit Association of Scientists Branch: C. J. Swartz.

F. D. Snell of the Foster D. Snell, Inc. Branch, Thomas J. Maresca of the General Electric Company, Ithaca Branch and E. F. Cox of the Whirlpool Corporation Branch had been officially designated as delegates but at the last minute were unable to attend.

In addition to these delegates from 20 branches there were present: George H. Boyd, President, and T. T. Holme, Executive Secretary of Sigma Xi, W. J. Coppoc and Peter King (all four are members of the Governing Board of RESA), Wallace R. Brode, President of AAAS and past Chairman of RESA, Hugh S. Taylor, Chairman of the Editorial Board of the AMERICAN SCIENTIST, Frank M. Carpenter, President-elect of Sigma Xi and Chairman of the Committee on Lectures, Roy L. Whistler, member of the Executive Committee of Sigma Xi and D. B. Prentice, Director of RESA.

No corrections or additions having been reported by the Branches,

It was *VOTED:*

to approve the Minutes of the 1957 Convention, held in Indianapolis on December 27, as circulated to the branches and clubs and published in the AMERICAN SCIENTIST for March 1958.

The Chairman reported that, since the 1957 Convention, he had installed branches of the Society at Dow Chemical Company, Midland, Michigan on January 22, Butler-Indianapolis at Indianapolis on February 19 and Monsanto Chemical Company, St. Louis on September 23. The two regular meetings of the Board of Governors, as prescribed by the Constitution, had been held in March and October in New York in conjunction with the corresponding meetings of the Executive Committee of Sigma Xi. The RESA meetings were held in the mornings and the Sigma Xi meetings in the afternoons with a joint luncheon for the two groups. This seems to be a very satisfactory arrangement as the Societies share several activities and there are some common problems.

The Chairman reported further that the rewording of the amendment to Article VII, Section 5, which amendment had been proposed by the Boulder Branch and approved in principle by the 1957 Convention, had been written by a Committee of the Board, Messrs. Flexser and Prentice, and approved by the Board at its October meeting. To make it a matter of record in the Minutes of this Convention the Chairman read the amended section, as follows:

> The Membership of a branch, subject to the provisions of Article XI, Section 3, shall consist of (a) regular branch members and (b) affiliate branch members, who also may be called branch affiliates. Regular branch members are defined as members and associate members of the Society who are connected with the organization(s) in which the branch is located and, at the option of the branch, members and associate members of the Society not connected with the organization(s) in which the branch is located. These latter members and associate members may be members of the Society by prior election (or by registration if members of Sigma Xi) or they may be elected by the branch. Branch affiliates are defined as members and associate members of the Society not connected with the organization(s) in which the branch is located who are accepted as affiliates in addition to the regular branch members. Branch affiliates have all the duties and privileges possessed by regular branch members except that of voting.

The Chairman reported that the special booklet about RESA and the Procter Prize Awards had been well received, an edition of 3000 was exhausted and a reprinting was planned, to include a record of the 1958 recipient of the Procter Prize, who would be Dr. C. Guy Suits, Vice-President and Director of Research of the General Electric Company. The Chairman announced that Dr. Suits would receive the award at the RESA-Sigma Xi luncheon which would follow the Convention and that he would deliver the annual RESA address at that time.

The Director reported that, in addition to the new branches installed by Dr. Weidlein, the Hampton Roads Branch had been installed on May 27, by Dr. Brode and the Navy Hydrographic Office Branch on November 21, also by Dr. Brode. On September 22, the Hercules Powder Company Branch had been installed by the Director. The installation of the Stromberg-Carlson Branch at Rochester, New York is scheduled for January 30, probably by Dr. Weidlein.

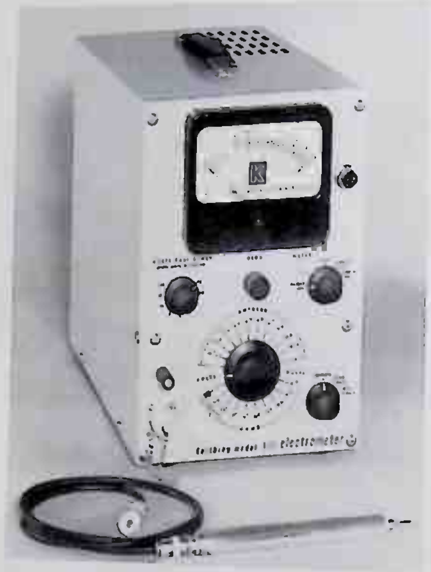
In addition to the above branches the Board of Governors has approved petitions for branches of the Society at the Shell Exploration and Production Laboratories in Houston, Texas, at Mead Johnson Company in Evansville, Indiana, at the Continental Can Company in Chicago, Illinois, at the Bakelite Division of the Union Carbide Company in Bound Brook, New Jersey, at Foster D. Snell, Inc. in New York City, and at the Whirlpool Corporation in St. Joseph, Michigan.

The Galesburg - Knox - Monmouth Club, which was installed in 1953 and has operated for five years, asked for a charter as a branch and the petition has been approved by the Board. A petition for a branch at the Cities Service Research and Development Laboratories at Cranbury, New Jersey had been received a few days before the Convention and was submitted to the Board for a mail ballot.

The year of 1958 has been the most active since RESA was established, with installations and petitions for sixteen branches. There will be 58 branches when the installations are completed. Current membership is about 8000, including about 1000 members-at-large. Current assessments have been

(*Continued on page 14A*)

Keithley electrometers for every dc laboratory test



Model 610, line-operated, 59 ranges

The Keithley 610 Electrometer is a laboratory workhorse, doing the work of several instruments. It covers the extreme spans of dc voltage, current, and resistance tabulated below, and is a useful preamplifier as well. It has precise gains to 1000, a dc to 500 cps bandwidth, and 10-volt and 1-ma outputs. Input resistance of the 610 is selectable from one ohm to over 10^{14} ohms. It checks its own resistance and voltage supply standards. Zero drift is comfortably within 2 millivolts per hour after warmup.

Model 600, battery-operated, 53 ranges

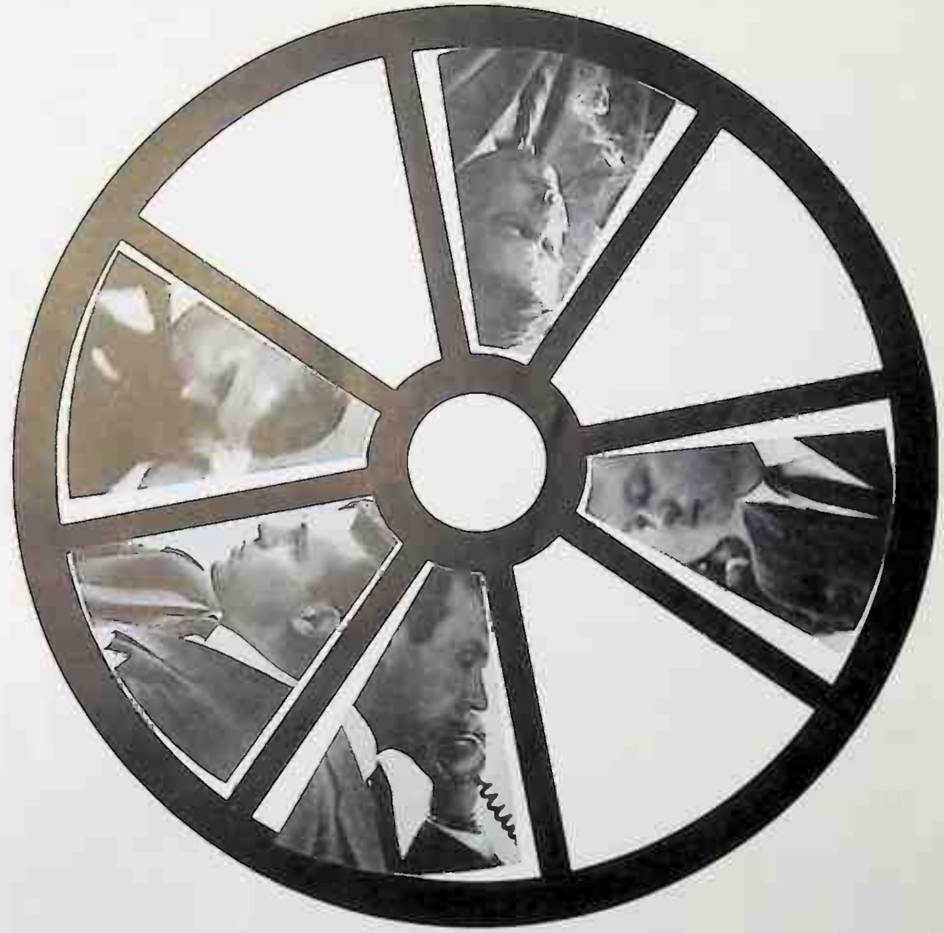
The Model 600 is a small, portable, battery-operated sister of the 610. Its many ranges also are tabulated below. Like the 610, its input resistance may be varied from one ohm to over 10^{14} ohms, permitting an optimum balance of low circuit loading verus minimum pick-up. Output is sufficient to drive potentiometric recorders directly, with a dc to 100 cps band-width, and zero drift is within 2 millivolts per hour. The 600 will check its own batteries; minimum battery life is 500 hours.

Need a few specials?
Recent Keithley developments include instruments for the Navy's Vanguard, for Army research balloons, and for Air Force research missiles. Your special problem will receive a prompt reply.

MODEL	FULL SCALE RANGES			PRICE
	VOLTAGE	CURRENT	RESISTANCE	
610	10 mv to 100 volts	10^{-15} amp. to 3 amperes	10 ohms to 10^{14} ohms	$465.00
600	10 mv to 10 volts	10^{-13} amp. to 3 amperes	10,000 ohms to 10^{13} ohms	$380.00

THREE ACCESSORY probes are available to facilitate measurements and extend the measuring range to 30 kv (Model 610) or 10 kv (Model 600). A convenient accessory test shield permits rapid checks of small components. Write today for more details.

KEITHLEY INSTRUMENTS, INC.
12415 EUCLID AVENUE · CLEVELAND 6, OHIO



MEN AND IDEAS IN MOTION: AERONUTRONIC

This is Aeronutronic—men, ideas, and the tools for research. Aeronutronic—a dynamic new name in science—created by the Ford Motor Company to meet the demanding technological needs of a nation on the move.

Aeronutronic is moving into the future and moving fast. Space sciences, missile technology and space vehicles... computers, electronics . . . tactical weapon systems... these are major research, development and manufacturing activities conducted at ASI's modern 200-acre Research Center under construction at Newport Beach, Calif.

Exceptional engineers and scientists are needed now. If you are forward-looking and want to be an important part of a forward-moving organization, you'll find a new challenge and rewarding future at Aeronutronic—*where men set ideas in motion.*

For information regarding positions, interests, facilities or products, write to Mr. K. A. Dunn, Aeronutronic Systems, Inc., Bldg. 3, 1234 Air Way, Glendale, California, or call CHapman 5-6651.

AERONUTRONIC
a subsidiary of FORD MOTOR COMPANY
NEWPORT BEACH, GLENDALE, SANTA ANA AND MAYWOOD, CALIFORNIA

OFFICE OF ADVANCED RESEARCH · SPACE TECHNOLOGY DIVISION · COMPUTER DIVISION · TACTICAL WEAPON SYSTEMS DIVISION · RANGE SYSTEMS DIVISION

Engineering careers grow better at Link

The climate is right for achievement at Link Aviation, Inc., for your career, and for your enjoyment of life.

Among the burgeoning projects at Link's two laboratories are those requiring cultivation by engineers who have a green thumb with *digital and analog computers.* If you have such talents, you will want to investigate a career at Link.

The professional atmosphere at Link is based on a recognition of technical competence and accomplishment. For those interested in advanced university study, Link will arrange tuition.

Salaries are excellent. The Link health, medical, and retirement plans are outstanding.

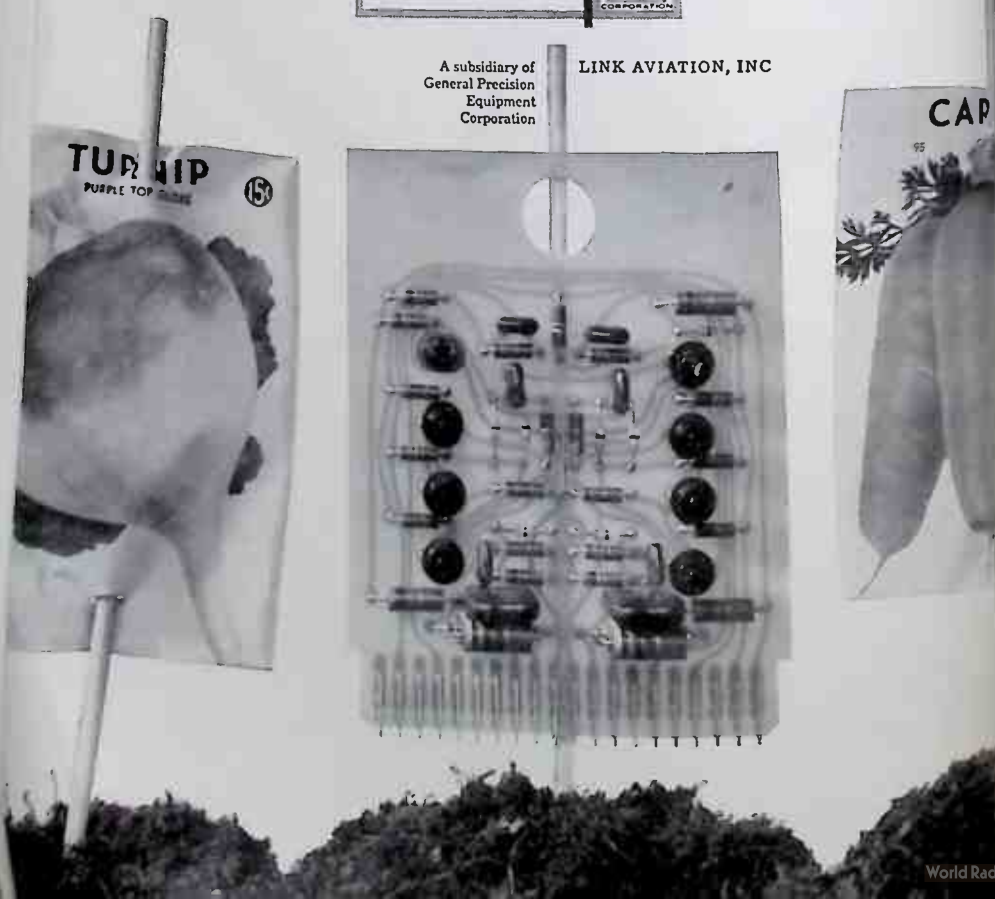
The charming suburban communities of Palo Alto, California and Binghamton, New York, are ideal locations for career advancement and for raising a family.

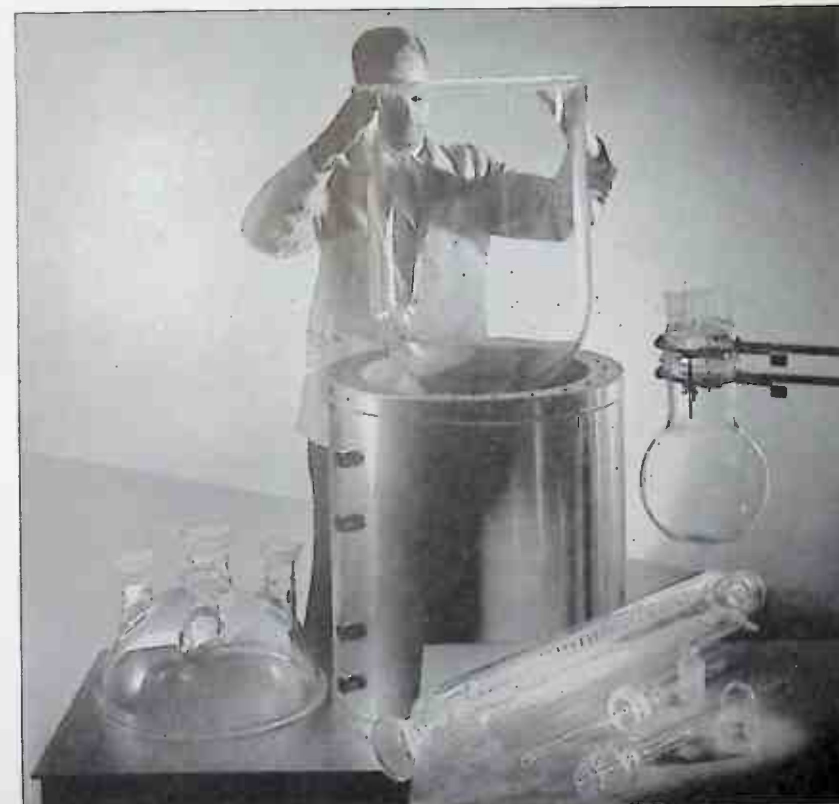
Link also has openings for engineers and scientists with capabilities in optical systems, electronic and electro-mechanical packaging, radar simulators, and automatic checkout equipment.

Your engineering career will grow better at Link.
Contact us at once. *Write to Mr. O. A. Larko, Link Aviation, Inc., P. O. Box 1318, Palo Alto, California or Mr. A. K. Wieland, Link Aviation, Inc., Binghamton, New York.*

LINK

A subsidiary of General Precision Equipment Corporation — LINK AVIATION, INC





Take much of the routine load off your pilot plant. Start a large-glassware section in your lab with one of these new PYREX reaction flasks. Multi-zone heating mantles, like the unit shown here, are recommended for all work with these flasks.

New giant PYREX® flasks give you wider choice of glass vessels for research and control reactions

Low cost PYREX reaction flasks give you many advantages over conventional units. Chemically inert, they're also heat resistant, easy to clean and maintain.

Now, check the added advantages of these three newly designed vessels:

Made in a widemouth, "urn" shape, they're available in 5, 10 or 20 gallon sizes. O.D. and height, respectively, are 12 x 16, 15⅜ x 20, 18 x 26 (inches). Covers have five openings and, like the accessories, can be had with pipe flanges or socket joints. Cataloged accessories include inlet tubes, condensers, blind caps, clamps and thermometer wells. Other components can be made to your specifications.

Other types . . . smaller sizes. Need a cylindrical flask? No. 6947, 500 to 4000 ml, has ground flanges and cover, four male ₸ joints. Another cylindrical flask, 500 to 3000 ml, but with pipe flange top is No. 93120. Covers have ₸ thermometer well and your choice of three socket, or female ₸, cover connections. Spherical flasks with pipe flange tops, 3000 to 72,000 ml, are listed under No. 93125. They take same covers as No. 93120.

Write to us at Corning, N. Y., for complete information on PYREX reaction flasks.

CORNING GLASS WORKS
CORNING MEANS RESEARCH IN GLASS

(*Continued from page 8A*)

received from all branches and will show an increase over 1957 of about 25%. Receipts from members-at-large have approximately doubled.

The Treasurer reported that the year's audited financial report would be published in the March issue of the AMERICAN SCIENTIST and would show a satisfactory addition to the operating reserve. The 1957 Convention approved an increase in the annual assessment from \$1.50 to \$2.00, to take effect in 1959. This increase is to meet the greater cost of the AMERICAN SCIENTIST which is to be enlarged fifty per cent. When RESA was established, it paid Sigma Xi the *increment* cost of printing and mailing the additional copies of the AMERICAN SCIENTIST required for RESA members. Two years ago the RESA budget covered the *average* cost of these copies and this cost is expected to increase by 50 per cent in 1959.

On motion of the Treasurer,

It was *VOTED:*

that the assessment per member or associate member for 1959 be \$2.00.

The Director reported that the Board of Governors recommended a corresponding increase of 50 cents in the first year fees for new members where a subcription to the AMERICAN SCIENTIST is included.

It was *VOTED:*

that the first year fees for 1959 be \$3.50 for new members, \$2.50 for inactive Sigma Xi members and \$1.00 for active Sigma Xi members who are already receiving the AMERICAN SCIENTIST.

The Board of Governors recommended a continuation of the present charter fees.

It was *VOTED:*

that the charter fees for 1959 be \$50.00 for a branch and \$25.00 for a club.

The Board feels that a continuation of the grant of authority of the Board to take final action on group petitions is very desirable in order to avoid delay until the year-to-year extensions of this grant by succeeding Conventions make prompt Board action possible and yet retain Convention control of the procedure, as specified in the Constitution.

It was *VOTED:*

that the Convention continue for one year the grant of authority to the Board of Governors to take final action on group and individual petitions for membership.

The Director reported that the Nominating Committee had proposed the continuation for another term of one year of the present officers of the Society and that accordingly the Board had elected, at its October meeting, E. R. Weidlein and D. B. Prentice for terms of one year from July 1, 1959 as Chairman and Director-Treasurer, respectively.

The Chairman called on Mr. G. B. Todd of the ALCOA Branch for the report of the Nominating Committee. For the Committee, which consisted of Norman Hilberry of the Board, Frank M. Moody of the Du Pont Branch and himself, Mr. Todd presented the nominations of W. J. Coppoc, Texaco Branch, Dean L. M. K. Boelter of the College of Engineering of U.C.L.A. and B. C. McKusick of the Du Pont Research Center Branch for membership on the Board of Governors for three-year terms commencing July 1, 1959.

The Chairman asked if there were any nominations from the floor. There being none,

It was *VOTED:*

that Messrs. Coppoc, Boelter and McKusick be elected members of the Board of Governors of RESA for three-year terms from July 1, 1959.

Dr. Hugh S. Taylor, Chairman of the Editorial Board of the AMERICAN SCIENTIST, discussed the present policies of the Editors, reported plans for the enlargement of the magazine and urged that members of RESA submit papers for possible publication. Dr. Taylor emphasized the dearth of papers reporting current research and stated the Editors' hopes of receiving many more of this type.

Dr. Frank M. Carpenter reported on the current operations of the lecture program with the list of speakers, subjects of addresses and schedules of tours. (All RESA program chairmen have received this material from the Lecture Committee.) Dr. Carpenter stated that next year the honorarium per lecture would be increased from $50.00 to $75.00 but without a corresponding increase in the charge to branches. The extra cost will be met from general society income. The $50.00 honorarium has been in effect since the beginning of the program more than a decade ago. As Dr. Carpenter becomes President of Sigma Xi in 1959 the Chairmanship of the Lecture Committee will pass to Dr. Harvey Neville, Vice-President of Lehigh University and a member of the Committee.

In the absence of Dr. Harlow Shapley, Chairman of the Committee on Grants-in-Aid of Research, Mr. T. T. Holme, Executive Secretary of Sigma Xi, reported on this activity. Mr. Holme stated that a complete list of grants had been published in the AMERICAN SCIENTIST and that the total for 1958 would approximate $27,000 as compared to $18,000 in 1957. He added that the Committee hoped to grant a total of $34,000 in 1959. Most of the gifts to the Research Fund are made by members-at-large of Sigma Xi and members of RESA are urged to increase their participation. The Committee feels that individual grants of $200–$750 meet a need not touched by the grants of the large foundations.

Under New Business the Convention considered a suggestion presented by Professor Howard W. Post of the Department of Chemistry of the University of Buffalo that RESA consider the feasibility and desirability of establishing branches outside of the United States. In particular it was suggested that a Japanese graduate student receiving his Ph.D. this year at the University of Buffalo might organize a branch in the government research laboratory at Tokyo, where he is a junior administrative officer.

After some discussion of the question,

It was *VOTED:*

that the Chairman be authorized to appoint a Committee to study the matter of foreign branches and to report to the next Convention through the Board of Governors. (Dr. Wallace R. Brode, Science Advisor to the State Department and a past-chairman of RESA, was named chairman of this Committee.)

The Director reported difficulties encountered by L. G. Balfour Company, official jewelers, in meeting the constitutional requirement of engraving name, date and electing branch on the back of a member's key. Some branch names are long and abbreviations are not informative. The delegates agreed, without vote, that the branch name might be omitted.

The Sequoia Branch had submitted a suggestion that the desirability of providing a corporate membership classification in RESA might well be considered. Such a grade of membership might carry an annual fee of $50.00 which would help carry the overhead expenses of a branch.

After discussion,

It was *VOTED:*

to authorize the Chairman to appoint a committee from the Board of Governors to study and report on this suggestion.

There being no further business,

It was *VOTED:*

to adjourn the 10th Convention of RESA.

(There was an attendance of 130 at the RESA-Sigma Xi luncheon which followed, and an attendance of 150 at the address by Dr. Suits, who received the 1958 William Procter Prize.)

Donald B. Prentice
Director

PROCEEDINGS OF THE FIFTY-NINTH ANNUAL CONVENTION WASHINGTON, D.C., DECEMBER 29, 1958

The Fifty-ninth Annual Convention of the Society of the Sigma Xi was held at the Sheraton-Park Hotel in Washington, D.C., on December 29, 1958, in accordance with the official announcement mailed to all chapters and clubs on October 24, 1958.

Call to Order

The Convention was called to order at 2:25 P.M., by the President, Dr. George H. Boyd, who appointed the following Committee on Credentials:

Dr. E. Ruffin Jones, Jr., University of Florida

Dr. Ward Pigman, Alabama Medical Center

Dr. Marion Maclean Davis, District of Columbia, Chairman

The Committee examined the credentials of the delegates and later the Chairman, Dr. Davis, reported to the Convention that 79 delegates representing 69 chapters, and 28 delegates representing 25 clubs, were in attendance as follows:

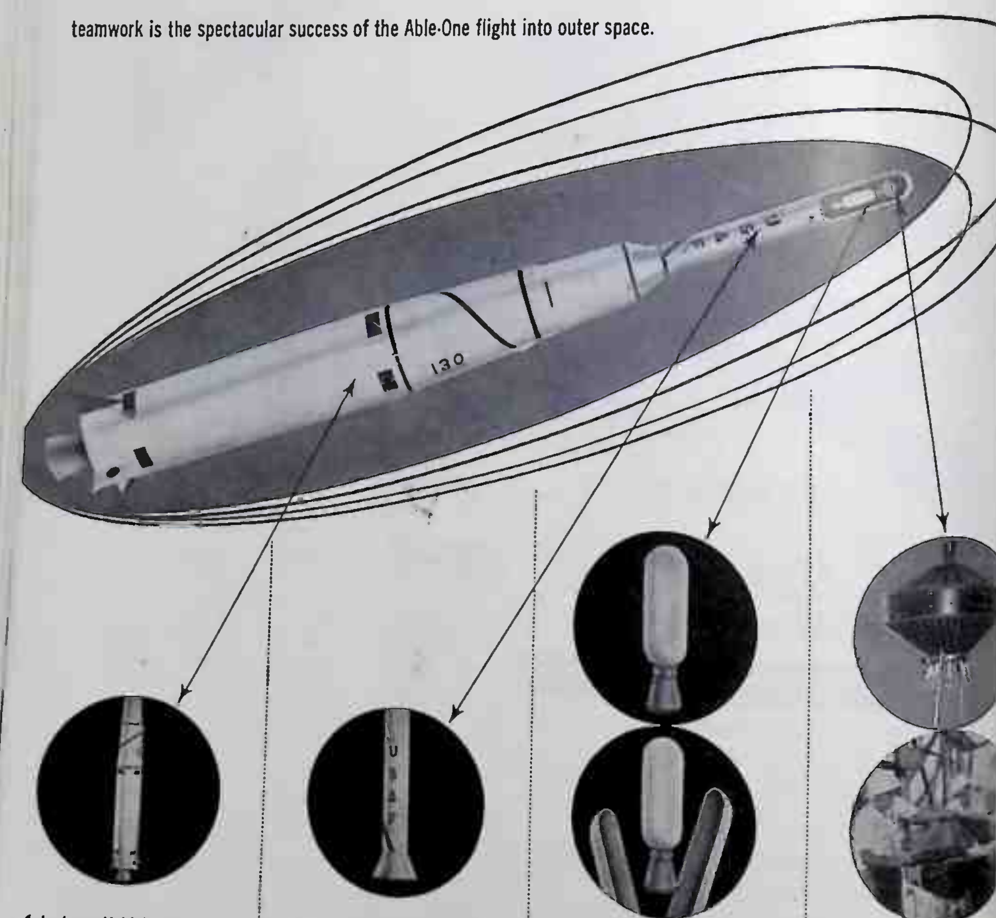
Chapter Delegates: Alabama Medical Center—Ward Pigman; Amherst College—H. H. Plough; University of Arizona—Albert R. Mead; University of Arkansas—P. M. Johnston; Boston University—Henry B. Russell; Brigham Young University—Wilmer W. Tanner; Brown University—Richard J. Gross; University of California, at Los Angeles—Stanley Ruttenberg; Carnegie Institute of Technology—Gilbert J. Mains; Catholic University of America—E. R. Kenedy, D. M. Marlowe, J. Wehner; Columbia University—A. W. Pollister; University of Connecticut,—N. T. Davis; Cornell University—Howard E. Evans; District of Columbia—Marion Maclean Davis; Emory University—C. G. Goodchild; University of Florida—E. Ruffin Jones, Jr.; Fordham University—Daniel Ludwig; George Washington University—R. E. Wood; Georgia Institute of Technology—W. T. Ziegler; Howard University—Walter T. Daniels, Roy C. Darlington; University of Idaho—J. W. O'Connell; Indiana University—Sears Crowell, Charles B. Heiser, Jr.; State University of Iowa—G. Edgar Folk; Iowa State College—Clarence H. Lindahl; University of Kansas—Joseph H. Camin; Kansas State College—Thomas D. O'Brien; University of Kentucky—Morris Scherago, John Carpenter, Mary Wharton; Lehigh University—Saul B. Barber; Louisiana State University—George Mickey; Loyola University, Chicago—Edward M. Nelson; University of Maryland—William Hollis; University of Massachusetts—John L. Roberts; Mayo Foundation—David T. Carr; Michigan State University—F. W. Snyder; University of Nebraska—Benjamin W. McCashland; University of New Hampshire—Marian H. Pettibone; New Mexico A. & M. College—M. G. Anderson; University of North Carolina—Martin Roeder; North Carolina State College—F. P. Pike; Northwestern University—Malcolm Dole; Ohio State University—Richard P. Goldthwait; University of Oklahoma—Teague Self; Oklahoma State University—Robert Fite, Pennsylvania State University—Marsh W. White; University of Pittsburgh,—Ralph Buchsbaum; Purdue University—Roy L. Whistler, E. J. McCormick; Rensselaer Polytechnic Institute—Roland Walker; University of Rhode Island—Emmett P. Christopher; University of Rochester—J. Lowell Orbison; Rockefeller Institute—Francis O. Holmes; Rutgers University—E. B. McLaurin; Smith College—George de-Villafranca; University of South Carolina—D. F. DeTar, A. P. French; State University of South Dakota—Benton W. Buttrey; Swarthmore College—Neal A. Weber; Temple University—M. Catherine Hinchey, Benedict M. Hall, Hazel M. Tomlinson; University of Tennessee, Medical Units—Clint L. Baker; University of Texas—John A. Wilson; College of Texas A. & M.—R. J. Bauldauf; Tulane University—Karlem Riess, Joseph Ewan; Union College—Leonard B. Clark; U.S. Naval Postgraduate School—Charles C. Torrance; University of Utah—Dorothea D. Mulaik; University of Viriginia—Jacques J. Rappaport; Virginia Polytechnic Institute—Robert C. Carter; State College of Washington—J. P. Meiners; Wellesley College—Delaphine Wyckoff; University of West Virginia—Leland Hart Taylor.

Club Delegates: Abbott Laboratories—Harry E. Sagen; University of Akron—Rolland R. Tougas; University of Alaska, Troy Péwé, John L. Buckley; Alfred University—Joseph L. Norton; American University of Beirut—Harry

(*Continued on page 24A*)

Able-One... a new apogee in scientific teamwork!

Preparation and execution of an undertaking such as the United States' IGY space probe demanded the participation and exceptional efforts of 52 scientific and industrial firms and the Armed Forces. The Advanced Research Projects Agency and the AFBMD assigned Space Technology Laboratories the responsibility for the project which was out under the overall direction of the National Aeronautics and Space Agency. One measure of this teamwork is the spectacular success of the Able-One flight into outer space.



1st stage: Vehicle, Douglas Aircraft Thor IRBM; propulsion, Rocketdyne; airframe, control, electrical and instrumentation, Douglas Aircraft; assembly, integration, and checkout, Douglas Aircraft.

2nd stage: Propulsion system and tanks, Aerojet-General; control, electrical, instrumentation, accelerometer shutoff, and spin rocket systems, STL; assembly integration, and checkout, STL.

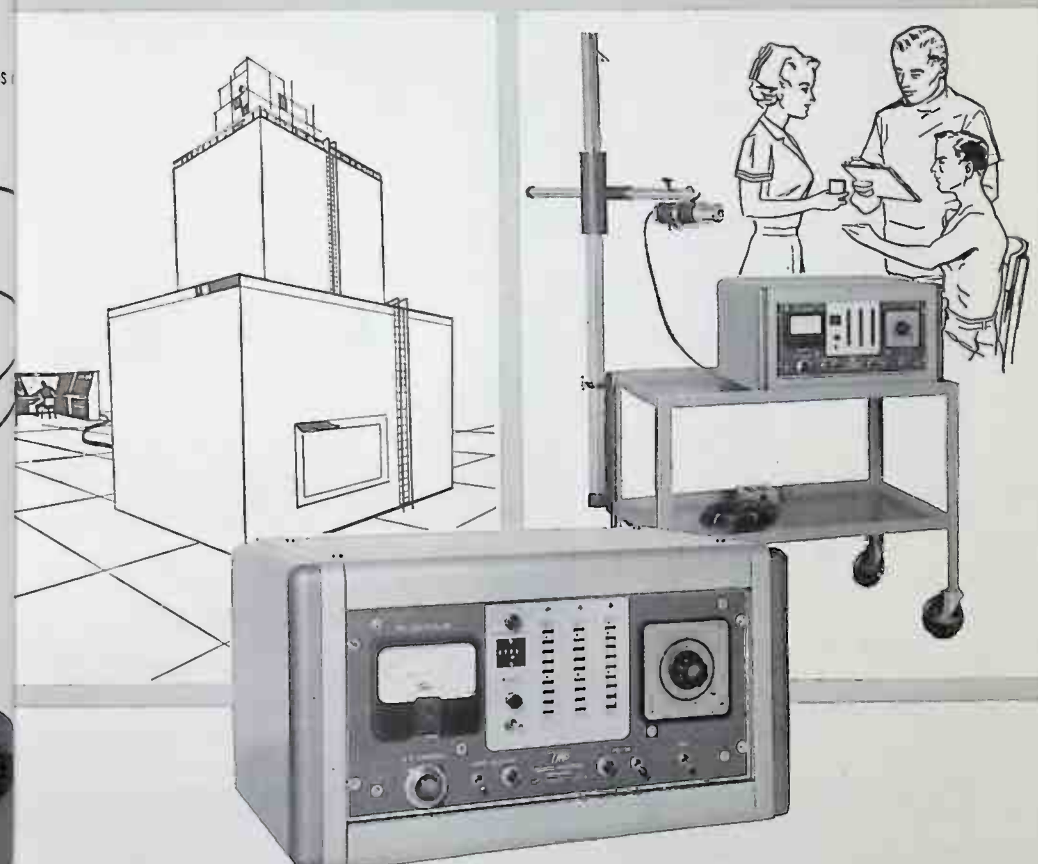
3rd stage: Rocket motor, U. S. Navy Bureau of Ordnance and Allegheny Ballistic Laboratory; structure and electrical, STL; assembly, integration, and checkout, STL; ground testing, USAF's Arnold Engineering Development Center.

Payload: Design and duction of Pioneer, th load of the Able-One ve was conducted by S addition to its overall nical direction and sy engineering respons of the Air Force Ba Missile Division pro This highly sophistic package included a NOT camera and transmitte Thiokol rocket motor.

Inquiries concerning openings on our staff will be welcomed by

Space Technology Laboratories, Inc.
5730 Arbor Vitae Street, Los Angeles 45, California.

FOR RELIABLE SCALING—
in *making* radioisotopes . . . or *monitoring* them



PERFORMANCE DATA

The SG-2A Scaler is a completely self-contained unit consisting of:

AMPLIFIER — Chase Higinbotham non-blocking type with selectable sensitivity — 1, 10 or 100 millivolts neg.

Rise time — 0.20 μs.

Dynamic range (at 1 mv. sens.) 10,000 to 1.

POWER SUPPLY

Dual Ranges 300 to 1000 volts, 300 to 2500 volts.

Stability — 2 volts low range, 4 volts high range (under normal operating conditions.)

(Available on special order with 5000 volt power supply for counters using high voltage gases. Dual ranges 1000 to 2500, 1000 to 5000 volts.)

SCALER — three etched wire decade strips followed by a precision four digit counter.

Counts — to 240,000 CPM with less than 1% coincidence loss.

Resolving time — 2.5 μs.

Auto-time — pre-set or elapsed time from 1 sec. to 60 min.

Accuracy ±0.2 sec.

(Model SG-2A4 also available with all electronic pre-set count from 100 — 10,000 counts.)

TMC MODEL SG-2A SCALER PROVIDES ACCURACY YOU CAN DEPEND ON—AT THE REACTOR, LABORATORY OR HOSPITAL

When the SG-2A is used as part of *reactor instrumentation*, the one millivolt sensitivity and wide dynamic range of its Chase Higinbotham non-blocking amplifier permit accurate measurement of neutron levels at start-up long before the less sensitive operating instruments detect their presence. For *medical diagnostic procedures* using radioisotopes, many hospitals find that the SG-2A with the mobile cart and detector arm (above), provides the reliability and good reproduceability that are particularly necessary for thyroid function studies and blood and plasma volume measurements using Iodine (I^{131}) tracers, or in determining red cell mass with radiochromium (Cr^{51}). Countless other applications of the SG-2A range from radioactivity protective monitoring systems in industry to experimental work in college laboratories — wherever there is need for accurate radiation measurement.

If you use radioisotope tracers, write for complete information on the SG-2A Scaler, or related TMC detectors, pulse height analyzers, ratemeters and other instruments. By describing your work and the radioisotopes being used, you will enable TMC to recommend the most suitable instruments for your needs.

TMC **TECHNICAL MEASUREMENT CORPORATION**
441 Washington Avenue, North Haven, Connecticut

Artist's concept of the guided-missile cruiser, U.S.S. *Long Beach*, the U.S. Navy's first nuclear-propelled surface warship, now under construction by Bethlehem Steel's Quincy, Mass., shipyard.

SHIELDING ENGINEERS
for Naval Nuclear Power Plant Development

Principal duties involve: design of primary and secondary shielding; performance of special radiation studies pertaining to operation and maintenance of nuclear power plants, development of computer programs, monitoring and health physics procedures.

The Central Technical Department of Bethlehem's Shipbuilding Division, an experienced organization recognized for its technical achievements, offers unusual opportunities for qualified shielding engineers to participate in our interesting and dynamic nuclear power program.

Minimum qualifications are:

1. BS in Engineering or Physics. 2. MS in Nuclear Engineering or equivalent. 3. At least two years' experience in design of shielding for mobile nuclear power plants or equivalent. 4. U. S. citizenship.

If you would like to make full use of your professional training in creative engineering, enjoy easy commuting and be close to New England's educational and recreational facilities, by all means write to us. Your inquiry will be considered confidential and acknowledged promptly.

Please address your résumé to:

BETHLEHEM STEEL COMPANY
Shipbuilding Division

CENTRAL TECHNICAL DEPARTMENT
Quincy 69, Massachusetts
Attention C. H. Goldthwaite, Assistant Manager

Illustrated above 6″ De Luxe Dynascope
* Reg. U. S. Pat. Off.

Excitingly New!

A Masterpiece of Design and Craftsmanship

CRITERION'S STANDARD DYNASCOPES*

6″ — 8″ — 10″ — 12″ — 16″ starting as low as $265

Custam built to meet a professional's exacting specifications priced within easy reach of the serious amatuer

Only Criterion could produce such magnificent instruments at such reasonable cost. Combining the latest advances in optical engineering with the old-world skill and patience of master craftsmen, these superb new Dynascopes, custom made,offer top-quality performance. Every feature necessary for superior viewing is precision finished to professional standards. Optical surfaces accurate to 1/10 wave.

The five matched eyepieces are the finest orthoscopic and achromatic. Mechanical features are carefully custom built to assure trouble-free operation. Construction throughout is of lifetime materials. There is both maximum rigidity and easy portability.

Above all, Criterion's strict control allows no room for chance. Each Dynascope is doubly tested and inspected, these rigid tests must be flawlessly met before shipment. Your satisfaction UNCONDITIONALLY guaranteed. Yet the price is amazingly low.

6″ Standard Model (80 lbs. shipping weight f.o.b. Hartford) $265
Also illustrated: Electric Clock Drive, $80; Setting Circles, $60; Permanent Pedestal, $70. Not illustrated, but now available single reflex camera with adaptor $89. Write today for full specifications and details of these superior reflecting telescopes. Easy payment terms available on request.

Compare these Superior Features Which Include

- f/9 Parabolic Mirror, accurate to 1/10 wave
- Fully Rotating Tube, for comfortable viewing
- Massive Equatorial Mount, adjustable for your latitude
- 5 Matched 1¼″ Eyepieces—3 Achromatic Ramsdens 45X, 76X, 152X)—2 Orthoscopics 228X, 343X)
- 8 x 50 Achromatic Finder Scope with crosshairs
- Secondary Support that minimizes diffraction
- Declination Axle, 1½″ diameter, with slow motion
- Polar Axle with Needle Bearings on 1½″ shaft
- Rack-and-pinion Focusing
- Double-draw Focusing Tube for any eyepiece
- Precision-fashioned 54″ Bakelite Tube
- An exclusive two way pier-Tripod, a massive 45 lb pier for permanent installation.
- A light weight tripod hidden inside pier for field trips.

CRITERION MANUFACTURING COMPANY
Manufacturers of Quality Optical Instruments
Dept. ASD8, 331 Church St., Hartford 1, Connecticut

(Continued from page 18A)

Smith; Boston College—Walter J. Fimian, Jr.; Bowman Gray School of Medicine—Charles McCreight; Chicago Medical School—James E. P. Toman; Chico State College—H. Courtney Benedict; Clarkson-St. Lawrence—Alfred Romer; Corning Glass Works—John F. Wosinski; University of Delaware,—R. R. Romkin, W. A. Connell; DePauw-Wabash—J. C. Polley; Georgetown University—James W. Johnston, Jr., Gloria C. Feeney; Milwaukee—Grace W. Gray; University of Mississippi—Theodore I. Bieber; State University of Mississippi—Denzel E. Ferguson; North Dakota Agricultural College—Franz H. Rathmann; Ohio University—Thomas S. Smith; Ohio Wesleyan University—Ronald R. Greene; Socony-Mobile—Alfred W. Francis; Southern University—Robert S. Beale; Southern Illinois University—Harvey I. Fisher; Texas Technological College—Earl D. Camp; Medical College of Virginia—William J. O'Malley.

Approval of Proceedings

The proceedings of the Fifty-eighth Annual Convention held at Indianapolis, Indiana, on December 27, 1957, were approved as published in the March issue of the American Scientist and distributed to the chapters and clubs.

Report of the President

President Boyd reported the following installations for 1958:

Chapters: Fordham University—April 30, Dean Taylor; University of South Carolina—May 10, Dean Boyd; Rockefeller Institute—May 13, Dr. Carpenter; U. S. Naval Postgraduate School—June 11, Secretary Holme.

Clubs: Los Angeles State College—January 24, Dr. DuBridge; Medical College of Virginia—January 28, Dr. Glass; Chico State College—April 25, Dr. Jones; Arizona State University—May 6, Dr. Bradbury; Pratt Institute—May 14, Dr. Korff; Mississippi State College—May 20, Dr. Auerbach; City College of New York—May 23, Dr. Townes; Northern Illinois University—23, Dr. Just; Idaho State College—May 26, Dr. Eyring; Drake University—Nov. 6, Dr. Turner; Texas Western College—Nov. 20, Dr. Williams; Hunter College—Dec. 3, Dr. Barr; Western Washington Experiment Station—Dec. 11, Dr. Fisher.

Dr. Boyd also announced anniversaries of chapters at the following institutions:

25th Anniversaries in 1958: Duke University and University of California at Los Angeles
50th Anniversary in 1958: Worcester Polytechnic Institute
25th Anniversaries in 1959: Tulane University and Massachusetts Institute of Technology
50th Anniversary in 1959: Purdue University

It was *VOTED that:*

Those Chapters observing 25th and 50th Anniversaries in 1958, be extended the congratulations of the Convention.

President Boyd then reported:

"As you well know, the work of this Society is done not by the President but by the Executive Secretary and his staff, the Treasurer, the Editor-in-Chief and staff of American Scientist, the Committee on Lectureships, the Committee on Grants-in-Aid of Research, the Committee on Membership-at-Large, and the many Chapters and Clubs scattered throughout the country. This makes it likely that, from the standpoint of things accomplished by him, the report of the President would be unimpressive. His work during the past two years has consisted of serving as installing officer for a new Chapter here and there, representing the Society at an occasional formal college or university function, meeting with the Executive Committee from time to time, and promoting the work of the Society in whatever way and whenever possible.

"Instead of reciting the activities of the President, I desire to express my appreciation to those persons who have done the work of the Society and to make mention of two or three matters

(Continued on page 32A)

SERVALL®

SUPERSPEED CENTRIFUGES

for continuous and batch operations

SEDIMENTATION DIRECTLY IN TUBES
FULLY AUTOMATIC CONTROLS

- 5 Rotors including Horizontal Field Aligning and Virus
- Exclusive Gyro-Action Self-Centering Direct Drive
- Dynamic Braking
- Automatic Acceleration and Timing
- Safety Design
- "Noisuppressor"
- Unmatched Versatility

We have been delivering these Superspeed Automatic Centrifuges and Continuous Flow Systems for almost two years—when you buy SERVALL, you buy work-proven instruments. Write for Bulletin AS-3CA3.

These unequalled Automatic and Continuous Flow features are now incorporated in the radically new SERVALL Superspeed Refrigerated Centrifuge. Before you specify any refrigerated centrifuge, contact us about this latest word in low-temperature instrumentation.

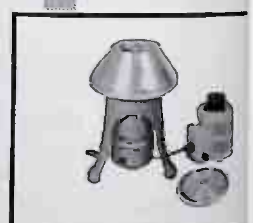
An Independent company: not connected with any other centrifuge manufacturer.
(Established 1934)

Ivan Sorvall, Inc.
NORWALK · CONNECTICUT

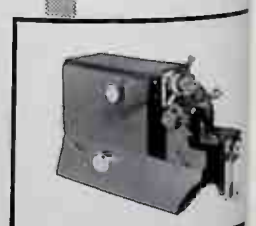
THE WORLD'S LARGEST MANUFACTURER OF SUPERSPEED CENTRIFUGES

Servall SS-4 Enclos
Automatic Superspe
to 17,000-rpm - 34,800

Servall Type SP/X
Medium Centrifuge
5,000 - 3,440 x G



Servall SS-1 Supers
Refrigerated Centri
16,000 rpm - 32,700 x G



Servall "Porter-Blu
Ultra-Microtome*
1/40 to 1/2 Micron

*Patented: U.S. & For

our world . . . and YOU

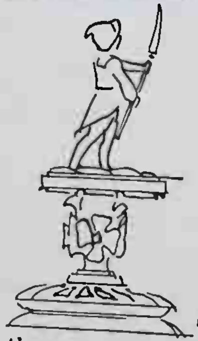
At Martin, you will find a world built around the creation of aircraft, missile and research systems for the Army, Navy, Air Force and Astro-Scientific branches of our government for operations on land, sea and air—and in outer space.

At the center of our universe in Baltimore, Maryland, you will work among world-renowned scientists and engineers probing the mystery of the unknown; unfolding secrets heretofore denied to 20th century mankind. Our staff of engineers and scientists aggregates better than 20,000 man-years of experience in scientific research and in the design and production of missiles, rockets and related electronic systems.

They constitute one of the country's most valuable resources devoted exclusively to the security of our world and its future progress in the exploration of space.

To investigate your personal opportunities, contact C. W. Spangler—Manager, Professional Employment—Dept. AS-2—The Martin Company, Baltimore 3, Maryland

MARTIN
BALTIMORE



your first move
CAN DECIDE THE GAME

your first job
CAN DECIDE YOUR FUTURE

That important first job can start you off in the wrong direction—*or* it can lead you straight toward your goal. If your ambitions are high, Motorola has a place that will give you the *finest* chance possible for the advancement you want. You'll get security and good salary, but, more important, you'll be working on projects with a *future*, like missile guidance, radar, and microwave. The door is wide open at Motorola, and the opportunity to fulfill your ambitions is *yours*.

If you are on ELECTRICAL ENGINEER, MECHANICAL ENGINEER or PHYSICIST, contact Motorola today.

CHICAGO, ILL.
Challenging positions in Two-Way Communications, Microwave, Radar and Military equipment, Television (Color) and Radio Engineering.

PHOENIX, ARIZ.
Outstanding opportunities in the development and production of Military equipment and Transistor products.

RIVERSIDE, CALIF.
This new modern research laboratory, located 65 miles from Los Angeles, needs men in Missile and Military equipment systems analysis and design.

For Assignments with a FUTURE UNLIMITED write:

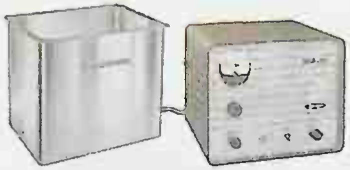
MOTOROLA Engineering Recruitment Office
4501 Augusta Blvd. • Chicago 51, Ill.

MAMMOTH narda SONBLASTER

America's first mass-produced industrial-size ultrasonic cleaner!

SAVE *7 ways over costly solvent, alkaline or vapor degreasing:*

• Clean faster, speed production! • Cut rejects, eliminate bottlenecks! • Save on chemicals and solvents! • Cut maintenance and downtime! • Eliminate expensive installation! • Save on floor space! • Release labor for other work!



G-1501 generator, NT-1505 tank.

MAMMOTH 5-GALLON TANK **$695**

Other models from $175. 2-year guarantee on all units.

SPECIFICATIONS: Interior Tank size (in.), 10W x 14L x 9½H
Tank Capacity, 5 gallons.

Now you can say goodbye to expensive chemicals, solvents, and degreasing equipment . . . reclaim valuable floor space . . . eliminate high installation costs . . . just by installing a Narda Series 1500 SonBlaster. At the same time, you'll get better, faster cleaning, and you'll need fewer people to do the job!

Get the tremendous activity of the new 200-watt Narda SonBlaster, with the largest transducerized tank ever made, at the lowest price in the industry! Choose from transducerized tanks or submersible transducers for use in any arrangement in any shape tank you desire. Up to 4 submersible transducers can be easily operated from the same generator at one time; load selector switch provided—an exclusive Narda feature.

Simply plug the SonBlaster into any 110-115 V AC line, and flip the switch. In seconds, you'll clean 'most any mechanical, optical, electrical, medical or horological part or assembly you can think of. Perfect, too, for brightening, polishing, radioactive decontaminating, pickling, quenching and plating; emulsifying, mixing, sterilizing, impregnating, degassing, and other chemical process applications.

Mail the coupon for free help in determining the model that's best for you.

The SonBlaster catalog line of ultrasonic cleaning equipment ranges from 35 watts to 2.5 KW, and includes transducerized tanks as well as immersible transducers. If ultrasonics can be applied to help improve your process, Narda will recommend the finest, most dependable equipment available—and at the lowest price in the industry!

Submersible Transducers

Model NT-604—Hermetically sealed heli-arc welded stainless steel case. Radiating face: 27 sq. in. Effective plane of radiation: 40 to 50 sq. in. (approx. 10" x 5"). Effective cavitation of volumes: up to 1200 cu. in. at 24" tank height (5 gal.) and 2400 cu. in. at 48" tank height (10 gal.). Swagelok tube fitting on side or end for internal tank wiring.

Model NT-605—Same as NT-604 except for bulkhead fitting on back for external wiring. Eliminates electrical conduits in solutions.

The Narda Ultrasonics Corporation
625 Main Street
Westbury, L.I., N.Y.
Department AS-15

Gentlemen:
Please send me more information about
☐ Series 1500 SonBlasters
☐ The complete Narda line

Name________
Organization________
Address________
City________
Zone____ State________

the narda ultrasonics corporation
625 MAIN STREET, WESTBURY, L. I., N. Y.
Subsidiary of The Narda Microwave Corporation

(*Continued from page 24A*)

that are on my mind as a result of my experience of the last two years.

"First of all, I would commend the officers and the committees of this Society for the fine services they have rendered during the past two years. Praise is due the Executive Secretary and the Treasurer of the Society, together with the members of their staff, for the efficiency, the wisdom and the loyalty with which they have conducted the work of their offices.

"Those of us who have been members of Sigma Xi for some years have watched with much satisfaction the progress of AMERICAN SCIENTIST. It has steadily improved in scope and quality, and I believe I represent the attitude of the majority of our members when I express pride in AMERICAN SCIENTIST as the official publication of the Society. We are profoundly grateful to Dean Taylor for his services so generously given as Editor-in-Chief during the past two years.

"Our lectureship program is an activity in whose benefits many members of the Society have shared and one in which great interest has been evinced from time to time. The selection and engagement of lectures and the arrangement of traveling and lecture schedules is a more difficult undertaking than many of us realize. I should like to thank Professor Frank Carpenter for the outstanding job which he and the members of his Committee have done in carrying out this program.

"Our Secretary, the members of his staff, and the Committee on Membership-at-Large have done splendid work with the important group of our members who no longer maintain their association with active chapters or clubs but who continue to have an active interest in the affairs of the Society. Within the past five years the number of these active Members-at-Large has grown from 5000 to 20,000. It should be mentioned that the contributions of this group to the funds of the Society constitute approximately ninety per cent of the fund for grants-in-aid of research, and the active and continuing interest of this group means much to the Society in a variety of ways.

"Mention of the fund for research leads me to mention the work of the Committee on Grants-in-Aid under the chairmanship of Professor Harlow Shapley. In this day of large grants and contracts involving many thousands of dollars, it might seem that there is little place for grants that provide only a few hundred dollars to the individual in research. The response to the grants-in-aid program leads us to feel that this is not the case. Apparently, there are many persons, located in places where support for research is not adequate, whose interest in research problems may be given impetus by no more than a small grant, and we want to thank Professor Shapley and the members of his Committee for their careful effort to see that these funds are wisely distributed.

"Finally, I want to express my appreciation to the members of our Executive Committee for their faithfulness and for the intelligent effort they have made on all occasions to serve the best interests of the Society.

"To be President of the Society of the Sigma Xi is important largely for the opportunity it affords for becoming educated in some small degree, at least, in the affairs of organization; and, if the President has profited as he should have from this opportunity, some of his impressions might be worth relating to you. Time will not permit me to outline them in detail, but I shall at least mention two matters in a very general way and a third one more specifically. With 133 Chapters and 70,000 members, Sigma Xi has become an organization of far-reaching influence; and, as the organization continues to grow, it would be easy for its procedures to assume something of a routine character and for some relaxation in its standards to occur. It is my impression that in the past history of the Society great carefulness has prevailed in the granting of charters by the Convention and in election to membership by the Chapters. There is evidence that the seriousness, the discrimination and the rigid standards for scientific achievement that characterized the pioneer days of the organization contributed much to the prestige and influence which it has had. If the Society is to continue to be an important force in the promotion of research in science, my own judgment is that its strict adherence to scientific standards must continue to be maintained.

"The second suggestion I would make is that the Society must be continually looking for new means by which it may further its cause and continually reexamining its practices and policies for their effectiveness in carrying out its traditional purpose of the encouragement of original research. Its program of grants-in-aid, its lectureship program, its election to membership, and all its other activities should be subject to constant review for their pertinence in an age that is undergoing rapid change in respect to science.

"The final, and perhaps the most important matter which I would mention is the relationship between the Society of the Sigma Xi and RESA. This is not a proposal for a merger of the two, but a suggestion that thought be given to the means whereby the two organizations may be placed in a sound relationship to each other. The matter of granting charters for the establishment of chapters of Sigma Xi in non-university laboratories was brought before this Convention, I believe, in 1947. The records indicate that the proposal was not considered acceptable at that time; and, instead of endorsing it, the Convention voted to establish RESA as 'a separate organization controlled by Sigma Xi.' It also instructed the Executive Committee to prepare a constitution and by-laws for the new organization to be considered by the 1948 Convention. Accordingly, members of the Executive Committee met in New York in June 1948, with representatives of some fifty industrial research organizations and undertook to carry out this directive. In the 1948 Convention the constitution and by-laws of RESA were approved.

"At the 1951 Convention of Sigma Xi President Hugh Taylor took occasion to speak to the delegates present on the importance of a single society, thereby eliminating the unnecessary duplication and confusion now apparent with the two societies, Sigma Xi and RESA, aimed at the same general purpose. He proposed two parallel divisions of a single society to be known as the Scientific Research Society of America, these divisions to be Sigma Xi and RESA. Sigma Xi would continue to be the society for the academic groups and RESA would continue to be that of the governmental and industrial groups. He urged delegates to suggest to their chapters the desirability of this one scientific society.

"In March 1952, President Taylor addressed the following letter to the officers of Sigma Xi Chapters and Clubs:

'The time has arrived when, in the judgment of the National President and Officers of Sigma Xi, serious consideration should be given to a consolidation of Sigma Xi Chapters and Clubs and RESA Branches and Clubs in one organization to be known as the *Scientific Research Society of America*. The consolidated Society would operate in two divisions, the one the Sigma Xi Division, the other the RESA Division. The Sigma Xi Division would be charged anew with the historic objectives of the Society of the Sigma Xi which, for nearly seventy years, has had the obligation of promoting the welfare of scientific research in the universities and colleges of the country. To the RESA division would be assigned a like measure of responsibility for the promotion of scientific research in the research laboratories of the nation, outside the universities and colleges, in industrial and government research organizations.

'Such a consolidated Scientific Research Society of America could be operated under a single group of national officers and under a single board of management and control in matters both scientific and financial. Each section would provide the tradition and experience accumulated in its past. RESA would provide an already assured financial stability and an opportunity for collaboration in the known objectives of Sigma Xi in the areas of industry and

(*Continued on page 42A*)

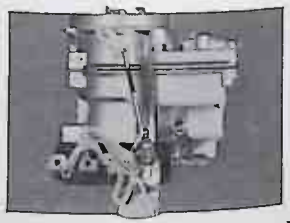
no. 6 OF A SERIES

TWO STEPS TO NEW FUEL ECONOMY

Unique Oldsmobile-developed two-stage automatic choke is a major step forward in improving automobile operating economy.

One of the important carburetor developments during the past few years was the automatic choke, a device that allows the automobile to be started in cold weather, and then keeps it running until the engine is sufficiently warmed up to sustain itself. Every automatic choke has two separate functions: 1) choking, which enriches the fuel-air mixture for starting, and 2) the idle speed control, which keeps the engine from stalling once it is started. In the past, and on all present carburetors except those used on the 1959 Oldsmobile, these two functions have operated simultaneously with the result that the engine ran on a rich mixture for the same length of time that the fast idle was "on". This resulted in excess fuel consumption.

With the introduction of the 1959 Oldsmobile, the two functions have been separated with a new and exclusive two-stage automatic choke developed by O engineers. An ingenious system of over-run allows the choke fly to open 75% sooner th ously required. The fast idle, however, remai the full warm so the engine stall. This ear tion of the function rep considerable f in cold wea numerous sh are made.



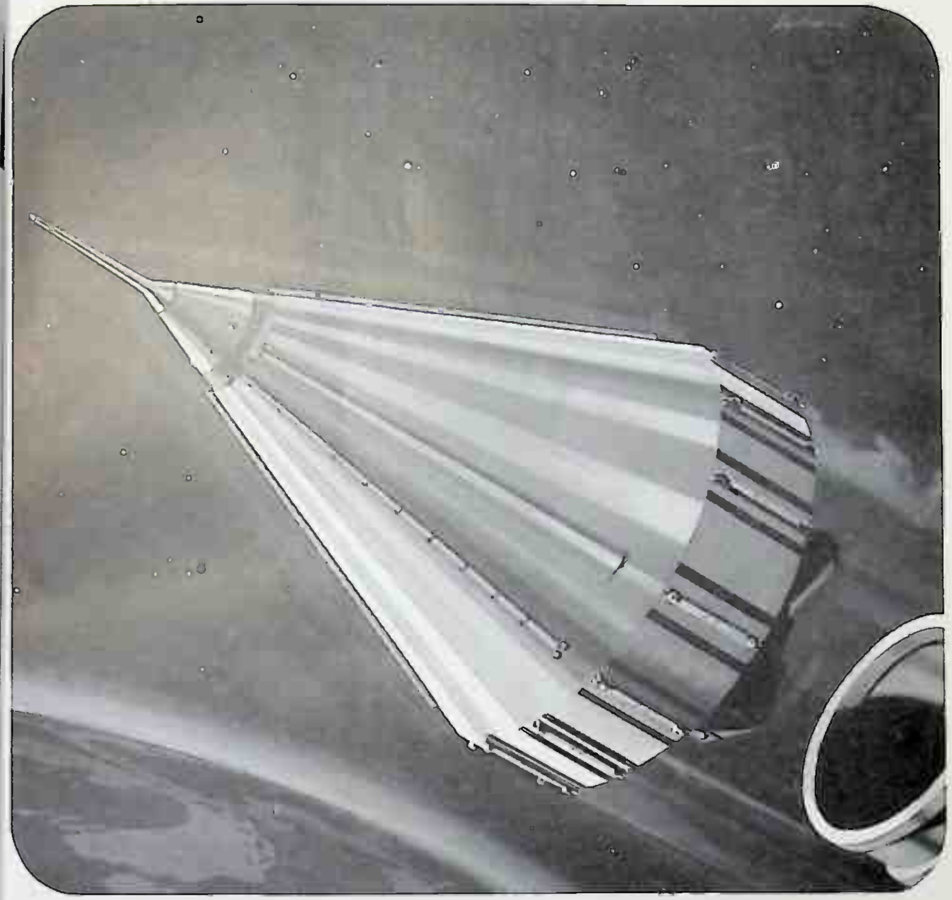
At Oldsmobile the Inquiring Mind is alway finding new and better ways to design, engineer finer automobiles for the most discriminating the Oldsmobile owner. Discover the difference by visiting your local Oldsmobile Quality Dealer a demonstration ride in a 1959 Oldsmobile.

OLDSMOBILE DIVISION, GENERAL MOTORS COR

OLDSMOBILE

Pioneer in Progressive Engine ...Famous for Quality Manufac

NOTABLE ACHIEVEMENTS AT JPL



PIONEERING IN SPACE RESEARCH

Another advance in man's knowledge of outer space was provided by Pioneer III. This, like others of a continuing series of space probes, was designed and launched by JPL for the National Aeronautics and Space Administration. JPL is administered by the California Institute of Technology for NASA.

During the 38 hours of its flight, Pioneer III was tracked by JPL tracking stations for 25 hours, the maximum time it was above the horizon for these stations.

The primary scientific experiment was the measurement of the radiation environment at distances far from the Earth and telemetering data of fundamental scientific value was recorded for 22 hours. This data revealed, at 10,000 miles from the Earth, the existence of a belt of high radiation intensity greater than that observed by the Explorer satellites.

This discovery is of vital importance as it poses new problems affecting the dispatch of future vehicles into space.

The study and solution of such problems compose a large part of the research and development programs now in extensive operation at the Laboratory.

CALIFORNIA INSTITUTE OF TECHNOLOGY
JET PROPULSION LABORATORY • Pasadena, California
A Research Facility operated for the National Aeronautics and Space Administration

Opportunities now open in these classifications ▶ APPLIED MATHEMATICIANS • ENGINEERING PHYSICISTS • COMPUTER ANALYSTS • IBM-704 PROGRAMMERS • AERONAUTICAL ENGINEERS • RESEARCH ANALYSTS • DESIGN ENGINEERS • STRUCTURES AND DEVELOPMENT ENGINEERS

$$\dot{x}_{j+1}(t)=\dot{x}_j(t-h) \text{ if } x_j(t-h)-x_{j+1}(t-h)=\beta S_c$$

● Problem: what doctrine for a motorized military convoy will mean the highest over-the-road speed? Solving such a problem by experimental, trial-and-error methods is difficult, long, and costly . . . yet answers to such questions are vital to our modern, mobile U. S. Army. Scientists of *tech/ops* solved this one by devising and applying a mathematical model to describe a convoy, programming this model for a large digital computer. *Result:* another application of *tech/ops'* research techniques to solve a problem whose solution by conventional means would have been prohibitively expensive . . . and a typical example of *tech/ops'* pioneering work in *operations research and broad scientific research and development* for *industry, business and government.*

Two other formulas complete this model:

$\dot{x}_{j+1}(t) = V_c$ if $\beta S_c < x_j(t-h) - x_{j+1}(t-h) \leq S_c$

$\dot{x}_{j+1}(t) = \frac{1}{T}[x_j(t-h) - x_{j+1}(t-h)]$ if $x_j(t-h) - x_{j+1}(t-h) > S_c$

The symbols have these significances: $x_j(t)$ is the position of the jth vehicle at time t; V_c is the assigned convoy speed; S_c is the assigned spacing between succeeding vehicles in the convoy; h the driver reaction time; β is a constant. Boundary conditions: $\dot{x}_j(t) \geq 0$; $\dot{x}_1(t)$ is a given (known) function.

Technical Operations, Incorporated

Central Research Laboratories / Burlington, Massachusetts

WASHINGTON, D. C. • MONTEREY, CALIFORNIA • FORT MONROE, VIRGINIA

tech/ops

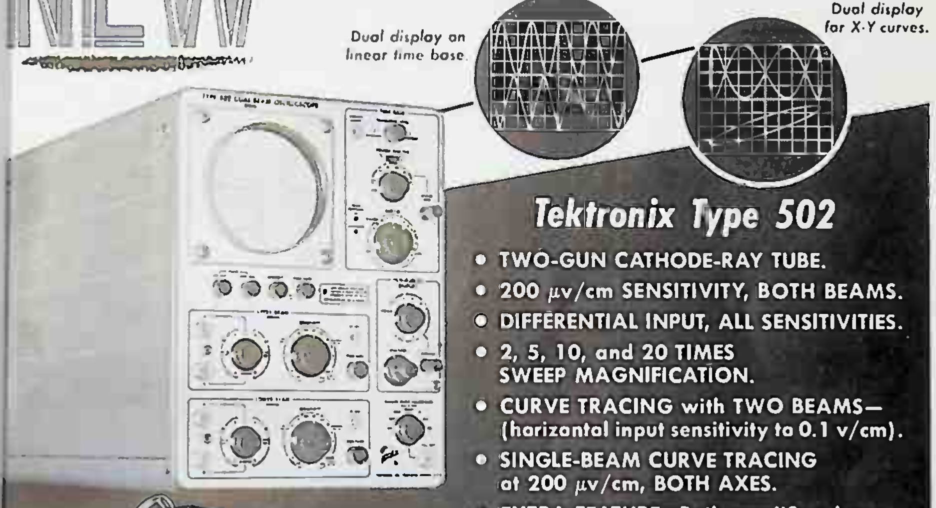
tech/ops personnel requirements

at Fort Monroe, Virginia, or Burlington, Massachusetts: Operations analysts experienced in industrial or military operations research, systems analysis, weapons systems evaluation, computer techniques, or related fields; training should be in mathematics or physical sciences, preferably on graduate level.

at Washington, D. C.: Programmers with substantial experience in the development of large digital computer programs; background should include experience with design and application of assembly programs, compilers, and advanced programming concepts.

at Monterey, California: Communications engineer or physicist thoroughly familiar with the principles of radio transmission and communications network analysis.

NEW DUAL-BEAM OSCILLOSCOPE



Tektronix Type 502

- TWO-GUN CATHODE-RAY TUBE.
- 200 µv/cm SENSITIVITY, BOTH BEAMS.
- DIFFERENTIAL INPUT, ALL SENSITIVITIES.
- 2, 5, 10, and 20 TIMES SWEEP MAGNIFICATION.
- CURVE TRACING with TWO BEAMS—(horizontal input sensitivity to 0.1 v/cm).
- SINGLE-BEAM CURVE TRACING at 200 µv/cm, BOTH AXES.
- EXTRA FEATURE—Both amplifiers have transistor-regulated parallel heater supply.

TEKTRONIX ®

TYPE 502 CHARACTERISTICS

HIGH-GAIN AMPLIFIERS

200-microvolts/cm deflection factors, both dc-coupled and ac-coupled. 16 calibrated steps from 200 µv/cm to 20 v/cm.

Passbands—dc-to-100 kc at 200 µv/cm, increasing to dc-to-200 kc at 1 mv/cm, dc-to-400 kc at 50 mv/cm, and to dc-to-1 mc at 0.2 v/cm.

Differential Input, Both Channels—Rejection ratios: 1000-to-1 at 1 mv/cm or less, 100-to-1 at 0.2 v/cm, 50-to-1 at 5 to 20 v/cm.

Constant Input Impedance (1 megohm, 47 µµf) Both Channels—from 1 mv/cm to 20 v/cm, for use with Tektronix P510A Probes. (2-P510A Probes furnished).

WIDE-RANGE SWEEP CIRCUIT (Common to both beams)

Single-knob control for selecting any of 22 accurately-calibrated sweep rates from 1 µsec/cm to 5 sec/cm.

Sweep Magnification—2, 5, 10, and 20 times, accurate within the maximum calibrated sweep rate.

Automatic Triggering—fully automatic, or preset with amplitude-level selection when desired. Sweep can also be operated free-running.

X-Y CURVE TRACING OPERATION

Horizontal-input amplifier permits curve-tracing with both beams simultaneously at sensitivities to 0.1 v/cm. For curve tracing at higher sensitivities (to 200 µv/cm) with one beam, one of the vertical amplifiers can be switched to the horizontal-deflection plates.

OTHER FEATURES

Amplitude calibrator, 1 mv to 100 v in decade steps—square wave, frequency about 1 kc.

3-kv accelerating potential on new Tektronix 5" dual-beam crt. 8-cm by 10-cm linear-display area, each beam.

Electronically-regulated power supplies.

Price . . . $825
f.o.b. factory

Here are a few uses for the Type 502:

IN ELECTRONICS—Use the Type 502 as a general-purpose oscilloscope and also to show simultaneously the waveforms at any two points in a circuit, e.g. input and output, opposite sides of a push-pull circuit, trigger and triggered waveform, etc.

IN MECHANICS—Display, compare, and measure outputs of two transducers on the same time base; plot one transducer output against another—pressure against volume or temperature for instance; measure phase angles, frequency differences, etc.

IN MEDICINE—Display, compare, and measure stimulus and reaction, or the outputs of two probes, on the same time base; use differential input to cancel out common-mode signals, or to eliminate the need for a common terminal; use in routine investigations, etc.

IN ALL FIELDS—The Type 502 can save you more than its cost in time—in as little as one application!

Tektronix, Inc.

P. O. Box 831 • Portland 7, Oregon

Phone CYpress 2-2611 • TWX-PD 311 • Cable: TEKTRONIX

TEKTRONIX FIELD OFFICES: Albertson, L. I., N.Y. • Albuquerque • Atlanta, Ga. • Bronxville, N.Y. • Buffalo • Cleveland • Dallas • Dayton • Elmwood Park, Ill. • Endwell, N.Y. • Houston • Lathrup Village, Mich. • East Los Angeles West Los Angeles • Minneapolis • Mission, Kansas • Newtonville, Mass. Orlando, Fla. • Palo Alto, Calif. • Philadelphia • Phoenix • San Diego • St. Petersburg, Fla. • Syracuse • Towson, Md. • Union, N.J. • Washington, D.C. Willowdale, Ont.

TEKTRONIX ENGINEERING REPRESENTATIVES: Hawthorne Electronics, Portland, Oregon, Seattle, Wash.; Hytronic Measurements, Denver, Colo., Salt Lake City, Utah.

Tektronix is represented in 20 overseas countries by qualified engineering organizations.

Engineers—Scientists
Required for many aspects
of General Electric's
Aircraft Nuclear Propulsion
Program...

PhD

APPLIED RESEARCH & DEVELOPMENT AT THE PhD LEVEL ON AN INTER-DISCIPLINARY BASIS

Opportunities now exist for a number of scientists and engineers with imaginative minds and established reputations to contribute to extremely sophisticated nuclear propulsion systems, where theoretical advances must be wedded to new standards of reliability and performance.

Qualified engineers and scientists are invited to inquire about these opportunities to pursue original work in a pioneering area with General Electric.

Metallurgy	Advise Materials Testing Laboratory on materials pr… lems and techniques. Direct experimental, physical, m… chanical and electrical tests. Analyze data.
Physics or Nuclear Engineering	Carry out reactor analysis and nuclear design of pow… plants, including formulation of nuclear design specifica… tions and test requirements.
Ceramics, Physical Chemistry or Ceramic Engineering	Supervise applied ceramic research including fabricatio… processes, testing of ceramic bodies; development of ne… ceramic materials.
Physics	Conduct theoretical investigations of effect of neutrons a… photons on matter.
Mathematics, Electrical Engineering	Engineering analysis of physical systems in electrome… chanical areas, deriving equations associated with system… study. Develop generalized digital program for parametr… study of these systems.
Mathematics	Apply advanced mathematical procedures and approach… in resolving complex problems in areas of reactor analys…
Physics or Nuclear Engineering	Advanced Reactor Analysis. Plan objectives of experim… tal reactor investigations, which adequately combine … tentialities of experimental techniques and improveme… of theoretical methods.
Physical Metallurgy or Mechanical Engineering	Undertake responsibility for studies regarding various m… terials used in moderator materials development.

Write in confidence, including salary requirements, to: Mr. G. W. Travers, Div. 77-QC.

AIRCRAFT NUCLEAR PROPULSION DEPARTMENT

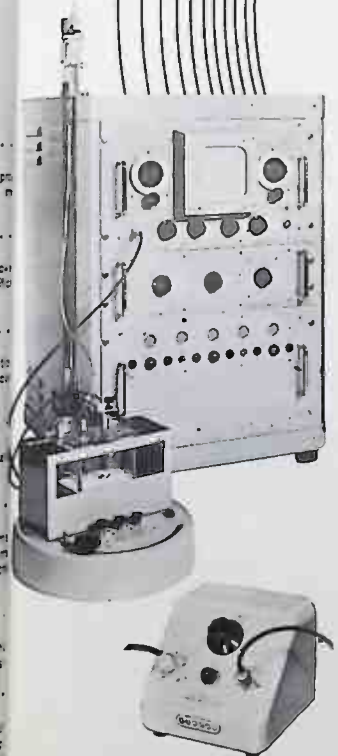
GENERAL ELECTRIC

P.O. Box 132 Cincinnati 15, Ohio

A number of additional positions in areas described above are also available to candidates with a Master's Degree.

A COMPLETE POLAROGRAM

Every Seven Seconds...



Polarotrace is a product of
SOUTHERN INSTRUMENTS
COMPUTER DIVISION
CAMBERLY, SURREY, ENGLAND
DETAILED INFORMATION
AVAILABLE UPON REQUEST
Our application laboratory is at your disposal to run test specimens.

The entire change of potential is effected during the lifetime of a single mercury drop...and the entire polarogram is reproduced on the screen once every 7 seconds, cutting analysis time to a fraction!

The Cathode Ray
POLAROTRACE
WITH ELECTRODE STAND

Determines concentrations OVER A THOUSAND TIMES LOWER than conventional types of polarographic instruments.

For DIRECT or Derivative Operations • Faster • More Sensitive • Greater Resolution • Easier to Operate • Ruggedly Designed.

Suitable for research or control operations, this highly versatile instrument opens up a new world of possibilities in polarographic analysis. Introduced just a short time ago, Polarotrace is now standard equipment in many leading institutions.

Now available with ANODIC CONVERSION UNIT to make possible measurement of anodic waves. Camera attachments also available.

Cat. No. AS-83725—POLAROTRACE, with Electrode Stand which includes a thermostatically-controlled tank and built-in gas manifold with stopcocks for simultaneous degassing of three cells, with complete set of glassware and ready for use (115 Volts A.C.) $4,500.00

STANDARD SCIENTIFIC Supply Corp. 808 BROADWAY NEW YORK 3, N.Y.

STANDARD SCIENTIFIC SUPPLY CORP. • 1888 •

LABORATORY APPARATUS
REAGENTS AND CHEMICALS

(*Continued from page 34A*)

government where universities and colleges cannot efficiently operate.

'With a single set of National Officers and a unified Board of Control, a large measure of duplicate effort could immediately be eliminated. The unit structure would enjoy a larger measure of prestige and influence in the promotion of a single objective: the encouragement of research.'

"President Taylor again called upon the officers of the Sigma Xi Chapters and Clubs to bring this matter to the attention of their local members for discussion, decision and action, and he proposed a joint Sigma Xi-RESA Convention to be held toward the end of that year to take action on the proposed consolidation and on the joint assumption of the normal duties of the two societies, including publications, annual lectures, award of the Procter prize, etc.

"The results of the circularization of the above-mentioned letter were announced at the 1952 Convention of Sigma Xi as indication that more information and more clarity of thought by the membership of Sigma Xi were needed before any action would be in order, and it was stated that the officers of the Society had decided not to take any further steps until the matter had reached a point of greater clarity in the minds of the members. The national officers had no thought at that time and have no thought now of pressing this matter upon the members of the Society, but the officers more frequently than any one else are faced with awkward situations that grow out of the present relationships of these two Societies.

"In a period of ten years since the approval of its constitution, RESA has become a flourishing organization with over 8000 members in about fifty Branches distributed from New England to California. Whether the members of that organization would be interested in a more intimate relationship with Sigma Xi than it now has is a matter quite unknown to me; however, there needs to be more adequate justification than I am able to see for the existence side by side of two organizations devoted to the same purpose, unless there is a basis of intimate cooperation between the two. My own background is a rather strictly academic one, and I am sure I embody the academic point of view and tradition; but the stated purpose of Sigma Xi is that of 'encouraging original investigations in science, both pure and applied,' and not that of promoting the academic tradition. It would seem unfortunate and unnecessary that rival organizations, or even cooperating organizations with separate offices and management, should come to exist where aims are identical and where the interests of the people involved are fundamentally the same. Therefore, I venture to suggest that we would do well to return to the question that was left unanswered in 1952 and give consideration to the possibility and the desirability of extending our traditional purpose beyond the limits of academic institutions through a merger of Sigma Xi and RESA.

"In closing my remarks please permit me again to express my sincere appreciation to every person who has contributed in any manner to the continuing progress of the Society of the Sigma Xi during the past two years. It has been a great pleasure to me to work with the organization."

Report of the Executive Secretary

Secretary Holme stated that the majority of his report would be made under other items on the agenda. He then gave a brief summary of the Society's operation and growth over the past five years. Calling attention to President Boyd's reference to 133 chapters, the Secretary explained the apparent inactivity of one chapter since 134 chapters had been installed. This "loss" of one chapter resulted from the merger of the Radcliffe College and Harvard University Chapters into the Harvard-Radcliffe Chapter. One operating problem was explained in detail—that of the canvass by the Committee on Membership-at-Large of Members and Associate Members dropped by chapters and clubs from active membership. The Secretary reported that some chapter secretaries misunderstood and felt this circularization was intended to proselyte on behalf of the Membership-at-Large at the expense of the chapters. This, however, was not the case as the Executive Committee and the Committee on Membership-at-Large felt that each member, if at all possible, should be active with a chapter or club to get, and to give, the maximum benefits of Society membership. It was imperative, however, that contact be made with a member just as soon as notice was received that he was not currently active with a chapter or club. The Secretary further stated that National Headquarters would be happy to supply any chapter or club with a list of those active in a given geographical area for purposes of invitation to be active with a particular group. The Secretary concluded with the statement that the Society was in a most healthy condition.

Report of the Treasurer

Treasurer Prentice announced that the complete and audited report for 1958 would appear in the March issue of the AMERICAN SCIENTIST. He wished to explain that any comparison of 1957 and 1958 might be somewhat misleading since one of the five issues of the AMERICAN SCIENTIST published in 1957 was paid for in 1958. The Treasurer also explained about the operation reserve which should reach \$100,000 by the end of 1958, and eventually \$140,000–\$150,000. Dr. Prentice stated he, too, wished to comment, as had the Executive Secretary, on the healthy growth of the Society. He did, however, feel that one feature of this growth—the great expansion of the Membership-at-Large (to 20,000 for 1959)—should be subject for some thought. This great segment of the Society's membership was not even represented at the Convention. This group, approaching one-third of those active—and responsible for 90% of the contributions to the research funds—has no voice in shaping policy for the organization. Dr. Prentice stated that this situation should be studied for possible future correction.

Dr. Prentice then reminded the Convention of its action at the 58th Annual Convention:

It was *VOTED that:*

The National Assessment for 1959 be increased from \$1.50 to \$2.00.

and presented recommendations of the National Executive Committee for Convention action as follows:

It was *VOTED that:*

1. The annual assessment on each chapter shall be due and payable on January 1, 1959, and that the amount of the assessment on each chapter shall be \$2.00 multiplied by the number of Members and Associate Members on the active membership roll of the chapter as of January 1, 1959.
2. The annual assessment on each club shall be due and payable on January 1, 1959, and that the amount of the assessment on each club shall be \$2.00 multiplied by the number of Members and Associate Members on the active membership roll of the club as of January 1, 1959.
3. That in sending notice of the 1959 assessment to chapter and club treasurers, the Treasurer of the Society be instructed to advise each chapter and club that the assessment is to be computed strictly on the number of Members and Associate Members on the active membership rolls as of January 1 without regard to whether Members and Associate Members have or have not paid current chapter or club dues, and to explain that this method of fixing the amount of the assessment on each chapter and club has been adopted by the Convention of the Society as the most equitable to all chapters and clubs.

Further, that each chapter and club should be informed that "the active membership roll as of January 1" is defined as the IBM Membership List as mailed to the chapter or club on or about December 15, 1958, modified by additions and deletions reported to National Headquarters prior to

January 15, 1959. If the edited IBM Membership List is not returned to National Headquarters by January 15, assessment will be due for all names on list as mailed by National Headquarters on December 15 and for all subsequent additions.

4. The annual assessment on each Member-at-Large and Associate Member-at-Large shall be $2.00 and shall be due and payable on January 1, 1959.

Regarding initiation fees, Dr. Prentice explained that an adjustment was necessary to compensate for certain changes in costs.

1. Cost of annual subscription to the AMERICAN SCIENTIST has increased for 1959 by 50¢—this is reflected in the increased National Assessments and individual subscriptions.
2. Cost of Initiate Booklet for each new initiate (20¢) to date has been borne by the National Society, but this should now be shifted to the new initiate.
3. Cost of individual certificate to Society is scheduled to increase by 15¢ in 1959.
4. Costs of the Society's operation are up and contribution of initiate to these costs should increase by 50¢.

These increased costs for 1959 were compared with those of 1958 as follows:

	1958	*1959*
AMERICAN SCIENTIST	$1.00	$1.50
Initiate Booklet	No charge	0.20
Certificate	0.65	0.80
Contribution to operation of Society	1.00	1.50
	$2.65	$4.00

The Treasurer then explained that the Executive Committee proposed that in 1959 a single increased initiation fee of $4.00 replace the present initial fee and certificate charge. Upon motion of the Treasurer,

It was *VOTED that:*

The initiation fee for 1959 be $4.00 per new Member or new Associate Member and include a subscription to the AMERICAN SCIENTIST, a certificate of appropriate grade of membership, and an initiate booklet.

After further explanation,

It was *VOTED that:*

The initiation fee for each Promoted Member (from former Associate Member status) be $1.00 and include a membership certificate and a current initiate pamphlet.

Upon completion and discussion of the Treasurer's Report,

It was *VOTED that:*

The Treasurer's informal report be accepted subject to audit and publication upon completion of the fiscal year of 1958.

Reports and Recommendations from the Executive Committee

Petitions for Chapter Status: In accordance with the recommendation of the Executive Committee,

It was *VOTED that:*

The Petition for a chapter to be known as the University of Delaware Chapter be *GRANTED.*

Following a discussion of the Executive Committee's motion that the petition for a chapter at Texas Technological College be *approved,*

It was *VOTED that:*

The Petition for a chapter to be known as the Texas Technological Chapter be *TABLED.*

Report of Editor-in-Chief of the American Scientist

Dr. Hugh S. Taylor, Editor-in-Chief of the AMERICAN SCIENTIST requested a show of hands which disclosed that approximately three-quarters of the delegates had received the December AMERICAN SCIENTIST prior to the Convention. The Editor stated that he was sorry a delay at the publishers had prevented everyone from having his copy in time but henceforth even greater effort would be made to have the December issue containing the Convention agenda out prior to the Christmas mail.

Dr. Taylor reported further that with over 80,000 copies of the American Scientist being mailed out four times in 1958, he found the membership of Sigma Xi most inarticulate. Although the rank and file were not writing in complaints, the Editorial Board was not satisfied with the publication and improvements could be made. The Editor felt that the fault was not with the Board but with the membership—not enough material was being submitted on experimental research in progress.

In 1957 with 5 issues, each issue had averaged 178 pages—94 pages of text and 84 pages of *News & Views*, *Bookshelf*, etc. In 1958 with 4 issues—each issue had averaged 200 pages—106 of text and 94 pages of *News & Views*, *Bookshelf*, etc.

Dean Taylor concluded his report with the statement that the Board of Editors had proposals to make to the Executive Committee which it hoped would result in better articles for the American Scientist, and with the support of the Executive Committee and the membership of the Society he expected a more cheerful report next year.

Reports of Other Committees

Membership-at-Large: In the absence of Dr. Frank Stanton, Chairman of the Committee on Membership-at-Large, the Executive Secretary reported that the Membership-at-Large in 1958 continued to increase and was now at a record 17,000 (2000 more than in 1957) with the prospect of reaching 20,000 in 1959.

The Secretary reported further that this group continued to give the major support to the Society's Grant-in-Aid of Research program, and he concurred in the Treasurer's suggestion that some arrangement should be made whereby this group could have a greater voice in the affairs of the Society.

National Lectureships: Dr. Frank M. Carpenter, Chairman of the Committee on National Lectureships, presented the following report:

"1. Report on Lectures for 1957–58: One hundred and seventy groups of Sigma Xi and RESA requested Lecturers; one hundred and sixty-two groups actually had lectures given, the remaining few having found it impossible to meet the Lecturer's schedules. The Lecturers and their subjects were as follows:

Northeast Tour:
Professor A. C. Zettlemoyer, Lehigh University
"Molecular Interactions with the Surfaces of Solids"
Mid-Atlantic Tour:
Dr. E. Cuyler Hammond, Yale University
"Smoking and Death Rates—a Riddle in Cause and Effect"
Mid-Western Tour:
Professor John Turkevich, Princeton University
"The World of Fine Particles"
Southern Tour:
Dr. Joseph W. Beard, Duke University
"Viruses as a Cause of Cancer"
Plains Tour:
Professor J. T. Wilson, Toronto University
"Geophysics and Continental Evolution"
Pacific Tour:
Professor A. C. Redfield, Woods Hole
"The Proportion of Things in the Sea"

2. Report on the Lectures for 1958–59: One hundred and sixty-five groups requested the Lecturers; one hundred and fifty-six groups are scheduled for lectures. The Lecturers and their subjects are as follows:

Northeast Tour:
Professor Harry Harlow, University of Wisconsin
"Intellectual Development of the Infant Monkey"
Mid-Atlantic Tour:
Professor Paul Delahay, Louisiana State University
"Electrochemistry and Kinetics"
Mid-Western Tour:
Dr. Bentley Glass, Johns Hopkins University
"Genes and the Man—New Vistas"

(*Continued on page 130*)

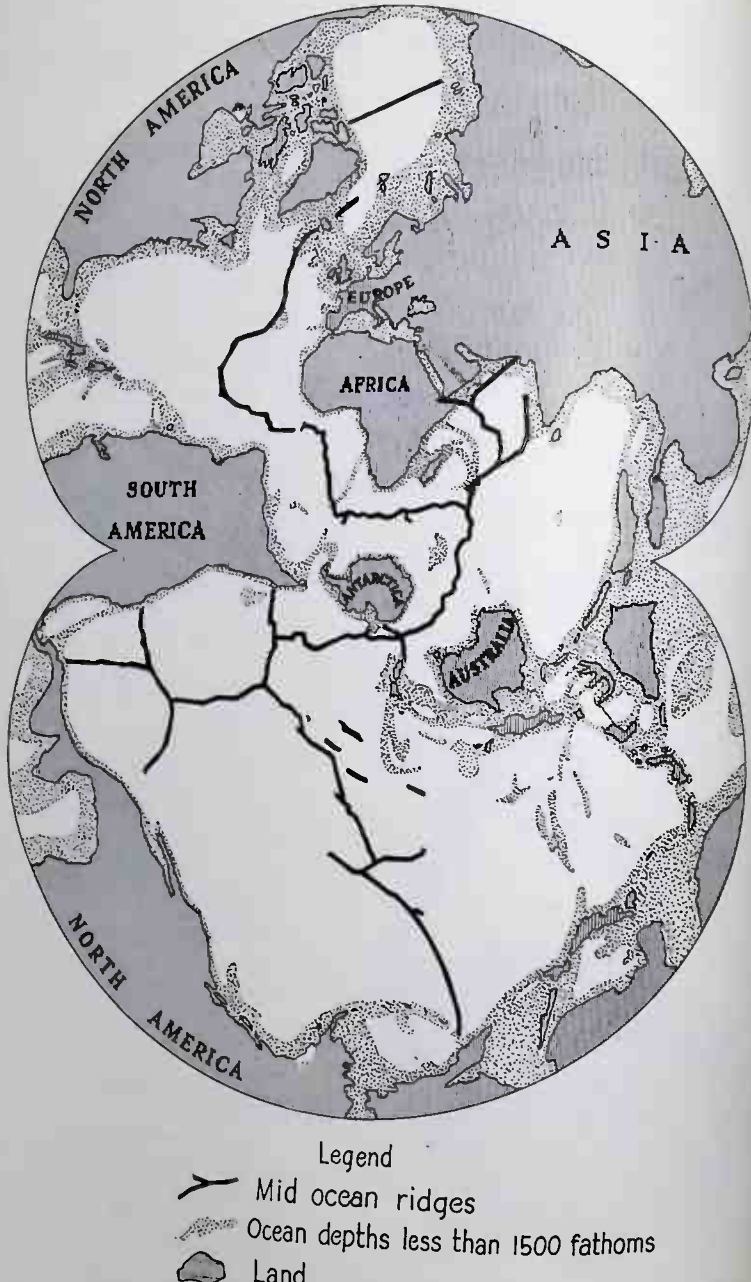


FIG. 2. FRONTISPIECE. It seems probable that a continuous ridge a few hundred miles wide, from 10,000 to 33,000 feet high and perhaps 40,000 miles long, winds its way across the central ocean floors. The discovery, only two years ago, of the continuity of this, by far the greatest mountain range on Earth, is an interesting commentary upon the state of our knowledge about some parts of the Earth. (Invert in order to view the Pacific Ocean clearly).

AMERICAN SCIENTIST

SPRING MARCH 1959

GEOPHYSICS AND CONTINENTAL GROWTH[1]

By J. TUZO WILSON

IS IT not remarkable that theories to explain the internal motions and behavior of stars and atoms are better developed and more widely accepted than similar theories about the Earth? The easy accessibility of the Earth compared with the remoteness of stars and atoms suggests that the reverse should be the case. The explanation lies partly in the fact that many stars and atoms radiate strongly and so reveal their nature, and partly in the great numbers of atoms and stars which enable us sometimes to observe one of them doing something unusual, such as undergoing a radioactive disintegration or exploding to form a nova. These disruptions are even more revealing. On the other hand, there is only one Earth; it radiates heat but feebly, and it is very stable.

Many geologists, when asked how theories of the Earth's behavior and of mountain building should be sought, reply that more and more detailed geological mapping is needed. When one considers that in a country so well mapped as Switzerland, the competent Swiss geologists have often been unable to predict in advance the precise arrangement of the strata to be found in railway tunnels a few thousand feet below the surface, one is forced to conclude that geological observations, while valid enough, are not of a type that lend themselves to extensive extrapolation. Furthermore, when it is realized that atomic, nuclear, and stellar theories were developed with no knowledge whatsoever of the surface details of atoms or of any stars except perhaps the sun, the value of more detailed mapping in solving this problem seems doubtful however valuable the mapping may be for other purposes.

The principal motions of the Earth are not surface phenomena, although they influence the surface. Since geological data give a poor indication of the internal nature of the Earth, additional information of other kinds is needed. Until the last few decades this problem was

[1] A Sigma Xi National Lecture, 1957–58.

FIG. 2. FRONTISPIECE. It seems probable that a continuous ridge a few hundred miles wide, from 10,000 to 33,000 feet high and perhaps 40,000 miles long, winds its way across the central ocean floors. The discovery, only two years ago, of the continuity of this, by far the greatest mountain range on Earth, is an interesting commentary upon the state of our knowledge about some parts of the Earth. (Invert in order to view the Pacific Ocean clearly).

AMERICAN SCIENTIST

SPRING MARCH 1959

GEOPHYSICS AND CONTINENTAL GROWTH[1]

By J. TUZO WILSON

IS IT not remarkable that theories to explain the internal motions and behavior of stars and atoms are better developed and more widely accepted than similar theories about the Earth? The easy accessibility of the Earth compared with the remoteness of stars and atoms suggests that the reverse should be the case. The explanation lies partly in the fact that many stars and atoms radiate strongly and so reveal their nature, and partly in the great numbers of atoms and stars which enable us sometimes to observe one of them doing something unusual, such as undergoing a radioactive disintegration or exploding to form a nova. These disruptions are even more revealing. On the other hand, there is only one Earth; it radiates heat but feebly, and it is very stable.

Many geologists, when asked how theories of the Earth's behavior and of mountain building should be sought, reply that more and more detailed geological mapping is needed. When one considers that in a country so well mapped as Switzerland, the competent Swiss geologists have often been unable to predict in advance the precise arrangement of the strata to be found in railway tunnels a few thousand feet below the surface, one is forced to conclude that geological observations, while valid enough, are not of a type that lend themselves to extensive extrapolation. Furthermore, when it is realized that atomic, nuclear, and stellar theories were developed with no knowledge whatsoever of the surface details of atoms or of any stars except perhaps the sun, the value of more detailed mapping in solving this problem seems doubtful however valuable the mapping may be for other purposes.

The principal motions of the Earth are not surface phenomena, although they influence the surface. Since geological data give a poor indication of the internal nature of the Earth, additional information of other kinds is needed. Until the last few decades this problem was

[1] A Sigma Xi National Lecture, 1957–58.

intractable because no other information was available, but, recently, several geophysical methods have been developed which tell about the interior and which can greatly supplement observations of the surface.

The irregular features mapped on the Earth's surface by geologists appear to be complex boundary phenomena which owe their irregularity to secondary causes such as erosion and transportation by water at the interface between the solid Earth and its fluid envelopes. This irregularity may serve to disguise the real simplicity of the Earth's internal motions.

If we could arrive at a solution to this problem and discover a valid physical theory for the behavior of the Earth, then we could tackle geological problems the other way about, using the theory as an aid in interpreting the complex observations made by geologists. This might be of considerable assistance in predicting underground structures and in suggesting places in which to look for ore bodies.

A belief that the Earth does operate in a manner which can be generalized in some simple way is justified because that is usual in nature. The Earth, furthermore, almost certainly operates in accordance with laws of classical physics which are already well known. We can say this with considerable assurance because, during the investigation of nuclear and stellar behavior, the limits have been discovered beyond which the laws of classical physics must be modified to take into account relativity and nuclear phenomena. Any conceivable state of the interior of the Earth lies well within these limits.

If this argument is correct, it might be thought that we should be able immediately to apply classical physics to suggest the way in which the Earth behaves. The matter, however, is not quite so simple. Unfortunately, we do not know the composition nor temperature of the different parts of the Earth's interior at all precisely, and the pressures prevailing in most of the interior are a few times greater than the highest which can be achieved in the laboratory.

Furthermore, the Earth could conceivably operate in any one of a number of different ways, and we need to examine the evidence for clues about the nature of these movements.

It may help us if we realize that the Earth is large, so that it has a very great heat capacity and, hence, its behavior in the past probably differed little from and changed but slowly to its present behavior. No theory is likely to be valid unless it can be shown to be applicable to ancient geological time as well as recent time.

In this article an analysis of the principal features of the Earth will be attempted with the thought that the main internal movements of the Earth may be relatively simple in character, that they are similar now to what they were in the past, that they should be subject to known physical laws and that it may be possible to discern them if we consider all the evidence already available.

The chief kinds of evidence available to us are the following:

(1) Topographical shape: lands; bathymetric shape: ocean floors

(2) Geology of the continents and of the ocean basins. In examining the geology we shall seek to find gross similarities and repetitions rather than to emphasize the true but secondary differences

(3) Seismic evidence of the thickness and distribution of subsurface layers and structures

(4) Seismic evidence of the distribution and direction of movement of earthquakes which give direct evidence of internal movements

(5) Absolute age determinations which provide a time scale and which can be greatly supplemented during the last one ninth of the Earth's history by relative paleontological age determinations

(6) Irregularities in the Earth's magnetic and gravity fields which, to the extent that they can be interpreted, provide evidence of the internal nature of the Earth

(7) Studies of flow of heat from the Earth

(8) Examination of the rates of emission of matter from inside the Earth through volcanoes and of the rates of other geological processes.

The relationship existing between the Earth and earth scientists may be compared with that existing between a patient and his doctors. In former centuries, a patient was often only examined by a doctor who looked at the surface of his face and those parts of his body not covered by clothing. Recently, it has been usual to observe all parts of the body and to investigate the interior by means of stethoscopes, x-rays, encephalographs, and laboratory tests. The description of the human body by anatomists has been extended by the study of its workings by physiologists. We need to realize that, although the detailed study of the anatomy of the land surface of the Earth remains today a valid, useful, and necessary pursuit, a proper understanding of the Earth demands much more, and it should be emphasized that, until we can discover the nature of the Earth's inner workings, we are no farther in the study of the Earth than physicians were in the study of the body before Harvey discovered the circulation of the blood.

If geologists are trained to be the anatomists of earth science, and if geophysicists are all too often thought of merely as specialists in some physical methods of prospecting, is there not a danger that too little attention is being given to the really basic part of earth science which is the physics or physiology of the Earth? How can we hope to make progress by just describing or measuring an Earth which we do not understand? We are now beginning to have some knowledge of the interior of the Earth and of its behavior, and of the formerly unknown three-quarters of the Earth's surface covered by ocean. Is it not time to study the whole Earth and make physics of the Earth the basis of our teaching and research?

The Principal Divisions of the Earth

The chief divisions of the interior of the Earth have been determined by studies of the manner in which the elastic waves generated by large earthquakes are reflected and refracted inside the Earth. These divisions are the liquid core, the solid mantle, and the solid crust and they are concentrically arranged in about the same proportions as the yolk, white, and shell are arranged in a soft-boiled egg, except that the crust is proportionately thinner than the shell.

The core is probably made of hot, liquid nickel-iron, and it may have a small solid part at its centre. It is very probably the source of the Earth's main magnetic field, but there is no evidence that it has much influence upon the Earth's surface features which are portrayed in the crust and molded by the actions of the outer part of the mantle.

The composition of the mantle is an intriguing question which cannot be quite definitely answered because it is not sure that any samples of the mantle have been seen. The suggestion has been seriously made that a deep hole should be drilled through the crust to obtain samples of the upper mantle, but that has not yet been done. Meanwhile the mantle is generally thought to consist of impure magnesium and iron silicates containing about 40 per cent SiO_2 and known as ultra-basic rocks. These would have the required density and elastic properties and resemble stony meteorites which may be fragments of another disrupted planet.

There is evidence that the mantle is concentrically layered, but each layer seems to be rather uniform. In the crust the situation is quite otherwise, for it varies a great deal in thickness and composition from place to place. Four main types of crust can be distinguished. Two types, the continental shields and the ocean floors, form stable patches which are crossed by narrower bands of greater activity, the mid-ocean ridges and the belts of young mountain and island arcs.

The Ocean Floors

The ocean floors are the simplest and perhaps the oldest type of crust and cover more than half of the Earth's surface. In those places where the thickness has been measured, the oceanic crust consists of a layer 5 km thick of rock through which seismic waves pass with a velocity of about 7 km/sec. At the Mohorovičić discontinuity forming the base of this layer, this velocity changes abruptly to 8.2 km/sec which is characteristic of the mantle. This layer is almost certainly basalt, the common lava of the ocean basins, having about 50 per cent SiO_2. Upon this layer rests a deposit of sediment of variable thickness, but usually only a few hundred meters thick, in part deposited by slow sedimentation and in part swept in from the continents by occasional rushes of muddy water impelled by gravity down the steep submarine slopes of continental margins and known as turbidity currents. The surface of the ocean floors is generally nearly 5 km below sea level (see Fig. 1).

The Mid-Ocean Ridges and Mid-Ocean Fracture System

Resting upon the ocean floors are the mid-ocean ridges, for a century vaguely known to exist around some remote islands, but so difficult and expensive to explore that their true nature is only being revealed during the International Geophysical Year. Ten years ago, Gutenberg and Richter published maps showing that most of the known ridges were the loci of shallow earthquakes. Ewing and Heezen, in 1956, combined the bathymetric and seismic evidence and first suggested that the mid-ocean ridges form a continuous system about the earth (see Fig. 2). This extends down the Atlantic Ocean from Iceland through the Azores, Ascension Island, and Tristan da Cunha. It turns to pass between Africa and Antarctica into the central Indian Ocean, whence branches reach north to the Gulf of Aden and to Pakistan and perhaps to the west coast of India. The main ridge continues south of Australia and New Zealand to cross the Pacific Ocean to Easter Island. Here it is least known but branches probably extend to South America through Juan Fernandez and Galapagos Islands and perhaps to Mexico and also through the Line Islands to the Hawaiian ridge and on probably to Kamchatka and toward the western Pacific. Another part may form the Lomonosov ridge discovered by scientists of the U.S.S.R. to cross the Arctic Sea. Thus, it seems probable that a continuous ridge a few hundred miles wide, from 10,000 to 33,000 feet high and perhaps 40,000 miles long, winds its way across the central ocean floors. The discovery, only two years ago, of the continuity of this, by far the greatest mountain range on Earth, is an interesting commentary upon the state of our knowledge about some parts of the Earth.

Most of this ridge is seismically active and, indeed, these earthquakes form the second most active seismic system on Earth. All of these earthquakes have shallow foci at depths of not more than 70 km. Where the ridge has been most closely studied, it has been found to have a central rift. Both rift and earthquakes suggest that the ridge follows a fault zone which we shall call the *mid-ocean fracture system*. The ridge is made up of basalt lava with some patches of ultra-basic rock and the islands along it are basalt volcanoes. All this suggests that the ridge has been formed by the escape of basalt lava along a fracture zone and that basalt has been formed by partial melting of pockets in the ultra-basic rock of the upper mantle at a depth of not more than 70 km. The lack of abandoned ridges and the slow rate of the ridge's volcanism suggest that it has been in its present position for a very long time, perhaps most of the Earth's history. Suggestions by proponents of isostasy and specifically by Betz and Hess, measurements of gravity by F. A. Vening

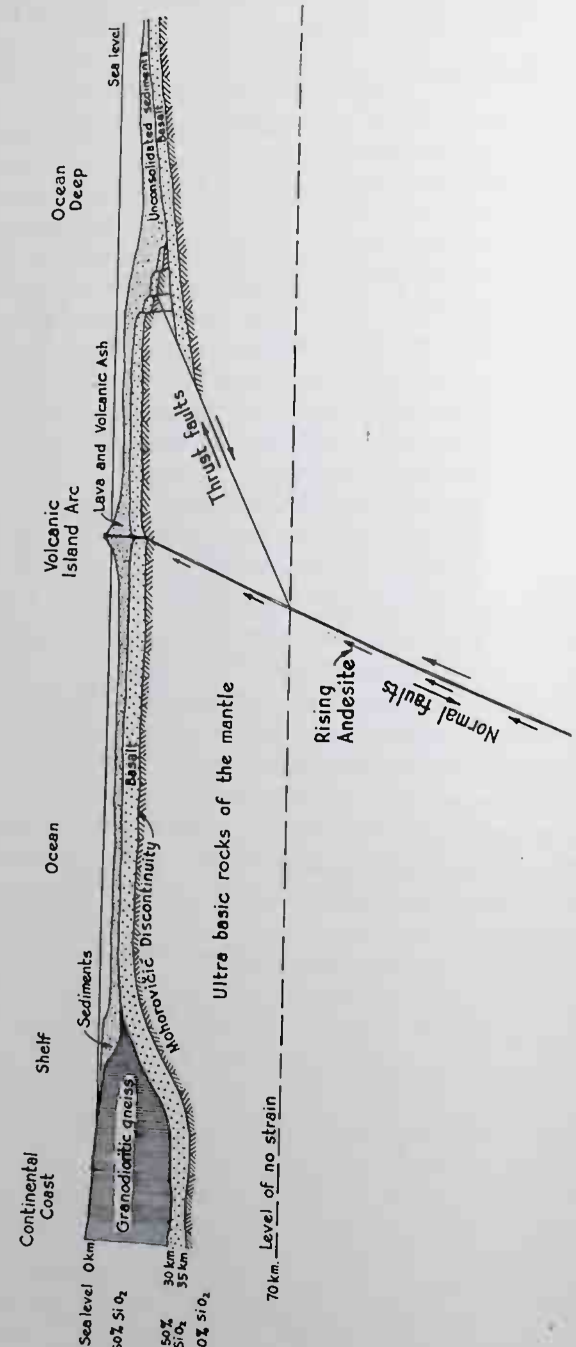


FIG. 1. Cross section across an active single island arc, the surrounding ocean floor and the adjacent margin of a continent.

Meinesz, G. P. Woollard, and M. Ewing and some marginal seismic data by T. F. Gaskell suggest that the ridge sank as it was formed and has, beneath it, a root perhaps as much as 30 km deep and that it is in hydrostatic balance.

The Active Mountain and Island Arcs and the Continental Fracture System

The other and more active system of earthquakes and volcanoes was recognized by E. Suess about 1880 and forms the active mountain system which, at present, forms two great belts about the Earth, each very roughly part of a great circle (Fig. 3). One passes from Antarctica

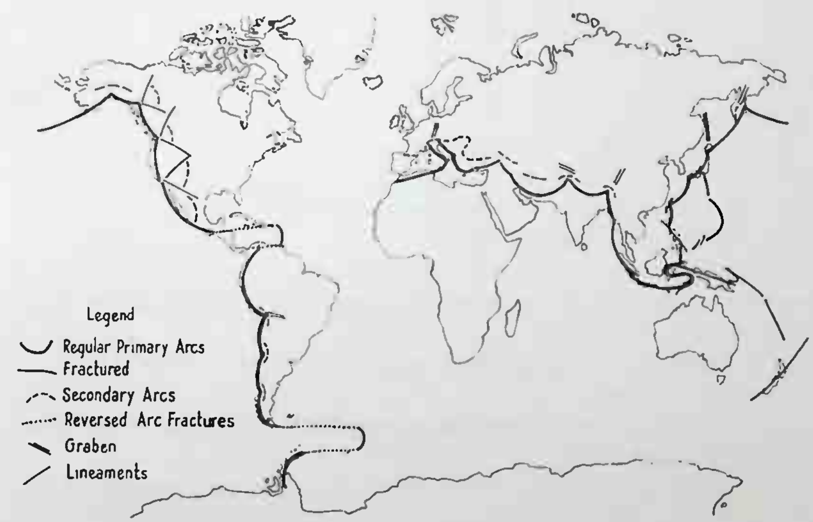


FIG. 3. Active mountain and island arcs and the continental fracture system.

through the Andes, the Cordillera, and the East Asian island arcs to Indonesia. The other, like the cross of a "T," extends from the Mediterranean through southern Asia, past Indonesia, New Guinea, the Solomons, and New Hebrides Islands to New Zealand. The two belts meet at right angles in Celebes Island. This system has more earthquakes and volcanoes than the mid-ocean system. The earthquakes are deep as well as shallow and occur at depths to 700 km. It is suggested that the active young mountains and the island arcs are but the surface expression of another and greater fracture system about the Earth, and that it is a deeper and more active system than the mid-ocean fracture system. It will be called the *continental fracture system* because so much of it lies along the margin of continents, to which it is clearly related.

Consideration of this system suggests that, although the various parts differ greatly in topographical and geological aspect, some features keep reappearing, for example, the arcuate shape of island chains and mountains, often marked by volcanoes and certain characteristic distributions of earthquakes and of gravity anomalies. In the following analysis, the various parts of the system will be divided into two main types, called primary and secondary arcs, because one of them is more fundamental than the other and forms first.

It is suggested that the variations in appearance of individual arcs are due to the fact that both primary and secondary arcs evolve so that arcs of either type in different stages of growth present varying appearances. For example, it will be seen that the Himalaya Mountains are but a later stage in the growth of an arc which once resembled the Kuril Islands or the West Indies.

It will also be pointed out that these two types of arc may be combined in three different ways to form the main types of mountain and island arc systems.

It is suggested that this plan provides a rough method of analysis for this system of volcanically and seismically active mountains and that a physical explanation of these features would be at least an elementary theory of mountain building and behaviour of the Earth.

Primary Arcs are volcanic and igneous arcs, whether like the Kuril Islands or Himalaya Mountains, which have these distinguishing characteristics:

(1.) In shape, they are generally parts of circular arcs, concave toward the nearest continent.

(2.) They are marked by recent andesitic volcanic and older granodioritic igneous activity.

(3.) They are accompanied by all the greatest deep trenches in the oceans.

(4.) Parallel to them there are large negative gravity anomalies occurring along narrow strips.

(5.) Most of the world's shallow earthquakes and all of the world's deep earthquakes lie in zones beneath them.

(6.) They rest upon no basement of more ancient rocks.

In detail primary arcs differ in appearance so that we can immediately distinguish the several types listed in Table I. The first four of these are regular and follow the same general pattern. There are believed to be stages in an evolutionary sequence by which small island arcs grow through volcanism and uplift to great mountain ranges. The cross section of a typical primary arc has these parts (Fig. 1):

(1.) An undisturbed frontland, often the ocean floor.

(2.) A deep arcuate trench (as off the Kuril Islands) or arcuate islands or mountains of greywache facies (as at Kodiak Island, Alaska, or in the Coast Mountains of the western U.S.A.). Under these regions or on their continental side are large negative gravity anomalies, and along these regions occur shallow earthquakes.

(3.) An intermediate zone between (2) and (4) which is about 150 km wide and beneath which earthquakes occur at depths of several tens of kilometers.

(4.) An arc of volcanic or batholithic mountains, beneath which earthquakes occur at depth of about 100 km. This is parallel to (2) and on the continental or concaveside of (3). Examples are the Kuril Islands or the main Aleutian Island arc or the Cascade Mountains plus the Sierra Nevada Mountains in the United States.

(5.) Farther toward the continent the earthquakes extend to depths of as much as 700 km in conical zones of which (2) marks the surface expression.

TABLE I

THE TYPES OF ACTIVE PRIMARY ARCS OR ELEMENTS AND THEIR CHIEF DISTINGUISHING FEATURES

	Name	*Shape*	*Outer Part Primary Arc*	*Inner Igneous Part Primary Arc*	*Example*
1.	Single island arc	Circular arc	Trench	Volcanic islands	Kuril Islands
2a.	Double island arc	Circular arc	Sedimentary islands	Volcanic islands	Aleutian Islands at Kodiak Island
2b.	Single mountain arc	Circular arc	Trench	Volcanic and batholithic ranges	Western Central Andes
3.	Double mountain arc	Circular arc	Sedimentary ranges	Volcanic and batholithic ranges	Coast, Cascade and Sierra Nevada ranges of U.S. (combined)
4.	Fractured arc	Straight	Features are irregular		Solomon Islands

Secondary arcs are another type of arc, often occurring with primary arcs, but less volcanic, less igneous, and less metamorphosed. They are chiefly made of uplifted older basement and folded and thrust sediments. Because of their lesser metamorphism their stratigraphy is usually simpler and they are often better mapped and better understood than primary arcs. Although they form some of the great mountains of the world, they are here considered as secondary because they have not the geophysical evidence of deep connections. They lack large negative gravity anomalies; they rest upon no older basement; they have little volcanic or igneous activity. They are not the surface expression of deep conical zones of fracturing as indicated by earthquakes.

They generally lie on the continental side of primary arcs, facing in the reverse direction, that is concave toward the ocean. Each secondary arc tends to lie opposite the junction of a pair of primary arcs.

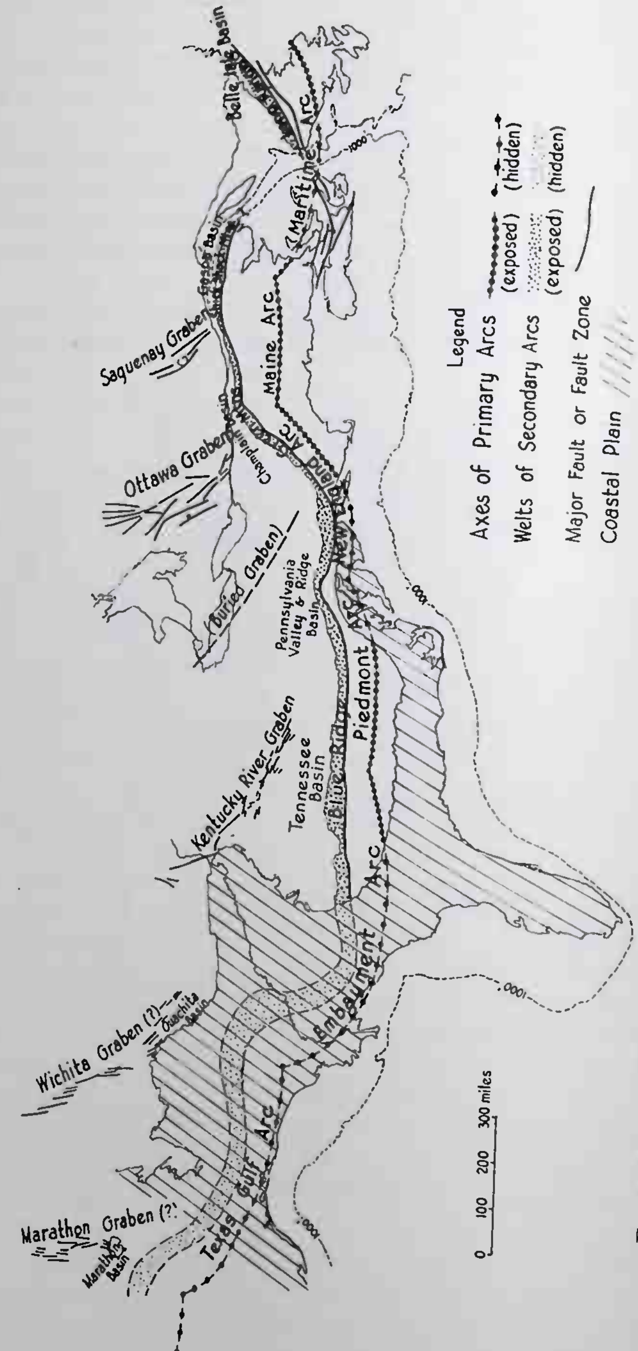


FIG. 4. The Appalachian Mountains, an inactive narrow-type mountain system with single graben junctions.

There appear to be two parts to each secondary arc; an uplifted massif or *welt* of older basement and a *geosynclinal basin* filled with folded, thrust, and uplifted sediments. The two parts are developed to different degrees.

Some examples of secondary arcs are the Alps, the Caucasus, probably the Pamirs and the high mountains of the Yunnan border, Taiwan, western (not eastern) Kamchatka, the Alaska range (which includes Mt. McKinley, the highest in North America), the several parts of the Rocky Mountains, and the eastern side of the Andes. The western side of the Andes is a series of primary arcs.

Anyone familiar with these mountains can in most cases distinguish the two parts—the geosynclinal basin is always on the convex side further from the primary arcs than is the welt. A very clear example is presented by Taiwan which has a welt of uplifted metamorphic rocks forming a high range of mountains along the central axis and eastern side of the island and a petroliferous geosyncline of Cenozoic sedimentary rocks on the western side.

The Appalachians provide other examples. For example, the Blue Ridge is the welt and the Valley and Ridge region the corresponding geosynclinal basin of a secondary arc. The White Mountains of New Hampshire and the Piedmont Region of the Carolinas are the roots of adjacent primary arcs (see Fig. 4).

Of *the three types of junction of arcs*, the simplest is the *reversed arc junction* where one primary arc is joined to two others facing in the opposite direction by large faults. Such is the case for the West Indies and for the South Sandwich Islands (see Fig. 3).

Another case is the *single graben junction* giving rise to the *narrow mountain system*. Here, the junction of two primary arcs is capped by a secondary arc beyond which one *graben* or broad structural valley radiates from the junction. The Appalachians provide examples shown in Figure 4 and so do Switzerland and Bolivia where the Alps and Puna block of the Andes form secondary arcs at the junction of primary arcs (in one case the Apennines and the Dinaric Mountains of Jugoslavia, in the other case the western volcanic Andes of Peru and of Chile). The two grabens are the well-known Rhine and Chiquitos grabens, respectively.

The third type of junction is the *double lineament junction* giving rise to a *broad mountain system*. In this case two lineaments radiate from the junction of two primary arcs and enclose the secondary arc between them at a distance of several hundred miles from the junction. A lineament may be considered to be a disturbed zone along which there is faulting and changes occur in facies and in structure.

The Cordillera of North America provide the best examples. The mountains along the west coasts of British Columbia and of the United States form two primary arcs meeting near Seattle. Thence, two lineaments extend inland respectively east-southeast to the Black Hills and north-northeast up the Fraser River. These two lineaments embrace and cut off either end of the secondary arc which forms the Canadian Rockies along the boundary between British Columbia and Alberta (see Fig. 3).

Thus to summarize, the continental fracture system extends from Antarctica to the West Indies as a series of primary arcs joined at either end by two reversed arcs and in South America by one clearly developed graben junction in Bolivia and one or two less clearly developed ones

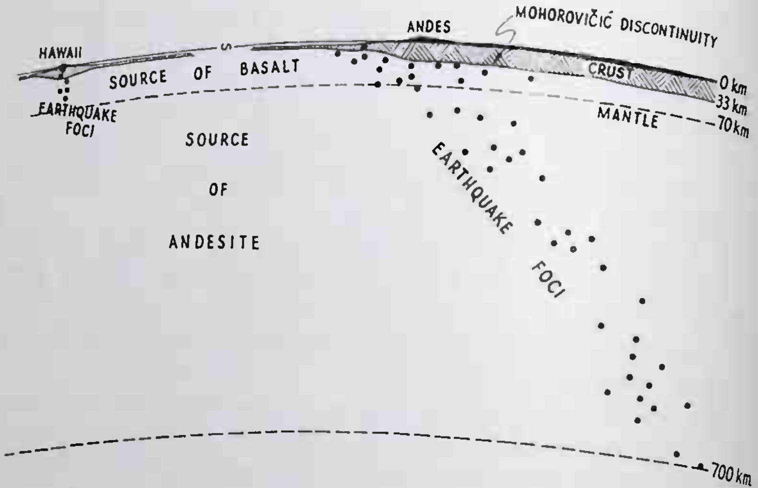


FIG. 5. Cross section of crust and upper mantle. (No vertical exaggeration.)

farther south. In North America, four primary arcs in Mexico, the United States, British Columbia and the Yukon are joined by three double lineament junctions. In Alaska and all the way along east Asia to the Philippines are graben junctions with the complication that, on either side of the Philippine Sea, there are two systems of primary arcs.

From the Philippines to New Zealand are fractured arcs with junctions hidden by the sea, which will be discussed later.

Across southern Eurasia the situation is complicated because there are continents on both sides of the system. But there appear to be graben junctions at the Alps, Caucasus, Pamirs, and in Yunnan.

The lavas emitted by the volcanoes of this system are both basalts and andesites, but the latter are more abundant. Andesites are lavas having about 60 per cent SiO_2.

The observation that shallow earthquakes are associated with only basalts in the mid-ocean ridges, while deep earthquakes are associated with both andesites and with some basalts along the continental fracture system suggests that all the lava is coming from the mantle and that the

source of andesite is deeper than 70 km and that the source of basalt is above 70 km. (This is shown diagrammatically in Figure 5.)

The Formation of Continents

It has long been recognized that andesites have about the same composition as the surface of continents and so it was formerly supposed that andesites arose along the margin of the continental blocks where they had been forced down to such depths that they had melted and were able to rise again as andesitic lavas. The continental blocks are a little over 30 km thick in most places but may thicken to twice that under some mountains. The upper part is gneiss overlain by a veneer of sediments.

In the western Pacific, a line, called the andesite line, was delineated just off the east Asian arcs which separated the andesitic island arcs from the mid-ocean basaltic islands, and some supposed that this marked the edge of the continent.

Seismic evidence, both from the study of waves generated by small explosions and refracted in the crust and from the study of surface earthquake waves of the type called *Lg* waves, has shown that the seas behind island arcs are like the ocean floor and are not part of the continental blocks. *Lg* waves will follow the surface of the Earth through continents, but are not transmitted by oceanic crust. They do not traverse the Bering Sea or the Tasman Sea, for example, any more than they will cross the Pacific Ocean.

It follows that, in the case of the Aleutians, New Zealand, and similar arcs, andesites are not being formed by remelting of continental crust. The thin ocean crust cannot form andesites. These lavas must be rising from the mantle. The lack of any andesites at all in those places where there are only shallow earthquakes suggests that the andesites rise from greater depths than the basalts. Many petrologists seem to object to this idea, but their objections appear to be based upon preconceived assumptions about the nature of the mantle and its origin and may not be valid.

It is of course not suggested that there is any layer of andesite itself inside the mantle. It is only proposed that the partial melting of a small fraction of the ultra-basic rocks of the mantle lying between 70 and 700 km produces andesite, which rises along the fractures associated with deep earthquakes. If only 5 per cent of this layer were partially melted, a layer of andesite approximately 30 km thick could be formed, and this happens to be about the thickness of the continental crust (see Fig. 5).

If andesites are not part of pre-existing continental crust which has been remelted, then they form new additions to the crust and provide the means for growth of continents.

The only careful study of the rate of volcanism was made by K. Sapper in 1927. He showed that, since 1500 A.D., the average rate of production of lava from all the world's volcanoes has been 0.8 km^3/year.

R. A. Daly and J. Verhoogen have suggested much lower figures, but this is because their estimates were based upon basalt only and upon the amount of basalt still preserved. Andesite is emitted by more volcanoes than is basalt, but, since it is usually in the form of ash or soft rock, it is quickly eroded and lost. A large error is introduced if this is not considered.

The emission of lava at the present rate of 0.8 km^3/year throughout the Earth's history of 4.5×10^9 years or even for the 3×10^9 years since the oldest known rocks were formed would have poured out lava of the order of 3×10^9 km^3 on the Earth's surface. This corresponds approximately to the volume of the continents (about 30 km times 1.1×10^8 km^2). A slightly higher rate of volcanism in the early stages of the Earth would allow for the emission of the oceanic crust as well.

It is of course not suggested that the continents are made of lava today. Especially, the more acid andesitic lavas which form the greater part have been repeatedly eroded, sorted, metamorphosed to granodioritic gneisses and eroded again and recycled. The alternative and older belief that the continents were formed by the accumulation or intrusion of large masses of coarse igneous granodiorite is faced with the difficulty that that process, added to the rise of lava, would have formed too much crust, and that it provides no method of disposing of the sedimentary rocks of past ages. According to any theory of uniformitarianism, vastly more sediments should have been formed in Precambrian than in subsequent time and they must have been destroyed since they are not now preserved. The simplest way to dispose of the sedimentary continental shelf and island arc deposits of former ages is to metamorphose them into coastal mountains. The geochemistry presents no problem if it is realized that the well-sorted sedimentary rocks exposed on continents are not typical. On the contrary, around the margins of continents, andesites break down to greywackes which are readily converted to gneisses of the same (i.e., granodioritic) composition.

Another reason for believing in the fresh generation of lava follows from the arguments of W. W. Rubey and others that all sea water is of volcanic origin and the similar arguments that the composition of the Earth's atmosphere indicates that the original atmosphere must have been lost and the present one emitted later. The fact that volcanoes emit lava as well as steam and other gases suggests that the continents as well as the oceans and the atmosphere may have been formed by volcanic activity. This would go a long way toward explaining the irregularity of the crust.

The Contraction Theory

If this happened, it follows that the Mohorovičić discontinuity represents the original surface of the Earth. Since this original surface is now overlain by an average thickness of 15 km of crust, it must have shrunk or been reduced in radius by that amount. The emission of the crust would therefore have produced about 100 km shortening in the circumference of the original surface which would be available to cause mountain building. If, in addition, the Earth is cooling, the estimates of D. W. Allan and J. A. Jacobs are that, at most, there might have been a further 50 km of shortening in any circumference, but this is not necessary. One can in fact visualize an Earth whose central parts are warming and expanding, whose outer parts may either be cooling or warming, but whose surface is in any case contracting as the result of the escape of volcanic matter. This indeed is the closest estimate which we can make at the present time to the probable behavior of the Earth.

An Explanation of Mountain Building in Terms of the Contraction Theory

Thus, it is suggested that the Earth, whether it is cooling or not, is contracting as the result of steady volcanism at about the present rate, and this contraction has produced two vast systems of fractures which form a network over its surface. One of the systems is feeble, shallow and hidden by the oceans; the other is active, deep and clearly the more important. No very striking characteristics have been observed in the mid-ocean system, but the continental system has one outstanding peculiarity: It is formed of a scalloped series of conical fractures. The basic problem of mountain building may thus be reduced in its simplest terms to one question in theoretical physics. How can a conical fracture be formed on the surface of a shrinking sphere? To this, A. E. Scheidegger has given an answer.

H. Jeffreys has discussed the behavior of a cooling Earth which he suggested should be thought of as consisting of three zones behaving in different ways. The innermost zone extending from the centre to within 700 km of the surface is, he thought, not changing in temperature nor volume. The next zone extending from 700 to 100 km is cooling most actively and hence contracting and stretching about the inner one. The outer zone above 100 km has already largely cooled and, since the intermediate zone beneath it is contracting, it is, as it were, losing support and hence is becoming compressed. The activity of the Earth is thus visualized as being chiefly due to the cooling and contraction of the intermediate zone.

This view is not altered in its essential points if we suppose that the activity of the Earth is due not necessarily to cooling but to emission of lava by, and hence shrinking of, the same intermediate zone. Nor is it altered if we use the levels at which earthquakes cease (700 km) an[d] which a marked change in their nature occurs (70 km) as the bound[aries] of the zones (see Fig. 6).

A. Ritsema has deduced from his studies of earthquakes that [the] stresses in the Earth are equal in all horizontal directions. A. E. Sch[ei]degger has shown that, under these circumstances according to [the] Mohr's theory, conical fracture can occur and that these cones sho[uld] dip at about 20° to 30° if caused by compression and dip at about 60°.

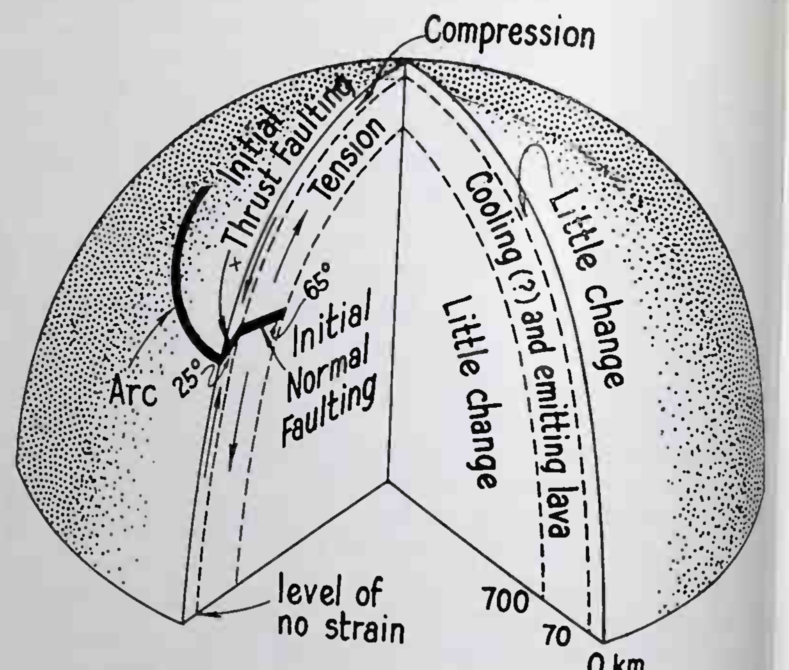


FIG. 6. Diagram of arc formation in a contracting Earth.

70° if caused by tension or reduction of pressure. Now the location o[f] earthquake foci is not very exact, but the evidence does suggest that th[e] shallower parts of most of the conical fracture zones below island ar[cs] do dip at about 30°, while the deeper parts do dip at about 60° whic[h] are the dips required to fit the contraction theory as postulated. Th[e] upper fractures were formed as thrust faults which may be seen on th[e] surface; the lower would have been normal faults. A. Ritsema from hi[s] studies finds evidence that some shallow and deep earthquakes do hav[e] the postulated directions of first motion, but to some extent he agre[es] with J. H. Hodgson that most earthquakes are due to neither normal nor thrust faults but to horizontally-moving shears called transcurrent faults.

This does not disprove Scheidegger's explanation because Mohr's theory and Scheidegger's only apply to movement at the moment of initial fracturing. Once the fractures have been formed, other forces come into play and different movements are channelled along the existing fractures. In particular, there are many arcs and they react on each other. Perhaps this is explanation enough for the rapid horizontal movements which can be observed at places on the Earth's crust, for example, along the San Andreas fault. Apparently, these movements are short-lived, geologically speaking, at any one place, for cases like the Great Glen fault across Scotland are known which once moved rapidly and which then became inactive.

In this connection we find an explanation by the contraction theory for the fractured arcs, each of which has along it evidence of a great transcurrent fault like the Alpine fault which lies along the axis of New Zealand.

It has been pointed out that the two active belts meet orthogonally to form a "T," folded about the Earth. Except for the small and irregular New Britain arc, all the arcs along one arm of the "T," from Celebes to New Zealand and no other arcs elsewhere are fractured arcs. It is not likely that this distribution is due to chance. The proposed explanation is that that arm, and only that arm, had to undergo two sets of movements approximately at right angles to each other, each corresponding to the general movements of one belt. The interaction of these forces has caused the complexity. This can be seen by reference to Figure 7, in which the two belts are shown diagrammatically. Along two of the three limbs of the "T," contraction below the level of no strain and compression above it can take place without horizontal shearing, but along the third arm shearing must accompany the shrinkage or compression. This can be demonstrated by arranging three books on a table separated by 1-inch gaps to form a T-shaped pattern and then moving them together or apart.

The direction of the fractured arcs will depend upon whether they are primarily formed by compression or by relief of pressure as is illustrated.

It is obvious from a comparison of Figures 7A and 7B with Figure 3 that the fractured arcs were formed by a relief of pressure, that is, in the zone between about 70 and 700 km, which has already been stated to be the source of the Earth's activity.

When it was pointed out that two principal types of mountain systems exist, Scheidegger was able to extend the contraction theory to explain the main properties of the narrow and broad types.

If, in the intermediate or contracting zone, two conical fractures meet and both are normal fault zones, then the two sides of these faults tend to move apart in the directions shown in Figure 8, but, at the junction, matter is being pulled in two different directions and this motion cannot occur unless there is at least one additional fracture.

Suppose first that there is only one additional fracture lying between normals to the arcs, then motion of the two sides of the fracture tends to create an opening along it as shown in Figure 8B. No opening can persist at those depths of from 70 to 700 km below the surface, so the rock above drops down and the single fracture appears at the surface as a large, down-faulted trench, known as a graben. It lies in a direction radiating from but not meeting the surface expression of the junction. In the compression zone above 70 km, a wrinkle forms between the junction and the graben. It forms the welt of the secondary arc like the Blue Ridge or the Green Mountains. The sediments tend to slide off its inner side as it rises and be crumpled in a geosynclinal basin

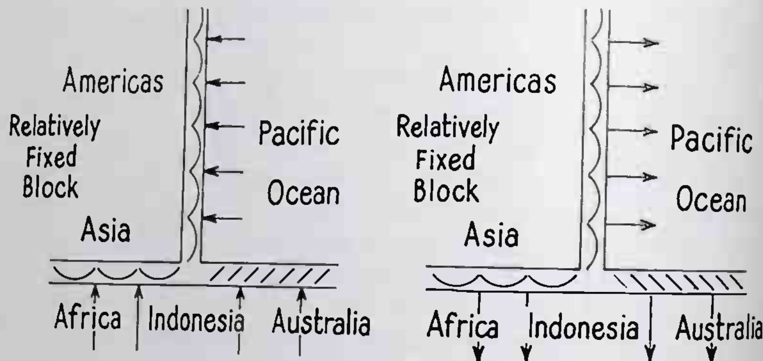


Fig. 7. Diagram of the pattern of continental fracture system which would result if created by forces (A) in the zone of compression above 70 km and (B) in the zone of relief of pressure between 70 and 700 km. The latter corresponds with Figure 3. Note the different orientation of the fractured arcs north of Australia.

on the continental side. Figures 4 and 8C illustrate this type of narrow system.

If, instead of only one additional fracture, two shears radiate out from the junction, then they do not form trenches or graben but dislocation zones, known as lineaments. Scheidegger has shown that, in this case, the secondary welt (and its accompanying geosynclinal basin full of slumped sediments) should lie at a considerable distance from the junction thus forming the broad type of range in which the primary and secondary arcs are well separated (see Fig. 8E).

Thus, the contraction theory can provide an explanation of many of the features observed in the Earth's main mountain system.

The transformation of island arcs into mountains appears to be largely due to metamorphism of the accumulated sedimentary rocks into gneisses after the manner suggested by V. Saull. The structure of batholiths and the uplift of mountains are consequences of this expansion *in situ*.

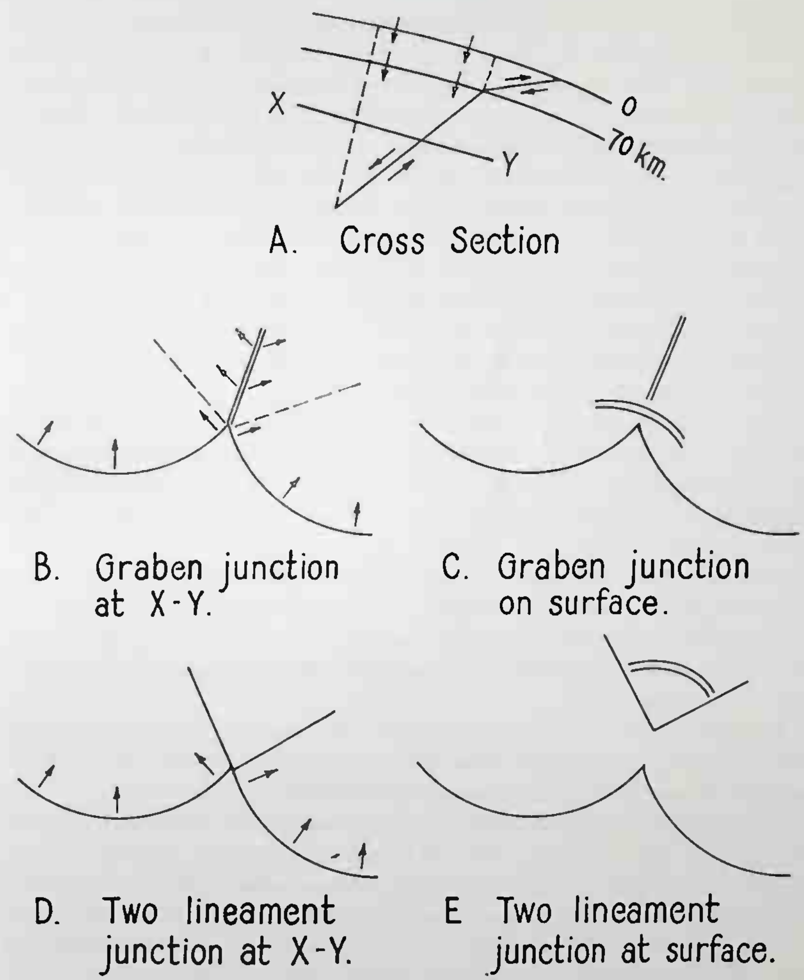


Fig. 8. Diagrammatic cross section through an island arc with plans of single graben and of two lineament junctions at depth and at the surface.

Inactive Mountains

It is apparent that not all parts of the active mountain system are of the same age. Some parts such as the Kuril and Aleutian islands may be only Cretaceous in age, but other parts such as Japan and British Columbia are at least early Paleozoic in age. It is probable that the position of the active belts moves. Now the belts are continuous at the present time and probably, as the earthquakes suggest, represent great fracture zones about the Earth. It may therefore be assumed that they were continuous in the past.

In that case we can suggest that, in about Cretaceous time when the Kuril and Aleutian arcs were formed, another part of the system was abandoned. This clearly lay from British Columbia through northwestern Canada across northern Alaska in the Brooks Range (which became inactive about Triassic time) and through the Paleozoic and Mesozoic mountains of northeastern Siberia to Japan.

Thus, the mountain systems can be thought to have changed position like a highway that has a detour in it or which is being shortened.

Another case of an abandoned mountain range is the Appalachians which became inactive in Permain time (last folding) or in Triassic time (last volcanism). Parts of it are now beneath the sea or are buried by the Atlantic coastal plain and that of the Gulf of Mexico which are Cretaceous and Tertiary in age. We can, however, still plainly trace the greater part of a narrow mountain system. The primary arcs, of which the roots are exposed as volcanic and igneous rocks, lay through central Newfoundland, Nova Scotia, New Brunswick, Maine, and the White Mountains of New Hampshire. They also formed the Piedmont region of the Carolinas (see Fig. 4).

The junctions of these arcs at which secondary arcs formed are in northern Newfoundland, in Gaspé of Quebec, in the Green Mountains of Vermont, in the valley and ridge and the Blue Ridge of Pennsylvania and of Tennessee, in the Ouachita and Marathon Mountains of Oklahoma and Texas.

Several of the graben are known; for example, along the Saguenay River, the Ottawa River and the Kentucky River and there are possibly two others west of the Wichita and Marathon areas.

The Appalachians are no longer active tectonically, seismically (except slightly) nor volcanically. They are being worn away.

If this process continues, we may expect to see the roots of the primary arcs and of the welts of the secondary arcs as a great belt of metamorphic rocks on the inner side of which lie some basins of little altered but folded sediments, representing the deepest part of the Valley and Ridge type of basins.

There might well be expected to be a major fault system separating the Appalachian Mountain system from the older basement of Precambrian age exposed in the Blue Ridge type of welts.

The Blue Ridge has been dated and is known to be of Precambrian, Grenville age (1.0×10^9 years). Thus, when the Appalachians are worn down, their roots can be expected to form two provinces of metamorphic rocks of different ages (Grenville and Paleozoic) which will be separated by a fault zone near which will be several basins of folded but little altered sedimentary rocks corresponding in age to the younger or Paleozoic metamorphics.

Precambrian Shields

When Precambrian shields were first studied, it was found that metamorphic rocks (which could not be easily mapped) were overlain with great unconformities by sedimentary rocks which were little altered and which were younger than the basement upon which they lay. These sedimentary rocks lay in separate basins. The assumption was made that all the basement was very old and that all the basins of sediments were younger. These two groups of rocks were regarded as representing two divisions of time and were called the Archean and Proterozoic eras.

There were many arguments, and in 1935, C. K. Leith suggested that the two types of rocks did not represent only two divisions of time, a view which had indeed been expressed earlier by the Canadian geologists, F. D. Adams and A. C. Lawson.

In recent years, when many age determinations became available, the doubts about the old classification became stronger. It has been found that the dates of the pegmatites of the metamorphic areas (called Archean) are not uniform.

At present the Canadian Shield can be divided into at least seven different archean areas, each with a characteristic and different age of pegmatites. At the same time, the ages which we have for the sedimentary rocks of Proterozoic age (although they are harder to obtain) also show great variations.

These Proterozoic rocks, although little altered, are always younger than the metamorphosed Archean rocks upon which they rest, but they may be older than Archean rocks elsewhere. Archean and Proterozoic are types of rocks and do not represent just two eras of time. Rocks of either type may be of any Precambrian age.

The early geologists distinguished the broad belt of Precambrian rocks along the St. Lawrence River as the Laurentian or Grenville rocks, differing from rocks to the northwest of them.

Age determinations have shown that these Grenville rocks have pegmatites uniform in age of about 1×10^9 years, while the pegmatites in the Keewatin rocks are about 2.5×10^9 years and those in the White Mountains and other Appalachian primary arcs are only 0.3×10^4 years.

Along the northwest side of the Grenville or Laurentian metamorphics are a series of sedimentary basins filled with Huronian and other Proterozoic type of sedimentary rocks. Such ages as have been determined relate them to the Grenville. Of course, they are younger than the Keewatin on which they rest. Between the Grenville on one side and the Huronian and Keewatin is a great fault. The relationships are like those in the Appalachians.

The situation is thus believed to be this: The Keewatin represents very ancient roots of mountains, while the adjacent Grenville is the roots of

In that case we can suggest that, in about Cretaceous time when the Kuril and Aleutian arcs were formed, another part of the system was abandoned. This clearly lay from British Columbia through northwestern Canada across northern Alaska in the Brooks Range (which became inactive about Triassic time) and through the Paleozoic and Mesozoic mountains of northeastern Siberia to Japan.

Thus, the mountain systems can be thought to have changed position like a highway that has a detour in it or which is being shortened.

Another case of an abandoned mountain range is the Appalachians which became inactive in Permain time (last folding) or in Triassic time (last volcanism). Parts of it are now beneath the sea or are buried by the Atlantic coastal plain and that of the Gulf of Mexico which are Cretaceous and Tertiary in age. We can, however, still plainly trace the greater part of a narrow mountain system. The primary arcs, of which the roots are exposed as volcanic and igneous rocks, lay through central Newfoundland, Nova Scotia, New Brunswick, Maine, and the White Mountains of New Hampshire. They also formed the Piedmont region of the Carolinas (see Fig. 4).

The junctions of these arcs at which secondary arcs formed are in northern Newfoundland, in Gaspé of Quebec, in the Green Mountains of Vermont, in the valley and ridge and the Blue Ridge of Pennsylvania and of Tennessee, in the Ouachita and Marathon Mountains of Oklahoma and Texas.

Several of the graben are known; for example, along the Saguenay River, the Ottawa River and the Kentucky River and there are possibly two others west of the Wichita and Marathon areas.

The Appalachians are no longer active tectonically, seismically (except slightly) nor volcanically. They are being worn away.

If this process continues, we may expect to see the roots of the primary arcs and of the welts of the secondary arcs as a great belt of metamorphic rocks on the inner side of which lie some basins of little altered but folded sediments, representing the deepest part of the Valley and Ridge type of basins.

There might well be expected to be a major fault system separating the Appalachian Mountain system from the older basement of Precambrian age exposed in the Blue Ridge type of welts.

The Blue Ridge has been dated and is known to be of Precambrian, Grenville age (1.0×10^9 years). Thus, when the Appalachians are worn down, their roots can be expected to form two provinces of metamorphic rocks of different ages (Grenville and Paleozoic) which will be separated by a fault zone near which will be several basins of folded but little altered sedimentary rocks corresponding in age to the younger or Paleozoic metamorphics.

Precambrian Shields

When Precambrian shields were first studied, it was found that metamorphic rocks (which could not be easily mapped) were overlain with great unconformities by sedimentary rocks which were little altered and which were younger than the basement upon which they lay. These sedimentary rocks lay in separate basins. The assumption was made that all the basement was very old and that all the basins of sediments were younger. These two groups of rocks were regarded as representing two divisions of time and were called the Archean and Proterozoic eras.

There were many arguments, and in 1935, C. K. Leith suggested that the two types of rocks did not represent only two divisions of time, a view which had indeed been expressed earlier by the Canadian geologists, F. D. Adams and A. C. Lawson.

In recent years, when many age determinations became available, the doubts about the old classification became stronger. It has been found that the dates of the pegmatites of the metamorphic areas (called Archean) are not uniform.

At present the Canadian Shield can be divided into at least seven different archean areas, each with a characteristic and different age of pegmatites. At the same time, the ages which we have for the sedimentary rocks of Proterozoic age (although they are harder to obtain) also show great variations.

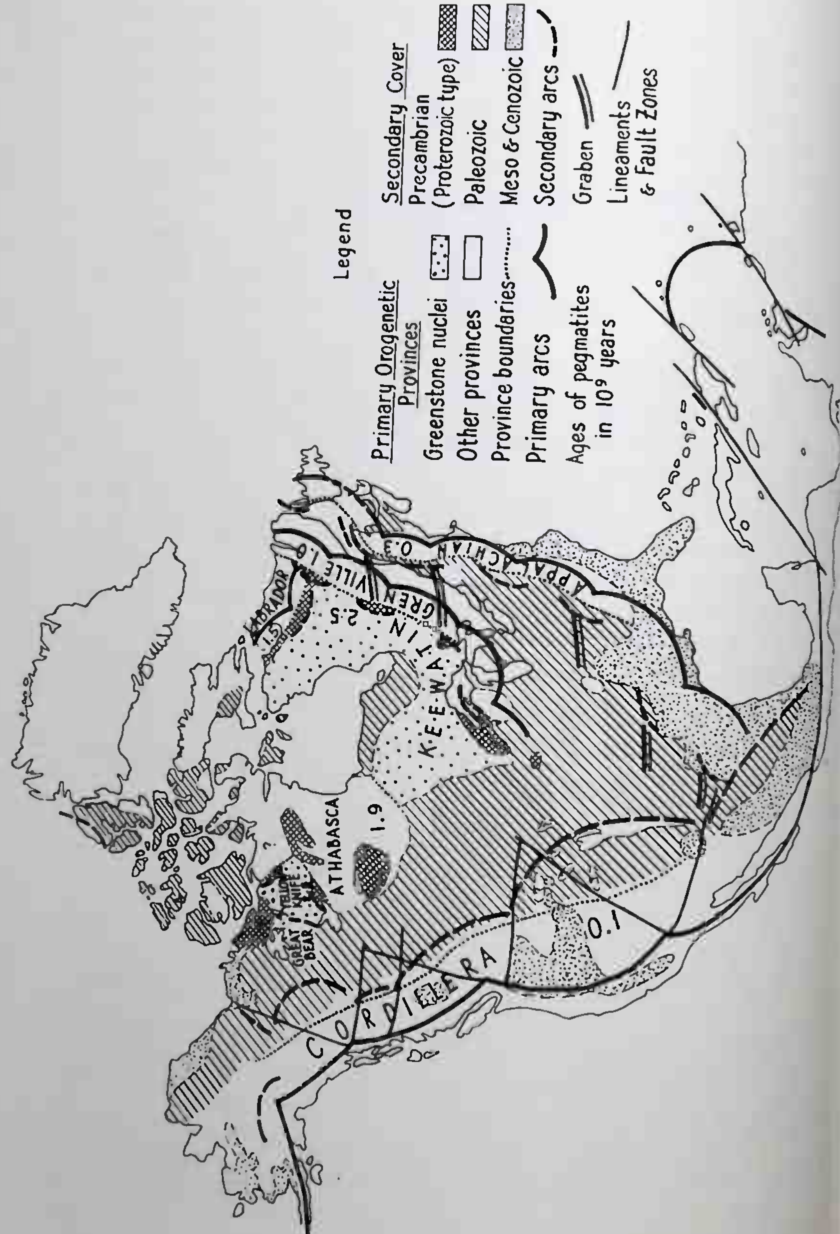
These Proterozoic rocks, although little altered, are always younger than the metamorphosed Archean rocks upon which they rest, but they may be older than Archean rocks elsewhere. Archean and Proterozoic are types of rocks and do not represent just two eras of time. Rocks of either type may be of any Precambrian age.

The early geologists distinguished the broad belt of Precambrian rocks along the St. Lawrence River as the Laurentian or Grenville rocks, differing from rocks to the northwest of them.

Age determinations have shown that these Grenville rocks have pegmatites uniform in age of about 1×10^9 years, while the pegmatites in the Keewatin rocks are about 2.5×10^9 years and those in the White Mountains and other Appalachian primary arcs are only 0.3×10^4 years.

Along the northwest side of the Grenville or Laurentian metamorphics are a series of sedimentary basins filled with Huronian and other Proterozoic type of sedimentary rocks. Such ages as have been determined relate them to the Grenville. Of course, they are younger than the Keewatin on which they rest. Between the Grenville on one side and the Huronian and Keewatin is a great fault. The relationships are like those in the Appalachians.

The situation is thus believed to be this: The Keewatin represents very ancient roots of mountains, while the adjacent Grenville is the roots of



another somewhat younger system of primary mountains to which the Huronian basins are the corresponding secondary arcs.

In similar fashion we find that all shields so far investigated are made up of provinces. The metamorphic provinces are separated by great faults and differ in age. Along the inner side lie sedimentary basins which are the roots of former secondary arcs. Each row of basins rests upon older metamorphics, but each is related (usually as secondary arcs are to primary arcs) to an adjacent province of younger metamorphics which lies on the seaward side.

Evolution of the Crust

The zoned arrangement of continents thus fits the concept that they have grown and that each zone represents the former position of part of the system of active mountains or in other words of a place where part of the continental fracture system once spewed out andesitic lavas.

It has been shown that the present rate of emission of lava is of the right order of magnitude to produce all the crust during geological time and that this would produce more contraction than any reasonable amount of cooling of the Earth. It has been shown that contraction could lead to fracturing of a kind which might show many of the principal features observed in existing and past mountains. A vast amount remains to be done, but no other theory can explain so much. Continental drift is without a cause or a physical theory. It has never been applied to any but the last part of geological time. Convection currents may exist in the mantle, but this is a pure hypothesis and it has never been shown how they form mountains with many of the properties of existing mountains. If they exist, it is possible that they are not the cause of mountain building. Polar wandering (that is, movement of the Earth as a whole about its axis) can have occurred without in any way affecting the arguments here put forward.

It is believed that contraction of the Earth due to its emission of lava and volcanic gases provides a tentative theory for the building of mountains and continents which is capable of explaining more of the details of these features than any other theory yet proposed.

The ocean basins are the original crust, only thinly covered with basalt. The mid-ocean and continental fracture systems are the breaks which allow the Earth to shrink. Conservation of energy tends to keep them as far apart as possible. The mid-ocean ridges and the continents are made of the material spewed out along these breaks during the Earth's long history. The feebler mid-ocean system has not migrated much, but the more active continental system has swung about from one side of the continents to another building them up, first in one place and then in another, so that the zoned continents are complexes of crustal scar tissue. The Earth has given birth to its own oceans, atmosphere,

islands, and continents and is continuing to generate them at its usu[al]
rate.

REFERENCES

Kuiper, G. P. (Editor). 1954. The Earth as a Planet. University of Chicago Pr[ess].
Poldervaart, Arie (Editor). 1955. Crust of the Earth. Geol. Soc. Ameri[ca]
Special Paper 62.
Scheidegger, Adrian E. 1958. Principles of Geodynamics. Springer-Verl[ag],
Berlin.
Wilson, J. T., J. A. Jacobs and R. D. Russell (in press). Physics and Geolo[gy].
McGraw-Hill Book Co., Inc., New York.

GRAVITATION—AN ENIGMA

By R. H. DICKE

I AM deeply appreciative of the opportunity to deliver this year's Joseph Henry lecture[1]. As a Princetonian I am particularly grateful, for although the Philosophical Society of Washington may claim Joseph Henry as its founder and first president, Princeton also has a claim. For it was much earlier as professor of natural philosophy at Princeton University that Henry made many of his important discoveries. He might have been the first and was certainly not the last scientist to sacrifice his opportunity for research in order to serve the United States Government as an administrator.

Is gravitation an enigma? I fear that many of my colleagues would say no! In fact, for the past 250 years there have been only two accepted theories of gravity, each of which has enjoyed widespread approval, of course, not at the same time. The approval of General Relativity is such that for the past 40 years most physicists have not felt a need for any new fundamental experiments on gravitation.

It is difficult to realize how venerable is the original Newtonian theory of gravitation. For example, in 1837, when Joseph Henry was experimenting on electromagnetic induction at Princeton, Newton's law of gravitation had already been stated 150 years previously. A very extensive literature on the dynamics of the solar system had been produced including the important works of Euler, Lagrange, and Laplace.

Prior to Newton's theory of gravitation there had been no theory in the modern sense of the word. An elaborate philosophical description, due to Descartes, attributed gravitation to complicated vortices in an ether-filled space called a plenum, but this was hardly a theory.

It is interesting to note that Newton's law of gravitation was not immediately accepted. As late as 1730, 43 years after the publication of Newton's *Principia*, Voltaire wrote, "A Frenchman who arrives in London will find philosophy like everything else very much changed there. He had left the world a *plenum* and now he finds a *vacuum*"[1]. The opposition to Newton's theory was largely philosophical. The inverse square law of interaction seemed to imply "action at a distance." The idea that two bodies could exert a force on each other at a distance and in a vacuum was highly objectionable, particularly to the French school brought up on the Cartesian philosophy.

Newton himself never regarded his theory of gravitation as more than an *ad hoc* formalism which seemed to work. In fact he wrote that,

[1] The 27th Joseph Henry Lecture of the Philosophical Society of Washington, delivered before the Society on April 18, 1958, and reprinted from the *Journal of the Washington Academy of Sciences*, *48*, No. 7, July 1958.

to suppose "that one body may act upon another at a distance through a vacuum, without the mediation of anything else . . . is to me so great an absurdity that I believe that no man, who has in philosophical matters a competent faculty for thinking, can ever fall into it"[2].

He conjectured that gravitation might be associated with a variation in density of the ether from one place to another, with a body being pushed from a region of high density to a region of low density[3].

General Relativity

Compared with the long period of difficulty experienced by Newton with his theory of gravity, Einstein's difficulties were short lived. For it is characteristic of Einstein's genius that he was able to construct a theory of gravitation which seemed to work and was esthetically satisfying. He was able to account for a small discrepancy observed in the orbit of Mercury as computed from Newton's laws. Further, his theory predicted an anomalous deflection of light by gravity, and this was later observed.

In a certain limited sense, Einstein's theory of gravitation, or the General Theory of Relativity, represent a return to the Cartesian philosophy, for Einstein's theory is a field theory, that is, the gravitational force acting on a body is assumed to have its origin in the space surrounding the body and not to be an "action at a distance." On the other hand, it must also be said on the philosophical side that General Relativity is not considered to be an ether theory of gravity, for the field is thought to represent the geometrical properties of the vacuum and not the properties of a "plenum." In fact the demise of the ether in 1905 was due in large measure to Einstein's contributions to the idea of Special Relativity and Lorentz invariance.

Einstein's theory of gravitation stems primarily from a single simple assumption. It had long been known that the gravitational acceleration of a body is quite accurately independent of its composition. Thus, all objects in a freely falling box fall with it, and an observer in this box would be unaware of the presence of the gravitational field. Conversely, an elevator accelerated in gravity-free space would seem to a passenger to be in a gravitational field. This suggests that a uniform gravitational field is completely equivalent to a uniform acceleration. This assumption is known as the "Principle of Equivalence."

From this point of view the common gravitational acceleration experienced by any small body at a given point in space is to be regarded as a property of the geometry of the four dimensional space-time continuum. Stated another way, a body projected at a given space-time point in such a way as to pass through another space-time point has followed a unique curve in the four dimensional space-time continuum between the two points and this curve is assumed to be determined by

the geometrical properties of space. Following Einstein, the obvious curves uniquely determined by the geometry of space-time are the geodesics, i.e., curves representing the "shortest" distance, more properly extremum, between the two points.

It is found that, to represent gravitation, the proper definition of distance between two space-time points in the space is such that the resulting space is Riemannian. That is, the infinitesimal distance (ds) between two neighboring space-time points is defined by

$$ds^2 = \sum_{ij} g_{ij} dx^i dx^j \qquad (1)$$

Here as usual, the indices i and j are summed over 1, 2, 3, and 4. dx^i are coordinate differentials of the four coordinates necessary to specify the space time point. The metric tensor, with elements g_{ij} which are functions of the coordinates, completely specifies the geometrical properties of both the coordinate system and the space. The trajectory of a particle between two space-time points is specified by the requirement that the distance along the trajectory between the two points be a minimum, i.e., by the variational principle

$$\delta \int ds = 0 \qquad (2)$$

where the integral is a line integral along the path. It should be emphasized that the definition of distance encompassed in equation (1) is assumed to be not completely arbitrary but to be the time or space interval which would be measured by actual material clocks or rods.

As stated above, in Einstein's gravitational theory there is no specific gravitational field quantity. The gravitational field is given wholly by the metric tensor g_{ij} which also specifies the various curvatures of space. The metric tensor is determined by a field equation in the form

$$R_{ik} - \tfrac{1}{2} g_{ik} R = \frac{8\pi G}{c^4} T_{ik} \qquad (3)$$

Here R_{ik} is a curvature tensor of the space. It is a function of the metric tensor and its first and second derivatives. R is the curvature scalar.

$$R = \sum_{ij} g^{ij} R_{ij} \qquad (4)$$

G is the gravitational constant, c the velocity of light, and T_{ik} the energy-momentum tensor of matter with the matter energy density as the 4-4 component. The field equation (3) results from a variational principle

$$0 = \delta \int (L + R)\sqrt{-g}\, d^4x \qquad (5)$$

where L is the Lagrangian density of the matter and $\sqrt{-g}\,d^4x$ is an element of volume in space-time. It should be stated that the equations of motion of matter are contained in the field equation (3) and that (2) is actually superfluous.

to suppose "that one body may act upon another at a distance through a vacuum, without the mediation of anything else . . . is to me so great an absurdity that I believe that no man, who has in philosophical matters a competent faculty for thinking, can ever fall into it"[2].

He conjectured that gravitation might be associated with a variation in density of the ether from one place to another, with a body being pushed from a region of high density to a region of low density[3].

General Relativity

Compared with the long period of difficulty experienced by Newton with his theory of gravity, Einstein's difficulties were short lived. For it is characteristic of Einstein's genius that he was able to construct a theory of gravitation which seemed to work and was esthetically satisfying. He was able to account for a small discrepancy observed in the orbit of Mercury as computed from Newton's laws. Further, his theory predicted an anomalous deflection of light by gravity, and this was later observed.

In a certain limited sense, Einstein's theory of gravitation, or the General Theory of Relativity, represent a return to the Cartesian philosophy, for Einstein's theory is a field theory, that is, the gravitational force acting on a body is assumed to have its origin in the space surrounding the body and not to be an "action at a distance." On the other hand, it must also be said on the philosophical side that General Relativity is not considered to be an ether theory of gravity, for the field is thought to represent the geometrical properties of the vacuum and not the properties of a "plenum." In fact the demise of the ether in 1905 was due in large measure to Einstein's contributions to the idea of Special Relativity and Lorentz invariance.

Einstein's theory of gravitation stems primarily from a single simple assumption. It had long been known that the gravitational acceleration of a body is quite accurately independent of its composition. Thus, all objects in a freely falling box fall with it, and an observer in this box would be unaware of the presence of the gravitational field. Conversely, an elevator accelerated in gravity-free space would seem to a passenger to be in a gravitational field. This suggests that a uniform gravitational field is completely equivalent to a uniform acceleration. This assumption is known as the "Principle of Equivalence."

From this point of view the common gravitational acceleration experienced by any small body at a given point in space is to be regarded as a property of the geometry of the four dimensional space-time continuum. Stated another way, a body projected at a given space-time point in such a way as to pass through another space-time point has followed a unique curve in the four dimensional space-time continuum between the two points and this curve is assumed to be determined by

the geometrical properties of space. Following Einstein, the obvious curves uniquely determined by the geometry of space-time are the geodesics, i.e., curves representing the "shortest" distance, more properly extremum, between the two points.

It is found that, to represent gravitation, the proper definition of distance between two space-time points in the space is such that the resulting space is Riemannian. That is, the infinitesimal distance (ds) between two neighboring space-time points is defined by

$$ds^2 = \sum_{ij} g_{ij} dx^i dx^j \quad (1)$$

Here as usual, the indices i and j are summed over 1, 2, 3, and 4. dx^i are coordinate differentials of the four coordinates necessary to specify the space time point. The metric tensor, with elements g_{ij} which are functions of the coordinates, completely specifies the geometrical properties of both the coordinate system and the space. The trajectory of a particle between two space-time points is specified by the requirement that the distance along the trajectory between the two points be a minimum, i.e., by the variational principle

$$\delta \int ds = 0 \quad (2)$$

where the integral is a line integral along the path. It should be emphasized that the definition of distance encompassed in equation (1) is assumed to be not completely arbitrary but to be the time or space interval which would be measured by actual material clocks or rods.

As stated above, in Einstein's gravitational theory there is no specific gravitational field quantity. The gravitational field is given wholly by the metric tensor g_{ij} which also specifies the various curvatures of space. The metric tensor is determined by a field equation in the form

$$R_{ik} - \tfrac{1}{2} g_{ik} R = \frac{8\pi G}{c^4} T_{ik} \quad (3)$$

Here R_{ik} is a curvature tensor of the space. It is a function of the metric tensor and its first and second derivatives. R is the curvature scalar.

$$R = \sum_{ij} g^{ij} R_{ij} \quad (4)$$

G is the gravitational constant, c the velocity of light, and T_{ik} the energy-momentum tensor of matter with the matter energy density as the 4-4 component. The field equation (3) results from a variational principle

$$0 = \delta \int (L + R)\sqrt{-g}\, d^4x \quad (5)$$

where L is the Lagrangian density of the matter and $\sqrt{-g}\,d^4x$ is an element of volume in space-time. It should be stated that the equations of motion of matter are contained in the field equation (3) and that (2) is actually superfluous.

Equation (3) has been solved for the case of a spherically symmetric mass distribution (Schwarzschild solution). The resulting metric tensor when inserted into equation (2) gives an equation of motion for the planet Mercury about the Sun which is in satisfactory agreement with the observations and accounts in a satisfactory way for the anomalous rotation of the periheleon of the orbit. It also predicts a deflection of very rapidly moving particles (e.g., photons) by the sun by an amount just twice what would have been expected from Newtonian mechanics. The gravitational red shift is also predicted. The first two effects seem to be in satisfactory agreement with observations and the third is qualitatively satisfactory.

We have here presented only enough of the framework of general relativity to form a basis for discussion, and we now return to the question previously raised. Is gravitation an enigma?

Gravitation as Geometry

As a first point we notice an enigmatic dichotomy. Gravitation is described in a completely different way from other interactions. Gravitation is ascribed to the geometrical properties of empty space, but all other interactions are traced to local interactions between particles.

This is a peculiar situation with an interesting history which requires elaboration. We have seen how the difficulties encountered by Newton toward the end of the seventeenth century in obtaining acceptance of his law of gravitation were due to the conflict between the concept of "action at a distance" and the concept of local action of an ether on matter.

Interestingly enough, after the success of Newtonian mechanics in dealing with planetary motion became known, the pendulum had swung the other way. Action at a distance became the accepted concept, and the ether-filled plenum was not required again until the beginning of the nineteenth century when Young's important experiments demanded a light wave for their explanation. To the physicist of this period, a wave without a medium to propagate it was unthinkable.

The idea of force fields in the form of local interactions of a tenuous medium (the ether) on matter was fundamental all through the nineteenth century. In fact, after Maxwell's electromagnetic theory, it became clear that light and all electromagnetic phenomena could be described in terms of waves in such an ether.

With Einstein's famous paper on special relativity in 1905, the ether concept was again in disrepute and for all practical purposes has been dead for the last 50 years. Its demise came about in a strange way. A number of electromagnetic experiments toward the end of the nineteenth century had been unsuccessful in measuring the velocity of the earth relative to the ether. It was shown by Lorentz and others that a contraction of matter in the direction of motion and a slowing down of the rate of clocks in motion, both by a factor

$$[1 - (v/c)^2]^{1/2} \tag{6}$$

where v is the velocity of the earth relative to the ether, would explain these results. Also, it was later shown by Lorentz that these contractions would in fact result from his ether theory of the electron with a certain reasonable assumption as to the electron's structure. He furthermore showed that the mass of an electron in motion should increase by a factor of the reciprocal of (6).

In 1905 Einstein suggested that since it appeared to be experimentally impossible to measure a velocity relative to the ether, all inertial frames were equivalent and an inertial frame with the ether at rest was physically meaningless. Once again serious doubt was expressed concerning the existence of an ether. He argued that if the concept of velocity relative to an ether is physically meaningless, the ether must not exist, otherwise this motion would be physically meaningful. We shall return later to this question.

By 1916 the vacuum had become complicated. The curvature of space represented the gravitational field. The vacuum also carried the electromagnetic fields. One suspects that, with empty space having so many properties, all that had been accomplished in destroying the ether was a semantic trick. The ether had been renamed the vacuum. This was a useful trick, however, as the mechanical connotation of the old name had been left behind.

After the discovery of the photon and the subsequent development of quantum mechanics in 1927, it was realized that what was formerly regarded as an abstract electromagnetic field in the vacuum could be visualized as a swarm of particles, the photons, being continuously created and destroyed. The electromagnetic forces could be traced to purely local interactions of the photons with charged particles. Thus, what had been thought to be a vacuum was not a vacuum but contained photons. Furthermore, it was then realized that there were other forces (e.g., nuclear) and these also could be attributed to purely local interactions between particles. Only gravitation remained as a field property of empty space.

In recent years space has been getting even more complex, with a charged particle being surrounded by a halo of both photons and electron-positron pairs in virtual states. These particles do not materialize as proper well-behaved specimens but have only a transient existence, being continuously replaced by new ones.

We thus see that all ordinary forces (e.g., electromagnetic and nuclear) are, from a physical point of view, believed to be due to purely local interactions of particles with each other. For example, the strong attraction between nucleons in the nucleus is believed due to collisions with virtual mesons (pions) which are being continually created and annihilated. Only gravitation and inertia are singled out as forces associated with empty space. This is an anomaly to which we shall return.

Gravitation a Weak Interaction

The second way in which gravitation is anomalous is in its strength. It is the weakest of all known interactions.

It might seem strange to call the force of 2×10^{18} tons, required to keep the earth in its orbit, a weak force. However, stated on a per particle basis, and compared with other forces, it is extremely weak. For example, in the hydrogen atom the gravitational force between the electron and the proton is only 5×10^{-40} that of the electrostatic interaction.

The fact that the strength of the gravitational interaction is so different from other interactions might tend to strengthen the belief that it is completely different from other forces.

There are some reasons for believing that this is true. For over 35 years, Einstein and many others attempted to extend the geometrical notions of General Relativity to obtain a theory of the structure of particles, and hence of other interactions, but without notable success. These attempts were called unified field theories. It is said that one well-known physicist, in referring to the unified field theories, commented, "Let no man join together what God hath put asunder."

On the other hand, there are empirical reasons for believing that the strength of the gravitational interaction is related to both the scale of the universe and the strength of the strong atomic interactions. By a strong interaction we shall mean an electromagnetic or nuclear interaction.

The empirical relations in question were first discovered by A. S. Eddington and are known as the Eddington numbers [4]. As the first of these dimensionless numbers we could take the ratio of the electrical to gravitational force between an electron and a proton. As was stated previously this is

$$\frac{e^2}{Gm_em_p} = 2 \times 10^{39} \tag{7}$$

In similar fashion the age of the universe ($T = 12 \times 10^9$ years) expressed in units of atomic time is

$$\frac{Tm_ec^3}{e^2} = 4 \times 10^{40} \tag{8}$$

Another significant number is the square root of the number of heavy particles in the visible part of the universe. With the assumption of an average density of matter in space of $\rho = 10^{-30}$ gms cm^{-3} this is roughly

$$\left[\frac{(4\pi/3)R^3\rho}{m_p}\right]^{1/2} = 1.7 \times 10^{39} \tag{9}$$

Here $R = Tc$ is the "Hubble radius" of the universe. The approximate equality of these three numbers would seem to imply some sort of a connection between gravitational and atomic quantities.

It should be noted that the first number represents a connection between quantities usually called physical or atomic constants whereas the other two involve astrophysical or cosmological quantities expressed in atomic units.

An interesting derived quantity is obtained by squaring (9) and dividing by (7) and (8). This quantity which can be called the ratio of gravitational energy to matter energy in the universe is

$$\frac{(4\pi/3)R^2\rho G}{c^2} = 0.036 \tag{10}$$

It should be emphasized that these numbers are inaccurate. In particular the accuracy with which the density of matter in the universe is known is very poor.

What interpretation can be made of the Eddington numbers? That these relations should be accidental seems unlikely, but what is their significance?

Dirac's Cosmology

A number of years ago P. A. M. Dirac [5] made an interesting suggestion regarding the Eddington numbers. He noted that one of the numbers (8) is proportional to the age of the universe. With the assumption that the rough equality of (7), (8), and (9) is independent of time, he concluded that, compared with electrical interactions, the gravitational interactions are steadily getting weaker. More exactly, the gravitational constant G varies inversely as the age of the universe. He also concluded from the equality of (7) and (9) that the number of particles in the visible part of the universe varies as the square of the age of the universe. This latter conclusion does not necessarily imply the creation of new particles as the boundary of the visible part of the universe may move outward to encompass more particles in the interior.

Dirac's hypothesis is very interesting as it suggests that the gravitational interaction between two particles is not a purely local phenomenon but depends upon distant matter. As the amount of matter in the visible universe becomes greater, the gravitational interaction becomes weaker.

If Dirac's hypothesis is correct, there are many important implications for geologists and astrophysicists. Some of these have been previously discussed [6]. We have merely summarized very briefly some of the important effects.

(*a*) The radiation rate of the sun varies as G^7. The sun must then have been hotter in the past.

(*b*) With a constant albedo (i.e., visual surface reflection coefficient) the surface temperature of the earth would vary with the age of the universe as the -2.25 power of the age.

(*c*) The albedo of the earth would actually change with the water vapor content of the atmosphere. This would tend to stabilize the earth's temperature for the past few billion years.

(*d*) For an age of the universe of 12×10^9 years, and an age of the solar system of 4×10^9 years, the increased radiation rate of the sun in the past would have required a conversion of roughly 20% of the hydrogen into helium. The rapid burning in the past would have been obtained through the use of the carbon cycle rather than the *p-p* reaction now the dominant one. This would lead to a convective core in the sun serving to mix the helium throughout the core. It is possible that a thorough mixing of helium throughout the whole sun might have occurred at some time in the past.

(*e*) In addition to the effects associated with an original high surface temperature, there are many other important geological effects. For example, the earth gradually expands with time. This could be an explanation for the apparent relation between the coast lines of the land masses bordering the Atlantic Ocean and the mid-Atlantic ridge.

(*f*) The gradual decrease of gravitation would lead to a continuous slowing of the rate of revolution of the planets and the moon compared with the rotation rate of the earth. This seems to be observed, although there are several other possible explanations, including an increased rotation rate of the earth produced by atmospheric tides, melting ice masses etc.

(*g*) The gradual reduction of pressure inside the earth leads to a continual slow convection of the eartb's mantle carrying heat to the surface. The computed rate of heat transfer is about 5×10^{-7} cal/cm^2 sec and this is compatible with the observed rate of heat flow after making allowance for heat from radioactivity. There appears to be some geological evidence that such a convection does actually exist in the earth's mantle.

By comparing these conclusions, and many others, with the observed history of the solar system, it may be possible to exclude Dirac's hypothesis. It would be truly remarkable if geology and astronomy could answer such a fundamental physical question about the gravitational interaction.

Although there are a number of reasons for accepting Dirac's hypothesis, there is also some evidence to the contrary. For example, there is evidence for glacial deposits over 1 billion years old. It is difficult to understand such deposits if the sun were substantially hotter at these times.

It is interesting to note, therefore, that Dirac's argument has a logical loophole. To infer the time dependence of the gravitational interaction requires more than a simple observation that the reciprocal of the gravitational constant and the age of the universe, when expressed dimensionlessly, are *now* nearly equal. It is also necessary to assume that *now* is a random time. But is it?

The present epoch is conditioned by the fact that the biological conditions for the existence of man must be satisfied. This requires the existence of a planetary system and a hot star. If we assume an evolutionary cosmology starting with the formation of hydrogen 12 billion years ago, there is an upper limit for the epoch of man which is imposed by the following two conditions: First, hydrogen is being continually converted to helium and heavier elements. Perhaps 20% has already been "burned." The formation of new stars is thus limited to perhaps the first 10^{11} years. Second, there is an upper limit on the radiating life of a star.

If the star is massive (10 times the sun's mass) it lives riotously, burning its hydrogen like a wastrel. For a light star ($\frac{1}{10}$ the sun's mass), hydrogen is burned slowly and the star is capable of living much longer than the sun, 100 times as long. However, if the star is much smaller than this, its central temperature never rises high enough to cause nuclear reactions to take place. Such a light star radiates until its gravitational energy is gone and then it cools off. It is seen therefore that the longest life of a star is very roughly 10^{14} years and this puts an upper limit on the epoch of man.

There is also a lower limit on the epoch of man. With the assumption that initially only hydrogen exists, it is necessary to produce other elements in the stellar caldrons and distribute them about the universe before a planetary system of our type can be formed. It is a bit difficult to estimate this time, but it would seem that 1 billion years would be a reasonable lower bound on the epoch of man.

It is thus seen that the epoch of man is not random but is very roughly delineated. Thus, it is possible that the strength of the gravitational interaction is not determined by the immense scale of the universe, but that conversely the scale of the universe and the present epoch is determined by the relative strengths of the strong and weak interactions.

From this point of view the Eddington number (7) is fundamental, (8) is determined by biological considerations and only (9) requires an explanation. Presumably, the significance of (9) is lost in the mystery of the original creation of hydrogen, for this number is determined by the density of matter. We shall return later to a possible explanation of this number.

One unsatisfactory feature of the above explanation of the numerical relations (7), (8), and (9) is that it does not give a satisfactory account of the enormous size of these numbers. It would be hoped that eventually an atomic constant such as the dimensionless number (7) would result from some theory of particles. But how could a number such as 10^{40} be expected to appear in a theory for which all other dimensionless numbers are small?

Here, Dirac's cosmology is more satisfactory. The large size of the Eddington numbers is considered to be purely accidental since the epoch *now* is taken to be random.

It is possible to construct cosmologies which lie somewhere between these two extremes. It could be assumed that the epoch *now* is not random but determined by the strengths of the weak interactions in the manner previously described. On the other hand, the strength of the weak interactions might be determined by the structure of the universe. For example, the number of particles in the universe might be large as a result of some chain reaction and this large number might determine the strengths of the weak interactions in some way not presently understood.

It is evident that all these cosmological models are highly speculative. However, one feature they have in common is not easily avoided. They all imply some sort of interrelation between the strong and gravitational interactions.

Absolute Space, Mach's Principle and Gravitation

It will be recalled that it was largely Einstein's influence which led to the rejection of an absolute space as an allowable physical concept. However, Einstein's theory of gravitation does make use of some of the concepts of an absolute space. For example, inertial coordinate systems are assumed to have a well-defined meaning even in the absence of gravitating matter.

Also, it appears to be a basic assumption of Einstein's theory that such geometrical concepts as points, curvatures, and geodesics are meaningful for a vacuum. The conceptual difficulty of defining these concepts and of describing motion relative to such an empty physical space is ameliorated by imagining space filled with a set of massless test particles. By moving on geodesics of the space, these particles serve to make evident the underlying geometrical structure of space. This underlying geometrical structure is thus regarded as a property of an absolute space. While it is true that the geometry of space is affected by the presence of gravitating matter, General Relativity is usually thought to give meaningful results even in the absence of matter.

The conceptual difficulties with ascribing physical properties to a vacuum have been long apparent. That there should be preferred coordinate systems (the inertial systems) in empty space is difficult to comprehend. This is a very old problem, and still without a solution. Many years ago, Mach [7] made an interesting suggestion which has come to be known as Mach's Principle. According to this principle it is not acceleration relative to empty space that produces an inertial force. This force is considered to be due to the acceleration relative to distant matter in the space. From the point of view of a strict relativist, the effect can also be said to be due to an acceleration of the distant matter relative to the stationary particle experiencing the inertial force.

The inertial force is not considered to be a new and different type of force but simply the gravitational interaction with distant matter. A particle at rest with respect to distant matter finds itself surrounded by a spherically symmetrical universe, and all gravitational forces cancel. If distant matter is accelerated, there is an unbalanced gravitational force acting on the particle in question and this force is called the inertial force [8].

For example, the earth accelerating toward the sun can be considered at rest in a particular coordinate system for which the total gravitational force on the earth is zero. This force is composed of the gravitational pull of the sun on the earth and the gravitational interaction of the remainder of the universe [8].

From the point of view of Mach's Principle, the inertial mass of a particle is determined by distant matter. From the point of view of Einstein's theory the inertial mass is a purely local property of the particle.

Interestingly enough, the actual structure of the universe appears to be compatible with Mach's Principle, at least to the extent that the structure is known. This is easily seen by considering an old problem solved by Thirring [9]. Thirring used the weak field solution to Einstein's field equations to find the rotation of the inertial coordinate systems inside a massive rotating spherical shell. He found that the inertial coordinate axes rotated with an angular velocity ω relative to distant inertial axes where

$$\omega = \omega_0 \frac{8Gm}{3aC^2} \qquad (11)$$

Here ω_0 is the angular velocity of the rotating shell, m is its mass, and a its radius.

In similar manner, from the point of view of Mach, the Coriolis force experienced in a rotating laboratory is due to the gravitational field produced by distant matter wheeling around in a set of concentric shells. In this case, the rotation rate of the inertial coordinate system ω is equal to the rotation rate of the distant matter ω_0 and Eq. (11) becomes

$$1 = \frac{32\pi}{3} \frac{G}{c^2} \int_0^R [\rho] r \, dr \qquad (12)$$

Here the distant matter is treated as a set of concentric shells out to the Hubble radius R. $[\rho]$ is an effective density of matter and is equal to the observed density only for $r \ll R$, for the weak field solution is valid only under this condition. It would be expected that equation (12) could be written approximately as

$$\frac{GM}{Rc^2} \sim 1 \tag{13}$$

where M is the mass in the visible part of the universe.

The relations, equations (12) and (13) appear to hold for the actual universe to within the accuracy of the observation. See equation (10).

It must be emphasized that General Relativity itself does not require a condition of the form of equation (12) but that such a condition is compatible with General Relativity. Hence, it is conceivable that some solutions of Einstein's field equations are compatible with Mach's Principle and that Mach's Principle determines which are physically meaningful solutions.

From this point of view it would be hoped that Mach's Principle should appear as a boundary condition upon solutions to Einstein's field equations, the allowed solutions being only ones which are compatible with Mach's Principle. Attempts of Komar and others to introduce such conditions into Einstein's theory have not yet been successful.

It is also possible that the absolute inertial properties of space are an essential part of Einstein's theory of gravitation and that the theory is incompatible with Mach's Principle. Indications exist that this is true; Einstein's theory may not be compatible with Mach's Principle. Consider, for example, the space inside a hollow spherical mass shell. According to the interior Schwarzschild solution of Einstein's equation, the space is flat and has inertial properties independent of either the mass or radius of the sphere or of the existence of other exterior concentric mass spheres. Thus, according to Einstein's theory, the presence of this matter has no locally observable effect on the interior.

However, consider again equation (11); it can be combined with equation (13) to give

$$\frac{\omega}{\omega_0} = \gamma \frac{m}{M} \frac{R}{a} \tag{13a}$$

with γ a number of the order of unity. This expresses the ideas of Mach very neatly. The ratio of the two precession rates is in the ratio of the masses and inversely in the ratio of the radii of the mass spheres. On the other hand, the ratio M/R, as seen locally, is easily changed. For example, the rotating sphere could be surrounded by a large fixed spherical shell, in addition to distant galactic matter, effectively changing the integral in equation (12) and hence the effective ratio M/R. According to equation (13a) such a change would result in a change in the ratio ω/ω_0, a locally observable effect. From equation (11) this implies a change in the quantity

$$\frac{Gm}{ac^2}$$

This change is to be traced to a change in the inertial forces. On the other hand, with local units defined in terms of atomic units, m and a are unchanged by definition. Also c could not change or there would be large observable violations of the Principle of Equivalence [10]. The only possibility is a change in G as it is observed locally. With this interpretation, equation (12) gives the locally observed value of G^{-1} as an integral over all the observed matter assumed to be isotopically distributed. With this interpretation, however, the numerical constant of equation (12) must also be questioned as it is based on General Relativity which is now invalid under this interpretation.

It should be noted that, from equation (12), the presence of nearby matter affects the local gravitational constant by an appreciable amount. Although the sun is hardly distributed isotropically about the earth, its presence should change G by an amount of the order of

$$\frac{\Delta G}{G} = -\frac{8}{3}\frac{GM}{rc^2} = -2 \times 10^{-8} \tag{14}$$

Another possible effect of distant matter on inertial mass would appear as a velocity dependence of the gravitational constant. The uniform motion of a laboratory relative to distant matter might result in a change in the gravitational constant observed in the laboratory. Any such change should be proportional to the square of the velocity relative to distant matter.

Mach's Principle provides a possible explanation for the Eddington numbers, equations (7), (8), and (9). We note first that equation (10), after making allowance for the inaccuracies in the numbers, is equivalent to equation (12). Secondly, the gravitational constant and the size of (7) is conditioned by the size of the universe, through equation (12). The number (8) is determined by biological conditions in the manner previously described. The number (10) is determined by (7), (8), and (11). The enormous size of (7) is related to the large number of particles in the universe and the reason for this will not be understood until a mechanism for the creation of matter is found.

Dirac's hypothesis that G varies with the age of the universe is understandable from the point of view of Mach's Principle. In fact, equation (12) implies that G is constant only if M/R is constant as the universe evolves. On the other hand, the structure of the universe may be such that the ratio M/R does not change with time. Thus, with the assumption of this form of Mach's Principle, G can be either varying or constant as a function of the time.

Mach's Principle and the Ether

As we have seen, only gravitation and inertia are considered to be field properties of the vacuum. All other forces are assumed to be caused by local interactions between particles, real or virtual. To ascribe such physical properties to the vacuum has long been considered objectionable.

It has been seen from the strength of the gravitational interaction that there are reasons for believing that gravitation is not completely different from other interactions. This suggests that gravitation be treated like other forces. It might be imagined that a gravitating mass is surrounded by a swarm of virtual particles and that the gravitational interaction involves only local interactions with these particles. It would be expected that a quantization of the gravitational field might result in such a particle picture of gravitation.

This picture of gravitation has several advantages. The dichotomy between gravitation and ordinary forces has disappeared, gravitation is *not* a property of the vacuum, local interactions between particles can be treated in accordance with the general relativity principle, and an action at a distance principle is not needed. The general relativity principle states that the only physically meaningful motions are of matter relative to other matter. Mach's Principle follows from the general relativity principle.

Inertia effects would be treated in the manner described previously as a gravitational interaction with matter at great distance. Thus the contribution to the density of the virtual particle field falls off as r^{-1} and distant matter is more important than nearby matter in determining the inertial properties of space.

Such a picture of space provides the extra degree of freedom needed for a variable gravitational constant (G). Two regions of space, both free of gravitational fields, may have different densities of virtual particles and different gravitational constants.

It will be noted that the picture of gravitation which is being drawn is very similar to the ether theories long rejected. In this connection it is interesting to note that the old objection of Einstein and others to the ether concept no longer holds. An ether consisting of a swarm of virtual particles, all in motion, may present the same appearance to two observers moving relative to each other. If it be assumed that the velocity distribution of the virtual particles is Lorentz invariant, the swarm of virtual particles (the ether or plenum) would look the same to all uniformly moving observers.

There is one difficulty with such an assumption. Such a distribution implies particles with arbitrarily high momentum states. If such momenta are excluded, Lorentz invariance would not be exact. Anomalies would show up at very high energies. From this point of view Lorentz invariance is not the simple fundamental invariance that is presently believed. It does not have the fundamental significance of the general relativity principle.

Models and Formalism

One can hardly drag the "ether," kicking and screaming, back into the laboratory for serious consideration without saying a bit about the role and significance of such a concept as the "ether" for modern physics. Since the time of Newton, there have been two diametrically opposed techniques for inductive reasoning in physics. Each has contributed a great deal to the development of physics.

Newton was probably the first exponent of the formal approach. In the *Principia* Newton states: "I have not been able to discover the cause of those properties of gravity from phenomena and I frame no hypothesis!"

In the formal approach to the physical world, only the physical observables are regarded as of physical significance, and the role of theory is one of giving a purely formal description of the relations between observables. The great advantage of this technique is that a problem is simplified to its bare essentials. A complicated substratum, existing perhaps only in the human mind, is ignored.

Thermodynamics is a very good example of a formal theory. The observables are such quantities as temperature, density, heat, energy, etc. The complicated substratum of moving molecules is ignored in constructing the theory.

A good example of the model approach to inductive reasoning is furnished by the molecular model of a gas and the resulting kinetic theory and statistical mechanics. This model is very old and was used by D. Bernoulli in 1738 to formulate a primitive kinetic theory of gases. The model was ignored for 100 years thereafter until the middle of the nineteenth century when it was again used to explain a variety of things. It should be remembered that practically the whole of the presently accepted kinetic theory of gases and statistical mechanics was originally based on a model which was not at the time subject to a direct experimental check.

For the past 30 years, formal theories have dominated physics, and this has been coupled with an insistence on an operational definition of the observable quantities. This is in marked contrast with the nineteenth century when the model approach was the dominant one. If the present atmosphere of a completely formal approach had been dominant during the nineteenth century, it is very doubtful that the kinetic theory of gases would have appeared until it had been forced by direct experimental observation of the molecules.

What then is the significance of a model such as the "ether"? This has a simple answer. To the extent that the model is beyond observation, it is simply a figment of the imagination that has no place in physics. However, to the extent that the model represents phenomena which by local interactions between particles, real or virtual. To ascribe such physical properties to the vacuum has long been considered objectionable.

It has been seen from the strength of the gravitational interaction that there are reasons for believing that gravitation is not completely different from other interactions. This suggests that gravitation be treated like other forces. It might be imagined that a gravitating mass is surrounded by a swarm of virtual particles and that the gravitational interaction involves only local interactions with these particles. It would be expected that a quantization of the gravitational field might result in such a particle picture of gravitation.

This picture of gravitation has several advantages. The dichotomy between gravitation and ordinary forces has disappeared, gravitation is *not* a property of the vacuum, local interactions between particles can be treated in accordance with the general relativity principle, and an action at a distance principle is not needed. The general relativity principle states that the only physically meaningful motions are of matter relative to other matter. Mach's Principle follows from the general relativity principle.

Inertia effects would be treated in the manner described previously as a gravitational interaction with matter at great distance. Thus the contribution to the density of the virtual particle field falls off as r^{-1} and distant matter is more important than nearby matter in determining the inertial properties of space.

Such a picture of space provides the extra degree of freedom needed for a variable gravitational constant (G). Two regions of space, both free of gravitational fields, may have different densities of virtual particles and different gravitational constants.

It will be noted that the picture of gravitation which is being drawn is very similar to the ether theories long rejected. In this connection, it is interesting to note that the old objection of Einstein and others to the ether concept no longer holds. An ether consisting of a swarm of virtual particles, all in motion, may present the same appearance to two observers moving relative to each other. If it be assumed that the velocity distribution of the virtual particles is Lorentz invariant, the swarm of virtual particles (the ether or plenum) would look the same to all uniformly moving observers.

There is one difficulty with such an assumption. Such a distribution implies particles with arbitrarily high momentum states. If such momenta are excluded, Lorentz invariance would not be exact. Anomalies would show up at very high energies. From this point of view Lorentz invariance is not the simple fundamental invariance that is presently believed. It does not have the fundamental significance of the general relativity principle.

Models and Formalism

One can hardly drag the "ether," kicking and screaming, back into the laboratory for serious consideration without saying a bit about the role and significance of such a concept as the "ether" for modern physics. Since the time of Newton, there have been two diametrically opposed techniques for inductive reasoning in physics. Each has contributed a great deal to the development of physics.

Newton was probably the first exponent of the formal approach. In the *Principia* Newton states: "I have not been able to discover the cause of those properties of gravity from phenomena and I frame no hypothesis!"

In the formal approach to the physical world, only the physical observables are regarded as of physical significance, and the role of theory is one of giving a purely formal description of the relations between observables. The great advantage of this technique is that a problem is simplified to its bare essentials. A complicated substratum, existing perhaps only in the human mind, is ignored.

Thermodynamics is a very good example of a formal theory. The observables are such quantities as temperature, density, heat, energy, etc. The complicated substratum of moving molecules is ignored in constructing the theory.

A good example of the model approach to inductive reasoning is furnished by the molecular model of a gas and the resulting kinetic theory and statistical mechanics. This model is very old and was used by D. Bernoulli in 1738 to formulate a primitive kinetic theory of gases. The model was ignored for 100 years thereafter until the middle of the nineteenth century when it was again used to explain a variety of things. It should be remembered that practically the whole of the presently accepted kinetic theory of gases and statistical mechanics was originally based on a model which was not at the time subject to a direct experimental check.

For the past 30 years, formal theories have dominated physics, and this has been coupled with an insistence on an operational definition of the observable quantities. This is in marked contrast with the nineteenth century when the model approach was the dominant one. If the present atmosphere of a completely formal approach had been dominant during the nineteenth century, it is very doubtful that the kinetic theory of gases would have appeared until it had been forced by direct experimental observation of the molecules.

What then is the significance of a model such as the "ether"? This has a simple answer. To the extent that the model is beyond observation, it is simply a figment of the imagination that has no place in physics. However, to the extent that the model represents phenomena which are observable or may be some day observable, and to the extent that the model suggests new experiments and new theoretical developments, it is useful. One can go too far in requiring a strict and immediate observable significance for all physical concepts employed. What is important is the realization that the end results should be subject to observation.

Summary

We started with Descartes' gravity, a maelstrom in a plenum, and ended with the speculation that both gravitational and inertial forces may be due to a bath of particles in rapid motion interacting with matter. We considered the mysterious Eddington numbers, Dirac's hypothesis of a time varying gravitational constant, and the implication of Mach's Principle. The remarkable fact was considered that the history of the solar system seen through the eyes of astronomers and geologists may enable fundamental conclusions to be drawn regarding physical interactions in the past.

The chief conclusion that can be distilled from the above discussion is that it is a serious lack of observational data that keeps one from drawing a clear portrait of gravitation. Each tiny fragment of information appears as a star shining through a murky haze. Conclusions regarding the most fundamental of physical concepts are based on numbers which may be off by a factor of 100. In any case, it appears clear that there is little reason for complacency regarding gravitation. It may well be the most fundamental and least understood of the interactions.

REFERENCES

1. WHITTAKER. History of the Theories of the Ether, Vol. 1, 29, London, 1951.
2. NEWTON. Principia.
3. WHITTAKER. Loc. cit., 28.
4. EDDINGTON. The Philosophy of Physical Science, Cambridge, 1939.
5. DIRAC. *Proc. Roy. Soc.*, *A165*, 199, 1938.
6. For a summary see R. H. DICKE, *Rev. Mod. Phys.*, *29*, 355, 1957.
7. E. MACH. The Science of Mechanics, ed. 6, 1904.
8. D. SCIAMA. M. N. *113*, 34, 1953.
9. *Phys. Zeit.*, *19*, 33, 1918; *22*, 29, 1921.
10. R. H. DICKE. *Rev. Mod. Phys.*, *29*, 363, 1957.

ALBERT EINSTEIN AND THE ROLE OF THEORY IN CONTEMPORARY PHYSICS

By CORNELIUS LANCZOS

IN OUR era of great activities the preponderance of scientific discoveries is noticeable. The advancements in physics and engineering, the spectacular explorations of nuclear phenomena, the sulfa drugs, antibiotics and vitamins in biology, the astonishing discoveries of psycho-analysis, all indicate that we live in an epoch of scientific development which has hardly a parallel in the entire history of mankind. Now science is partly experimental, partly theoretical. Sometimes the one, sometimes the other phase is in the focus of interest. What can we say about the role of *theory* in the contemporary evolution of physics? Is it something of the ordinary or the extraordinary kind? Have we advanced in the last 50 years merely at an accelerated pace, or has something basically new happened?

In order to analyze this question we will take our bearings from another historical period which in its scientific evolution shows many remarkable analogies to the present stage of development. It is the era of awakening intellect, the decline of the Middle Ages and the dawn of exact scientific thinking, within the framework of Western civilization. It is the great epoch of learning which starts with Copernicus and leads through Kepler and Newton to the scientific philosophy of the eighteenth century.

Here is Copernicus with his great opus in which he demonstrates the superiority of the heliocentric view compared with the geocentric view. The earth rotates around its axis and revolves around the sun. The earth drops to a second rate planet compared with the mighty sun. And then, only two generations later, Galileo demonstrates with his telescope the miracles of heaven, the mountains of the moon, the moons of Jupiter, the phases of Venus, the sunspots—awe-inspiring cosmic spectacles compared with which our own spectacle of the Atom bomb with its devastating consequences shrinks to insignificance. The intellect opens its eyes and, lo and behold!, it discovers a new world whose boundaries seem to recede to infinity.

We are wont to consider Galileo as the father of the modern scientific method by his insistence on the experiment. Against the old Aristotelian approach of trying to comprehend nature on the basis of logical reasoning we should *observe* nature and find its laws by *experimentation* rather than speculation. But is this view actually justified? In the first place, Aristotle was a very keen observer, as his famous biological studies on the evolution of the chicken egg demonstrate. Had he had the refined instruments of Galileo at his disposal, would he not have made the same discoveries? Furthermore, there is a peculiar paradox about the experimental method: "You should not trust your reason but your observations!" But look at Copernicus, who taught the motion of the earth. The reason that people did not believe him was not the religious motive that the Scriptures teach us that the earth is at rest and the heavens move. By no means. Our *observations* disprove him. Here is a man who is crazy enough to tell us that the earth moves. But you look out of the window and see that the earth does *not* move. Then why does he tell us that it moves? The experimental, observational method is here definitely not on the right side.

Galileo was keenly aware of the difficulty. He expressed his undying admiration for those minds who were able to shake off the shackles of the senses and see the truth *in spite* of the testimony of the senses. Galileo is in fact the father of the *theoretical* rather than the experimental method. Of course, he insisted on the necessity of experimentation, not however as a final goal in itself, but as a *means to an end*. And what is that end? The end is the discovery of the *law*, because the universe is a lawful organism in which nothing can happen by chance. Moreover, that law must take a very definite form, it must be a *mathematical* law. Galileo points out that we cannot understand a poem written in a strange language with which we are not familiar. But the poems of nature are written in the symbol language of mathematics and thus we must familiarize ourselves first of all with this language, become conversant with it, and then apply it to the formulation of quantitative relations which can be found between the various parts of the vast organism called nature. This was basically the return to the old Pythagorean view that "number is the essence of things." Aristotle taught that quantity is only *one* of the categories by which the nature of things can be conceived. Galileo rejected this view, in favor of a shadowy world of numbers into which the entire universe could be resolved. Do not trust your senses, said Galileo; they deceive you with a multitude of *qualities* such as taste, odor, color, heat, and so on, while in actual fact all that exists in reality is the relation between *numbers*.

This was a new program, never since abandoned. It became *the* program of the exact sciences. We see it clearly stated by Copernicus in relation to astronomy. We see it strongly emphasized in Kepler. But it comes to the foreground as the basis of *all* nature in the work of Galileo. A new religion starts here, a new revelation, a new program for the thinking mind. Galileo poked fun at his colleague in philosophy who refused to look into his telescope because he did not want his belief in the categories of Aristotle to go to pieces. But let us assume that somebody had shown Galileo a world in which the laws of science go topsy-turvy, in which gravity works as repulsion instead of attraction, in which all the ordinary laws of motion and mechanics are put out of action. He would have refused to look into this world or to take it seriously. He

would have dubbed the whole apparition a bad dream, a nightmare, which just could not be true. One religion was put in the place of another religion, and this has to be recognized, lest we think of Galileo, Kepler, Newton, and all the other great minds of modern science as sceptical philosophers who refused to pin themselves down and wanted to obtain everything as the result of observations. The possibility of resolving all qualitative differences into quantitative relations—something already attempted by the old atomists Democritus and Leucippos—was put down as a *basic postulate* without which exact reasoning is unthinkable.

If, as we asked before, whether Aristotle had he had the same refined instrumentation at his disposal as Galileo, would have come to the same conclusions, we have now to deny this proposition. It was not only the improved instrumentation which arrived during the time between Aristotle and Galileo. There was the further fact that the mathematical sciences had taken in the meantime a tremendous stride forward. The Greeks were masters of geometry, but did not possess the mathematical technique of formulating quantitative statements in symbolic language. Archimedes, the towering scientific genius of antiquity, did make unprecedented discoveries by applying the mathematical method to physical phenomena, inventing the concept of center of mass, density, moment of force, hydrostatic pressure. But he could not create singlehanded a new dynamics. At the time of Galileo and Kepler the science of mathematics was already well advanced and became a cornerstone in the exploration of the physical universe, although it is quite astonishing to see that the mathematical method could be established as *the* scientific method in an epoch which did not yet possess the powerful tools of calculus. It was only the calculus, invented by Newton and Leibniz, which made the formulation of physical laws in adequate form possible, namely in the form of differential equations. And yet, both Kepler and Galileo could go a long way in discovering some of the fundamental quantitative relations in the realm of motion; Kepler in the realm of planetary motion, Galileo in the realm of general motion.

This predominance of mathematical thinking did not mean, however, that the experimental side of science was neglected. The experiment had its place because only by careful experimentation was it possible to find out *what* mathematical law was realized in nature. In this respect Galileo and Kepler differed in their temperament. Galileo was anxious to refine his measuring instruments in order to make accurate observations before the proper law emerged. Thus, he could establish the time of oscillations of a pendulum which he could then use as an exact time piece. Similarly, his experiments on the inclined plane established the laws of uniform acceleration which govern the motion of bodies under the action of gravity.

Kepler, on the other hand, was a man of mystical inclinations. He saw in the number relations of nature the revelation of divine wisdom, and thus he experimented with number mysticism in order to explain the basic structure of the universe. Only five planets were known in his day beside the Earth, viz., Mercury, Venus, Mars, Jupiter and Saturn, and great was his delight when he discovered the "music of the spheres," by inscribing the five regular bodies into spheres, or circumscribing them around spheres, and then finding that the radii of these spheres harmonize with the approximate mean distances of the five planets. These speculations of Kepler were in fact nothing but fancy, but he believed—and that is a remarkable thought—that the mathematics realized in nature cannot be of an arbitrary kind, must have a deeper meaning. Not until Einstein did this thought bear fruit on a grandiose scale.

We come to Newton, who succeeded in formulating the laws of mechanics with sweeping generality, thus including a tremendous variety of single observations in one universal law. On the basis of the laws of mechanics, with the added law which controlled the force of gravity, the entire constellation of the planetary system could be calculated if one started from a given initial position and initial velocity of the planets. Newton took over the scientific method of the Greeks, brought to perfection in their geometrical theories. Just as Euclid starts out with definitions, axioms and postulates, so does Newton. The only difference is that in geometry the basic postulates are suggested to the reasoning mind by the principle of simplicity and symmetry, while in physics the basic postulates have to be deduced from carefully conducted experiments. But, after establishing the basic postulates, the rest follows by exact mathematical deductions, in physics no less than in geometry.

Newton was a pragmatist in his scientific philosophy and again and again he came out against the idea of making hypotheses. "Hypotheses non fingo" (I do not make hypotheses) was what he emphasized again and again—a somewhat strange statement from somebody who was so bold in setting up scientific hypotheses. But what Newton had in mind was that he did not consider it important to have a definite *model* for the explanation of the mathematical law. The task of science is to establish the mathematical law and, if that law is found, the scientist has done his job. In his optical investigations, he expressed the opinion that *it looks* as if light were caused by very fine particles which leave the light source and travel in all directions with great velocity. But, he took a strong stand against those who invoked his authority for a corpuscular theory of light. His argument was that the mathematical law is all that counts, and that even if the corpuscular model fails, the laws of optics will remain what they are. What models we invent for the explanation of these laws are, strictly speaking, *outside* the realm of exact science.

The remarkable feature of Newton's system was that so much could

be explained from such a scarcity of fundamental concepts and fundamental postulates. It seemed that there are a few fundamental building blocks from which the entire edifice of nature could be erected. Newton covered, of course, only those phenomena which were already well investigated in his time and they were essentially of the *mechanical* type.

One hundred and fifty years later a similar development took place in the realm of *electricity*. The English physicist, James Clerk Maxwell, formulated the fundamental experiments of Faraday in the form of two mathematical equations which describe the interrelations between two basic quantities, the electric and the magnetic "field strengths." These equations were very different from those with which Newton operated. The picture of little hard balls which were concentrated in small portions of space and represented "material particles" disappeared and gave place to the concept of a "field." This "field" meant that the physical action is not concentrated in the particles alone but takes place *everywhere*. In fact, the particles could be looked upon as those portions of the field in which the lines of force terminated and which came into motion if the electric tensions which tried to pull the particle in every direction, did not balance each other, thus giving a resultant force. The basic postulates of the new theoretical scheme looked very different from the Newtonian scheme. Instead of singling out the material particles as the source of physical action, it was now the *entire space* which participated in the physical action and the particles were not more than *indicators* for measuring the action. But, that action was there even in the absence of matter. The mathematical abstraction had gone a long way from the much simpler Newtonian scheme, but the picture remained strictly mathematical. In fact, the picture could never have come into being without that great edifice of mathematical development which took place in the realm of mathematics after the invention of the calculus, particularly in the field of *partial differential equations* which came into the focus of mathematical interest around the beginning of the nineteenth century.

Maxwell did much more than merely find the mathematical equivalent of Faraday's experiments. There was one particular experimental fact that puzzled him greatly. In electric measurements it was convenient to operate with two different kinds of units, depending on the physical circumstances, viz., the "electrostatic" and the "electromagnetic" units. These units are in a definite relation to each other, and the conversion factor by which one unit could be converted into the other, could be measured experimentally. The result of the experiment was that the conversion factor came out numerically very nearly equal to the velocity of light. The measurements had nothing to do with any velocity or any light. It was merely a strange mathematical fact that the conversion factor had the dimension "length over-time," i.e., the dimension of a velocity. How astonishing that this velocity seemed to agree with the propagation velocity of light in empty space! What is the meaning of this, Maxwell asked himself. Could this be a mere coincidence? Isn't it much more plausible that this result expresses some deep-seated affinity between light and electricity? Is light a form of electricity or electricity a form of light? He looked at his equations and studied their structure. He noticed a peculiar asymmetry in the two sets of equations. In the one equation one term was missing. When he put in this term, the two sets of equations became entirely symmetric and balanced each other. The missing term was in fact so small that its direct observation was impossible at that time. But, after putting in this term, an astonishing thing happened. The equations now had the consequence that electricity did not propagate with infinite velocity but with that peculiar velocity which before appeared as the ratio of the two kinds of electric units and which was experimentally known to be light velocity. Thus Maxwell arrived at his electromagnetic theory of light by which he could derive all the fundamental properties of light as a consequence of electromagnetic action. Light ceased to be an independent chapter of physics, it became an appendix of electricity.

What Maxwell did was, of course, sheer divination; it was scientific black-magic of the highest type. Later on, after he completed his equations by sheer clairvoyance, he could give all kinds of physical reasons why that term had to be present. In actual fact it was an ingenious guess and experimental physics was not sufficiently advanced in Maxwell's time to give an actual demonstration of the existence of the "displacement current," as Maxwell called it. Fifteen years later, the ingenious German physicist, Heinrich Hertz, demonstrated in his laboratory the existence and propagation of electric waves and showed that they have exactly the same properties as light waves. Today, through the invention of the radio, the electric waves—which differ from light waves only through their much longer wave length—are common household property all over the world. They have shaped our lives perhaps more decisively than any other invention of the human mind. From the purely theoretical standpoint, the astonishing feature of the Maxwellian equations is that so few equations should describe such an overwhelming variety of physical phenomena. What Newton did for mechanics, Maxwell accomplished for electricity. Moreover, we observe that the kind of mathematics involved in Maxwell's scheme was of a much more abstract type than that employed in Newton's scheme.

In 1905 a young and, up to that time, unknown physicist published three papers which are today recognized as landmarks in the history of science. The one was on Brownian motion, the other on the hypothesis of light quanta, the third on the electrodynamics of moving bodies. It was the third paper which became of greatest importance for the future development of theoretical physics. The hallowed name of Albert Einstein will forever be associated with that "principle of relativity" which was first enunciated in this paper.

What Einstein did was not a formal accomplishment. He did not approach the problem from the standpoint of finding some mathematical equations which will describe a certain group of phenomena. Something much more fundamental was at stake, namely, the critical evaluation of the entire foundation of theoretical physics. Certain things which were always taken for granted, were put under scrutiny and their falseness proved. This was no longer mere physics and formal mathematics. This was a combination of philosophy and physics and mathematics. The experimental side of physics was by no means neglected. But the point where the experiment came in was somewhat shifted. Before we make any experiments, said Einstein, we should be clear about what we are doing in an experiment. We *measure* in an experiment. True enough; but, in order to measure, we must have a *platform*. We cannot be suspended in thin air, we must have firm ground under our feet. Where shall we take such a stand? We characterize the position of a point by its three coordinates, corresponding to length, width and height, and in physics we add time as a fourth piece of information. But, is there only *one* possible frame of reference? By no means. We can put our instruments here or there. We can make our measurements on earth. Undoubtedly the very same phenomenon could also be measured from some distant point of the universe. Would all those measurements agree with each other? By no means, since different observers use different coordinates. First of all, therefore, we have to know how these different coordinates are related to each other since we do not doubt that there is only *one* physical universe in existence but infinitely many *platforms* from which it can be described. But, how are we going to find the relations between these platforms, or "reference systems" as they are called? Here is the point where experiment has to be invoked. We have to ask our physical experience what kind of relation exists between the various reference systems. Up to a certain point we can make predictions on purely mathematical grounds. We can assume, for example, that the relation between the two types of coordinates will be of a *linear* type. This is not enough to establish the relationship. We need further experimental evidence. That evidence was provided by the Michelson-Morley experiment, which had clearly demonstrated that there is no "absolute motion" in nature. The negative outcome of this experiment, which was such a tremendous headache to the previous phase of theoretical physics, appeared in Einstein's analysis not as a paradox but as a natural result. Why should we expect the existence in the universe of an absolute frame of reference when it is in the very nature of coordinates that they cannot be absolute? The new feature of the transformation equations which related the various reference systems to each other was that there existed a fundamental velocity, namely light velocity—that very same velocity which was so important in Maxwell's electromagnetic theory—which had the character of an *absolute quantity* in each reference system. No matter in what reference system we made our measurements, light always propagated in every system with the same constant velocity of 300,000 kilometers per second. On the other hand—and this was the strangest feature of the new theory—time ceased to be an absolute quantity. There is no "time as such" in the universe. Every reference system carries its own time with it and if we change our frame of reference, we have to change not only the space coordinates but the time coordinate as well.

This was only the beginning of the road. In a few years, Einstein added so much to his previous accomplishments that he became universally recognized as one of the leading scientific minds of his era. But he himself was not satisfied. He was driven forward by a demon which followed him day and night with the same problem: "General Relativity." The relativity of a certain group of reference systems, as it was established in his paper of 1905, could not be enough. There must be a solution to the problem that "all reference systems must be equally good for the formulation of the laws of the universe."

Theoretical physics entered here a new phase of its development. Previously, the question was only to find *some* mathematical law which will fit the experiments. Here was a philosophical mind which asked fundamental questions and which was firmly convinced that an arbitrary scheme of mathematical description which cannot be justified on logical ground is only a makeshift solution which will not stand the test of time. "Hypotheses non fingo," said Newton, and this meant that he did not inquire into the inner workings of nature. If he finds a mathematical equation which satisfies the experiments he will accept that equation without asking further questions. Einstein was temperamentally akin to Kepler, who experimented with number mysticism in order to lift the veil of the innermost secrets of nature.

Einstein started to experiment with the mysticism of "Absolute Calculus" and in his efforts to find the proper mathematical apparatus he was happy to learn from his friend Grossmann in Zurich that this apparatus had been worked out by the mathematicians of the last century to great perfection. Here started that dogged uphill fight of Einstein which lasted for ten years and which is perhaps unparalleled in the entire history of the human mind: a fight which did not arise from any experimental puzzle but from a purely philosophical puzzle of the mind. One thing was clear to Einstein from the beginning: if he succeeded with the principle of general relativity, he would also have solved the riddle of universal gravity, as it was formulated but not elucidated in Newton's

mathematical theory. There was one feature of the phenomenon of gravity which, he knew, he would be able to solve on the basis of relativity, namely the puzzling fact, known already to Newton and yet never explained before, that *all matter* is subject to the force of gravity and that every mass has two aspects: it acts as an *inertial mass* and yet it acts also as a *gravitational mass;* these two masses are necessarily the same, although their operation is so completely different.

In ten years the great work was finished and established a victory of the philosophical mind, armed with the mighty tools of higher mathematics, which is without peer in the history of civilization. The Newtonian law of gravity was no longer *some* mathematical law, deduced on the basis of observations and confirmed by the motion of the planets. It was a law which had to be *that* law and could not be anything else, in consequence of the geometrical structure which is realized in nature. A new universe was created of a much higher inner meaning than that which we had before. The geometry of the world ceased to be of that simple Euclidean type that we had conceived as the only possible basis of the physical universe. The abstract tools of higher mathematics presented us with a curved world in which matter does not exist anymore as a physical entity but as a portion of space in which the curvature of the world becomes high. "God always geometrizes" said Plato. "God always geometrizes" said Kepler. "God always geometrizes" said Einstein. The dream of the old philosophers became reality, the universe was explained as a magnificent mechanism of lawful interactions and the law itself became the manifestation of a supreme mathematical wisdom which operates in the universe.

If we look at the edifice of physics before the advent of Einstein, we observe activities on *two floors*. Down in the basement is the experimental physicist, making observations with very refined instruments and establishing functional relations between quantities of various kinds. Here is one quantity x and another quantity y. If x changes like this, y will change like that. The results can be tabulated in the form of numerical charts, a more vivid picture being obtained if we condense our measurements into a graph. These graphs look somewhat like the charts we see sometimes in advertisements which show how the sale of a certain toothpaste or automobile has increased as the years went by. The difference is only that the graph need not go up all the time. It might go up but it might also go down and generally the functional relation between x and y may be of very complicated nature. If we go around in this basement laboratory, we find on the walls a large number of complicated charts of all kinds.

Now we move up to the *first floor*. Here the experimental instruments have disappeared. The hum of the laboratory gives place to the silence of the scientific study. The blackboards are full of mystical symbols, small and capital letters of the Roman and the Greek alphabet, sometimes even Gothic letters, and in addition some peculiarly shaped *d* and *s* symbols together with many brackets and parentheses. We recognize also the ordinary plus, minus, multiplication, and division signs of our ordinary arithmetic, but we see very little in the line of ordinary numerical calculations. The quantity appears here in its *abstract* aspects, as an entity with which we can operate and which satisfies certain basic properties, codified in the fundamental laws of *algebra*. Here is the domain of the *theoretical physicist* who takes the results of the experiments and formulates the mathematical laws which are behind the measurements of the downstairs laboratory. "Here is a mathematical equation between x and y which replaces completely the chart you have made on the basis of your experiments." The mathematical equation is a much briefer and much more powerful description of a quantitative relation than a chart. One can draw conclusions from it which would not be possible on the basis of a mere chart. Furthermore, the mathematical equation is unlimited in its accuracy while our experiments are always of limited accuracy, although the accuracy constantly increases with the gradual refinement of our measuring instruments, made possible by the steady increase of technological perfection. Of course, the equation deduced by the theoretical physicist, need not fit the experiments unlimitedly; it may happen that an equation which was good for hundreds of years becomes invalidated by later experiments, made under more refined or more general conditions. For example, the Newtonian equations of motion broke down when it came to velocities comparable to light velocity.

What is the method by which the theoretical physicist obtains his results? How can he replace a bewildering variety of numerical tables and graphs by one simple equation? He uses the method of *abstraction*. By the careful study of experimental results he enunciates certain basic *postulates* which seem to fit the experiments. These postulates are of a sweeping nature and thus have to be checked constantly by newer and newer experiments which may either corroborate or disprove the basic postulates. It is perhaps more adequate to say that later experiments may demonstrate the *degree of accuracy* to which a certain postulate holds or the conditions necessary that a certain postulate may hold. We can never hope to have found a postulate which will hold unlimitedly since the possibility is given that later experiments, made under more accurate or more general conditions, will show a disagreement between the results of the experiment and the prediction of the theory.

It is astonishing to see how theoretical research succeeded in finding relations between originally unrelated things, thus reducing the number of basic relations to a constantly diminishing number. Newton's equations of motion, augmented by the law of universal gravity, explained at

one stroke the motions of the heavenly bodies. The mechanics of fluids and solid bodies could be reduced to a small number of basic equations. The phenomena of heat found their codification in the laws of thermodynamics. The incredible diversity of electric and magnetic phenomena could be included into a system of eight partial differential equations, called the "Maxwellian equations." These equations contained the light phenomena as well, explaining light as an electromagnetic phenomenon. It seemed, around the end of the last century, that we had advanced very far in describing the phenomena of nature in terms of a few well-understood mathematical equations which left little room for further basic research, and seemed to indicate that theoretical physics had essentially attained its goal and could from now on rest on its laurels and devote its efforts to the exploration of details.

Yet, things were not so completely rosy. There was the puzzling Michelson-Morley experiment hovering like a tiny storm-cloud over the horizon, and what a storm it became a few years later when Einstein reshaped the entire Universe on the basis of this experiment. There was the other storm-cloud: the discovery of the electron as the basic building stone of electricity and the experimental discovery of the atomistic structure of matter, long suspected theoretically but not established before in uncontestable terms. How much more *complicated* was everything than what we had thought before. And yet, it was exactly during these years that the fantastic possibility ripened in Einstein's mind to erect on the existing *first floor* of the edifice of physics a new *second floor*, a *superstructure* of theoretical speculation in which the mathematical law is no longer accepted as a more or less accidental description of natural events, deduced solely from the experiments, but becomes explained on the basis of some sweeping *philosophical principles* which are realized in the universe and which can be recognized on the basis of abstract speculative thinking.

It is of interest to observe this metamorphosis of Einstein from a positivist of a Machian type to a religious mystic of the Keplerian type. When he started out on his career, he felt quite at home in the cool and sceptical atmosphere of the "first floor." He took the Michelson-Morley experiment for granted and argued that this experiment forces us to assume—without any questions as to why and how—that light propagates in every direction with the same velocity c, whether we are at rest or in uniform motion. Let us draw the mathematical consequences of this postulate. This was Einstein's Special Relativity. But he could not help feeling that much more must be at stake. Is it really true that we should accept the non-existence of an absolute reference system on purely *experimental* grounds? Is it not much more adequate to accept the validity of all uniformly moving frames of reference on *philosophical* grounds, simply because it is in the *nature* of a reference system—as a mere tool of description—that it cannot have an absolute but only relative significance? But if that is so, why should the reference systems in *uniform* motion be of such fundamental importance? Why not reference systems in *any arbitrary motion?* A scientist of purely positivistic ilk would never ask such a foolish question, and, even if he had asked it, he would shrug his shoulders and say: "It is not up to me to speculate on the principles which underlie the structure of the physical world. I do not know whether there *is* such a structure. My task is no more and no less than to *describe* the observable facts of the experiments in proper terms. To *speculate* on the more or less reasonable nature of that description cannot lie within my competence." But Einstein was haunted by the idea of the relativity of reference systems. He became an eager student of "absolute calculus" which deals with the problem of general coordinates in a formal way. He noticed that in all formulas of absolute calculus a peculiar quantity entered to which nobody had paid any attention from the standpoint of physics: the "metrical tensor." It was this quantity which decided the geometrical structure of the universe; who would have doubted that that structure could only be of the Euclidean variety? But, said Einstein, if this metrical tensor enters the equations of absolute calculus as a field quantity, why should it not be characterized by field equations of a similar structure as, for example, the Maxwellian field equations which characterize the electromagnetic tensor? Of course, he knew that Riemann showed the way by discovering the "curvature tensor" which is the basic mathematical quantity of a metrical universe. If this quantity vanishes everywhere, the manifold becomes Euclidean; but is it reasonable to demand that the *full* Riemann tensor should vanish? No, said Einstein, because the full Riemann tensor cannot be considered as an adequate tool for the characterization of a certain geometry. It possesses too many components. In four dimensions the line element possesses 10, the Riemann tensor 20 algebraically independent components. We cannot characterize 10 quantities by 20 equations since then our system is overdetermined. The natural thing is to determine 10 quantities by 10 equations. What can we do then to relieve the overdetermination? A very natural algebraic operation offered itself, called "contraction." The resulting tensor, called "the contracted curvature tensor" or "Einstein tensor" had wonderful properties. In the first place, it had exactly the right degree of determination by having 10 instead of 20 algebraically independent components. More than that, a slight algebraic modification gave a tensor which had all the earmarks of the so-called "matter tensor" of physics. It was a symmetric tensor of second order, similar to Maxwell's "electromagnetic stress tensor" which described the matter aspects of electricity. Moreover, in full analogy to that tensor, its divergence was zero, which is the mathematical description of the "conservation laws" of theoretical physics, that is the law that energy and momentum can never be lost but only *transformed* in the physical universe. The Einstein tensor thus gave a purely *geometrical* interpretation of matter in curvature terms. If the world were completely void of matter, its curvature would vanish everywhere and we would have that Euclidean world that we used to postulate as the true geometry of nature. The material particles, however, modify the picture. A material body of the size of the sun, for example, makes with its tremendous mass an *indention* on the flat background of the universe and thus sits there like a spider in its web, causing the planets to revolve in their orbits *as if* they were subjected to a specific force, Newton's "force of gravity." In fact *no* such force existed. It was solely caused by the fact that we described a non-Euclidean phenomenon in terms of Euclidean geometry. In its proper setting the "force of gravity" melted away, while the "force of inertia" remained and the age-old mystery of why every inertial mass has necessarily a gravitational aspect associated with it found its solution in completely unsuspected terms. Mass means, by its very definition, curvature, and, if that curvature is there, the straight lines immediately bend into ellipses. They remain straight lines in their proper setting, a world of Riemannian structure, but they appear as ellipses to us who project this Riemannian world into a flat Euclidean world.

If we look at the various stages of Einstein's development—the philosophical principle of the equivalence of all reference systems and, subsequently, the discovery of the importance of the metrical tensor, the experimentation with Riemann's curvature tensor and its reduction to the matter tensor—all this was done essentially on *speculative* grounds, urged on by the desire to *understand* and not by the desire to *describe*. No wonder that the originator of all these fantastic ideas changed his philosophy from the cool and sceptical positivism of his early years to the transcendental rationalism of his later years. Erecting a *second floor* on the edifice of theoretical physics, on which the mathematical laws distilled from the experiments and codified into postulates are now examined on philosophical grounds and are *re-discovered*, not as descriptions of experiments but as necessary emanations of the logical structure of the universe. This was an achievement that made the originator of this theory the founder of a new scientific philosophy and the fountainhead of a new era in the history of human thought.

What was the impact of these remarkable developments on the science of our days? Was the discoverer of General Relativity hailed as a new Messiah and recognized as the leading spirit of his age? History shows that the intellectual leaders of revolutionary periods are seldom recognized by their contemporaries and we can hardly expect that it should be different in our days. Indeed, when Einstein appeared on the tivity, the other from quantum theory. The one equation is the famous relation $E = mc^2$, the relation between mass and energy, deduced by Einstein from the general principles of relativity as early as 1907. As soon as it was established that space and time are not separate categories but form one unified manifold, it was inevitable that certain quantities previously unrelated should attain a new status and appear in a new light. At the time, when Einstein announced the new principle that every mass by its very existence is the seat of a tremendous amount of energy, there was not the slightest experimental evidence which would have supported this purely theoretical conclusion. Almost forty years later, the first atomic explosion demonstrated on an awe-inspiring scale the correctness of the deduction, and thus the logical operation of the forces of nature. Consider, on the other hand, the relation $E = h\nu$, the equation which associates a definite energy content with every vibration of the frequency ν. This equation is forced on us by empirical evidence and is indispensable for the elucidation of many otherwise completely incomprehensible phenomena, such as, for example, the photo-electric effect, explained by Einstein on the basis of the light-quantum hypothesis, coupled with the relation $E = h\nu$. Here is a law which we have to accept at its face value on the basis of experimental evidence. If we try to understand it we are at a complete loss, since we cannot find any mental picture which could reconcile us to the fact that a wave motion which has extension and amplitude should spend its energy at one single point and that its energy should depend on the frequency alone. We have become conditioned to the equation by its manifold applications, but the basic mystery associated with it has never diminished.

So, we see the two great theories of contemporary physics, Relativity and Quantum Theory, developing from opposite corners and viewing each other with a certain hostility. Einstein was convinced that nature cannot be reasonable in one corner and unreasonable in the other. The victory in the field of gravity was admittedly only a partial victory, but still it was a victory of no small dimensions. If it was possible to decipher the cryptic messages relayed to us through the planetary motions, thus discovering the Rosetta Stone which explained the inner meaning of nature's hieroglyphics, there was good hope that we could perhaps enlarge our foothold by cautious generalization. Perhaps, by reading the small print with the proper intuition, we could decipher more than what meets our eye at first inspection. Perhaps the illumination obtained in the field of gravity could be extended to the realm of electricity or even the quantum phenomena? Einstein's dream was thus to enlarge the field of relativity with the final goal of including *all* phenomena of nature in one unified scheme. Again and again, he thought that he arrived at the hoped for solution. While the world press reported with excitement each of these attempts, the professional physicists greeted the same efforts

with stony silence and indifference. The same Einstein who, a few years earlier, was hailed as the greatest scientific genius of all times, now had difficulties in even publishing his results. Soon he faded out altogether from the scientific arena of his time. What is the explanation of this astonishing transformation?

The explanation frequently offered is that the modern formulation of quantum theory demands Heisenberg's "uncertainty principle" while Einstein was unwilling to depart from the strict determinism of classical physics. In actual fact, the antagonism has much deeper roots. It is associated with that complete change of intellectual and cultural values which took place in the course of Einstein's life. During the years between the two world wars, unprecedented social upheaval occurred in the history of mankind. The dictatorships established in those years, whether of the Russian, Italian or German variety, had, in spite of ideological differences, certain basic patterns in common. They were all equally aggressive, equally destructive, equally ruthless, equally arrogant and power-mad. Moreover, they were characterized by an extreme pragmatism which denied any shade of transcendental orientation. The Nazi aggression made Einstein homeless in more than one sense. With Hitler's ascension to power, a whole era came to an end. It was the end of that refined and aristocratic culture of freedom in which the individual had the right to maintain himself unreservedly and without fear of suppression. It was the end of that era of cosmic reverence in which the exploring mind was awed by the overwhelming majesty of the cosmic forces and humbly admitted his own impotence. The new era was very different. It is inevitable that in defending ourselves against our enemies we will unconsciously adopt some of their attitudes. The postwar period, while conquering the spirit of cruelty and aggression, became not less pragmatic and mass-minded than the regimes against which the war was fought. If Einstein said occasionally that, in his opinion, the mystical feeling of cosmic reverence is the most beautiful and profound emotion that man can experience, such a statement appears to us as if it came from the dark ages. In our enlightened times, there is no place for mystical emotions, at least not in science. The increased demands of industry and technology have put new obligations on the scientist, apart from a truly frightening increase in their number. The scientist has to solve certain problems, delineated in contracts from industry or military establishments. The scientist has to do a job. He has to do research. He has to write papers and progress reports. He has to discover something at least once every year. In Einstein's time, this "has to" did not exist. One did not "have to" do anything. One was fascinated by the mysteries of the universe and tried to find an answer to them. Science was not a career, not a social position, not a job, but an inner obsession, which enslaved one's mind and fixed it with a fanatical one-sidedness on one particular aspect of thinking.

The changed social scenery had profound effects on the scientific climate of the times. Quantum theory provides us with a certain mathematical mechanism which allows us to go through certain mathematical manipulations and obtain the desired answer. If we do not understand what we are doing, why should we bother, as long as the answer is correct? The only important thing is to predict the results of experiments. "Savoir pour prévoir" (to know in order to predict) was the program of the positivist August Comte, inaugurated around the beginning of the nineteenth century. It holds sway without any change in our days. The idea that there is a "meaning" in the laws of nature is rejected as not worthy of serious consideration. "Meaning" is a metaphysical concept which has to be banned from the vocabulary of a physicist. What Einstein did, was no more than shadow-boxing. It is regrettable that such a great mind fell victim to metaphysical dreams and wasted his time on completely useless and irrelevant speculations. Quantum field theory has still its limitations, but we can imagine that the day will come when we will be in the position to give an answer to all reasonable questions. We will then have a scheme comparable to a language whose words we do not understand, but whose syntactic rules we possess to such an extent that we can construct flawless sentences in that incomprehensible language. We will then have everything we need because we will be able to predict, at least in the statistical sense, the results of any conceivable experiment. We may ask why it is so important to predict the results of experiments. The answer is that the steadily expanding industry and technology, the design of newer machines and gadgets, the perpetual rise of our standard of living, leans heavily on the predictability of physical action. Only by a thorough mastery of the laws of nature can we expect to cope with the steadily rising pressure for bigger and better machines, whether their aim is the increased comfort or entertainment of mankind, or the need to forge weapons for aggressive or defensive purposes. The social significance of science is thus on the increase but so also is the social pressure on the aims and methods of science, until a climate is reached which cannot be dissociated from the social values which dominate our historical period.

Under these circumstances, it is clear that our survey of the role of theory in contemporary physics can be censured as ill-balanced if the word "contemporary" is taken in too literal a sense. In a more balanced presentation, the importance of quantum theory should have been greatly extolled, while the contributions of the theory of relativity could have been telescoped into a few polite phrases. I realize that by not following this course I lay myself open to the charge of loading the dice in the wrong direction. It is, nevertheless, my belief that I can justify my procedure on stronger grounds than mere personal preference. First of all, science is not a fashion show which changes every six months. In relation to science the word "contemporary" cannot be interpreted in too narrow a sense. The cleansing storm which swept through our basic concepts of physics on account of the advent of relativity is still too vividly in our minds to dismiss it lightheartedly from our conscience. Secondly, in spite of all mechanization, in spite of the threatening spectre of automation, it is still true that we ourselves are not automatons but human beings. Man lives not on bread alone but has his spiritual needs as well. We have our transcendental yearnings and this is something that at least we who live in the free world have still the right to admit. The majority of us will never believe that the Cosmic God acts on the basis of mere chance or that the laws of reason are not applicable to cosmic happenings. The majority of us will not be willing to accept the particle-wave dualism, even if we are provided with a perfect catalogue which contains exact prescriptions when we should deal with a particle as a particle and when as a wave. In spite of all discouragements on the part of the epigones of modern physics, we will continue to look for a unified structure which we feel must exist behind the appearances. Thus, we are back at Einstein's program, although the actual solution of the puzzle may look quite different from what he envisaged.

Finally, let us assume that it is in fact true what the representatives of modern physics claim, namely, that the present foundations of quantum theory are absolutely the final ones which will never change to the end of the days, and that therefore any deterministic and completely field theoretical description of matter is in advance doomed to failure. In that case, Einstein's work assumes a significance which will never be repeated again. Pragmatism and positivism have been with us for hundreds of years. But it happened in our own day that somebody dared to challenge the physical universe with philosophically sound questions and obtained philosophically sound answers. A modern Jacob wrestled with the angel and did not let him go before he was blessed by him. The work of Einstein has shown that refined abstract thinking, armed with the mighty tools of advanced mathematics, can penetrate into territories which were never before open to the human eye. In an historical period of unprecedented destructiveness, in which irrationality and the aggressive instincts of mankind celebrated triumphs of appalling magnitude, we find a quiet oasis of peace and silence in which the mind plays on sunny meadows and demonstrates the invincible priority of the power of logic and reason. If a future historian will look back on our days, trying to find the dazzling light which outshines all the tragic deficiencies of our age, his gaze will not focus on the atom bomb, but on the humble figure of a soft-spoken man who, by a unique combination of physical, mathematical, and philosophical ingenuity created a new world picture of sweeping magnificence: *Albert Einstein and the Theory of Relativity.*

EDUCATION AND SCIENCE

By GUY SUITS

IT IS with deep appreciation that I accept the honor you have bestowed upon me. I would be less than sincere if I did not acknowledge a feeling of both pride and humility in hearing a citation such as the one just given. But the recipient of an award should also acknowledge an obligation—an obligation to meet the *future* challenge that is implicit in an award *today*. Perhaps that stimulation is the real purpose of a prize such as this one.

There are two aspects of this honor that make it especially gratifying to me. First, it memorializes the late Dr. William Procter, an eminently distinguished American. Dr. Procter's scientific talents were supplemented by the broad interests of a true humanitarian. Second, this award is made by the Scientific Research Society of America, an organization which, through its unique "family relationship" with Sigma Xi, has helped provide an essential bond between two key elements of our modern civilization.

It is easy to assume that a bond between *education* and *science* is completely natural and would exist even if no one gave the matter any particular attention. We take for granted the ideas that only education can provide scientists, and that science, as a primary force re-shaping the modern world, is automatically worthy of the educator's interest. However, as I will indicate, the bond between the two is somewhat frayed.

Historically, a mutual interest between education and science has *not* always existed. We have not yet celebrated the hundredth anniversary of the first granting, in America, of a Ph.D. in science. This latter event took place at Yale University in 1861. Many more years passed before science really became an integral part of American education, if indeed this has actually come to pass. It is a fact that many people graduating from college today have little or no understanding of science, and that not many decades ago it was customary—and necessary—for young Americans bent on a scientific career to take their advanced research training in Europe.

The mutual interaction of education and *industrial* research is new if

for no other reason than the fact that industrially-sponsored science is, of itself, a comparatively recent addition to the American scene. It seems apparent that a good educational system and a thriving pattern of industrial research have some common interests. It all sounds very pat: the schools provide training for the bright young men; industrial research provides opportunities for careers in science for the bright young men, who make discoveries, which make possible new industrial enterprise, which increases the nation's standard of living and thus provides the source of support for more and better educational institutions.

Of course, education and science—particularly industrial science—should work together because of these common interests. And, thanks to organizations like RESA, they do. But are we, as a nation, doing the job as well as we should? Have both scientists and educators fully identified the joint obligations, responsibilities, needs, and particularly, the *opportunities* in this relationship between education and industrial science? Has industry fully recognized the problems of modern education, including its financial support, and determined industry's appropriate role in this issue? Has education understood the motivation of industrial science and its tremendous impact on the economy, the standard of living of all of the citizens, and the national defense? Both fields abhor the thought of "losing," as we put it, our most skilled people to each other, yet our common interests would probably be served best if we had a much greater exchange of experienced people between the campus and industrial research centers. Most important of all—have educational administrators truly understood the need to teach science as a part of our culture? Have any of us—educator or scientist—really come to grips with the fact that, in the most scientific age in the history of the world, science itself has not become a matter of real interest and understanding to the individual citizen?

Please do not believe that I am going to provide any answers today. My purpose is to stimulate inquiry by making some personal comments about education and science. This audience provides an unusual opportunity because it is made up of men and women who are *doing something* about maintaining and strengthening the bonds between classroom and industrial laboratory.

Frequently I have been amazed to find so many similarities between *individual education* and industrial research. For example, both are popularly regarded as "good things." One trouble with a good thing is that it is so easy to believe that, if it's good, twice as much would be twice as good. We might, for example, conclude that to meet the Soviet threat we should train scientists and engineers at twice our present rate. Many of my waking hours are spent trying to determine how much scientific research is appropriately done by my Company. That question is about as complicated, and has some of the same elements, as the question of how much education is optimum for an individual. Perhaps an individual's investment in education should be in proportion to his intellectual capacity or his ability to absorb it, and perhaps a company's investment in scientific research should be in proportion to its technological capacity and hence its ability to absorb the results of research in furthering its economic objectives.

Passing from the education of the individual to education in a broader context, I note some similarities to industrial scientific research, also. The college president evaluating the relative merits of strengthening the Physics Department versus adding courses in Russian must feel some kinship with the industrial research director trying to divide a given amount of money between Information Theory and Polymer Chemistry. College presidents and corporation presidents could, with only a few changes in the words, interchange speeches on the subject of "growing costs"—in education and in research. There are some discussions in my business about the appropriate role of the federal government—and government money. It is reliably reported that educators also have varying opinions on this issue as it applies to their institutions. The management of research and the administration of education have some common denominators. In particular, they both deal with many intangibles, with strongly intrenched practices, and inadequate measurement. Both have problems in productivity that we are reluctant to face realistically. Both have basic problems of balance, with only intuitive methods for determining the desired answer.

Let's take a quick look at the question of measurement. In scientific research, we have superlative ability to measure natural phenomena and almost no ability to measure any of the economic attributes related to the work, except cost—let alone the political and sociological consequences. This lack was not serious when scientific research was only a ripple on the economic pond, but this activity now makes a big splash and it may become a tidal wave. The results of research have great leverage on the economy; they are a prime factor in the national defense; they affect the life of every citizen. Most importantly, with an eight billion dollar annual bill for research, and its resultant development and engineering, this activity has leverage on Mr. Citizen's pocketbook.

In industry, research and related technical activity is becoming a big component of cost, and it promises, or perhaps I should say threatens, to become much larger in the future. Hence we must be more businesslike in conducting this big business, and we are becoming dissatisfied with traditional intuitive methods of administration and management.

I am no expert on educational matters, but as an outside observer I would guess that there may be some need for measurement in education, too, and for similar reasons. For education has become a big business and business administration rests on measurement. The techniques of measurement are difficult enough, but the problem is a lot more serious if we don't know what to measure; and I am not sure any of us, in our groping for measurement, are really measuring the right things. For example, when educators stamp their product *magna cum laude*, have they measured the right things? I can think of some specific instances that would at least support doubts. When we measure the output of an industrial research venture, we frequently have our fingers crossed, too. I am not sure that we are measuring the right things either. I would like to leave the distinct impression that we both have a good deal of unfinished business in the area of measurements.

Let's focus a bit of attention on a word that is too much of a stranger in our midst—productivity. This word has all of the elements of irritation built into it, for its mere mention implies that there may be some dissatisfaction with the subject. Yet, in any economic system, this is a basic concept, and preoccupation with productivity in the economy as a whole increases each year, as we add new and more efficient machinery, manufacturing processes, better engineering methods, more automation, and more power to lengthen and strengthen the arm of each worker. Let's look at the result purely on the basis of power alone, which is by no means the whole story. In American industry, the installed horsepower per worker is now about 15. It is easy to determine that the sustained rate of effort for an individual is in the range of $\frac{1}{10}$ to $\frac{1}{20}$ horsepower, and, for a small side bet, I could identify a few individuals who work at a still lower rate. But, let's say that one horsepower equals ten manpower and go on from there. The average worker in industry thus has at his disposal about 150 manpower.

This basic relationship of power and mechanization to individual productivity is fundamental to the high standard of living in the industrialized countries of the world. Productivity paces the economy; not, however, without creating a few problems in the cultural areas of the same economy. Cultural pursuits may, by their very nature, not be adapted to mechanization, and may thereby price themselves out of reach of all but a few of our citizens. This appears to be a serious problem in mass higher education, and one in need of thoughtful study. The economic status of the teacher, especially the college teacher, has not been keeping up with the economy as a whole, and productivity is one of the factors. There has been little or no increase in teaching productivity to match the progressive increase in productivity in the manufacturing industries, in agriculture, in transportation, communications, distribution, and in the service industries.

The class of 15 or 20 students has been a constant denominator of college education for about two centuries with good cultural results, but with obvious economic consequences. One of the most serious problems here is tradition and habit and, until recently, no one had the courage to raise this question and to examine it carefully for possible solutions. The Fund for the Advancement of Education of the Ford Foundation has for a number of years, however, sponsored a series of studies of this subject. This work now appears to be yielding some exciting and very significant results, some of which have been incorporated into the teaching plan of the sponsoring institutions. I recommend that educators look into this work.

The rising costs of scientific research and the resultant and related technical work, have focused attention on the productivity factor in this kind of activity, also, with some distress in certain quarters. Since, in most cases, we can't satisfactorily measure the output of scientific or technical effort, we aren't entirely sure what is happening to productivity in this kind of work. But in many typical laboratory research operations, it is difficult to identify elements of organization, mechanization, services and facilities which would support firm claims to substantial and progressive increases in productivity in research operations. So industrial researchers, also, have some unfinished business and consequent opportunities in this area.

While I have been discussing some difficult problems of education and science, I would not leave with you the impression that negligible progress is being made in their solution. I might digress for a moment to describe the work of one of my associates, Ned Landon, whose tongue fits neatly in his cheek, and who recently produced what he calls "The Chief Executive's Utterly Exact Method for Measuring Scientific Research"—which goes as follows:

> I multiply your projects by the words I can't pronounce,
> And weigh your published papers to the nearest half an ounce;
> I add a healthy bonus for research that's really pure
> (And if it's *also* useful, your job will be secure.)
>
> I integrate your patent-rate upon a monthly basis
> And I figure what your place in the race to conquer space is;
> Your scientific stature I weigh upon some scales
> Whose final calibration is the company's net-to-sales.
>
> And so I create numbers where there were none before;
> And thus have facts and figures and formulae galore—
> And these volumes of statistics make the whole thing very clear:
> Our research should cost *exactly* what we've budgeted this year!

Now, more seriously, in the spirit of this interchange of ideas, I would like to discuss some additional relationships between education and science.

Science training in our schools must, first of all, be based on a better understanding of science and its practitioners. A primary motivation of the true scientist is a desire to understand nature. He believes that nature is orderly and that by properly formulating questions, he may learn the pattern and organization of natural phenomena. The starting point in

technological progress is often provided by scientists of this type, working at the growing edge of science, pioneering in areas never before explored.

The number of persons with aptitudes for this truly exploratory work is relatively small. Generally, they are found among the brightest students. It is essential, then, that they be identified at an early age, and encouraged and motivated so that they will realize their great potential. This, however, is *not* to say that all bright students should be encouraged to become scientists. Nothing could be more wrong—or more dangerous. Every field of human progress needs its share of the very best minds. Scientific progress in the future is not likely to be limited by nature, but rather by our progress in the social sciences, the humanities, the political and cultural order—in short by the total environment. We need to make *balanced* progress in many fields, not in one.

I have already mentioned that the greatest single objective in teaching science in the schools—and the one that has largely been overlooked—is the matter of making science a part of our culture.

It is a fact that there are still in existence "liberal" education programs that virtually ignore science. In a time when science is shaping the lives of all men, everywhere, this subject *must* be a matter of real interest and understanding to the individual citizen. For the individual citizen is the key element in the total environment, and he can no longer be isolated from science. Of course, no citizen can be expert in every field of science, and most citizens not in any of them. But it is vitally important that the level of public comprehension of science and its impact on people be raised materially.

Today's student will be tomorrow's citizen. As a citizen, he will have responsibilities as well as rights. As a responsible citizen he should know something about the history, dimensions, and trends of our technological civilization so that he can intelligently help to shape our total environment tomorrow.

The increasing role of science in government is particularly apparent to us in these days—and here (Washington). Our government depends on an informed citizenry; in a democracy we cannot rise to heights much greater than the levels of public enlightenment. In the long run, public opinion is a controlling factor. Each of us has been born into the most complex technological civilization the world has ever known. Unless people, individually, come to have an understanding of the methods, objectives, capabilities and the content of science—the basis of technology—their elected representatives cannot intelligently represent the "will of the people."

Further, if the representatives themselves do not have at least a modicum of scientific knowledge, they will not be competent to select advisors nor to evaluate technical advice.

Many of the decisions of government are made in highly technical areas that are fully understood only by technical specialists. Someone must choose the specialists. The future may depend upon how well the choices are made.

I will not excuse the scientist from a substantial responsibility in this task of developing a public understanding of science. Too often his contribution to the understanding of scientific progress, often based on his own work, has been less than optimum. Too often he succumbs to the temptation to mystify, when he should be bending every effort to translate and interpret and explain his progress to his fellow citizens.

Education, then, has a tremendous challenge. It must accomplish the double purpose of providing a continuing supply of scientists, and an informed citizenry, including representatives in government. This means that science must be taught not as a set of mystical facts, but as a living, breathing, and growing component of modern civilization. Much more must be done to interpret science and technology to the public.

Can I tell you how this should be done? No. But it is obvious where we must start—with the *recognition* and *identification* of this important problem. It will take the contributions of many—scientists and educators alike—to make progress here. And it would be presumptuous for scientists to claim an all-knowing role in the planning, for scientists have frequently been remiss in their duty to interpret science for the public. We often have shown ourselves ignorant of how the mind of the public works—at least as ignorant as the public has been of the workings of the scientific mind. Better communication is certainly indicated. And in the long run, we must look to the schools to do the job. Are we passing the buck? Perhaps, but we can justify our buck-passing if we give our schools the support they deserve.

The same science that is the source of so much of today's progress could also lead to mankind's destruction in the years to come. There are many things we must do in a material way for the survival of our civilization, but all of them will avail us nothing unless we somehow learn to live with each other. That's why *all* the bright young men and women must not become scientists. We need, somehow, to achieve advances in the social sciences comparable with those that have been made in the physical sciences. And I am certain that, to an increasing extent, knowledge of *physical science* will be an essential part of the education of every *social scientist*.

Alexis de Tocqueville was a Frenchman who came to America in 1831 to study democracy. Shortly thereafter he published a book called "Democracy in America." Many are familiar with his fantastic prophecies about the coming struggle between America and Russia. Another less well-known passage predicted a future for industrial research nearly a century before industrial research really began to come into its own.

"You may be sure," Tocqueville wrote, "that the more a nation is democratic, enlightened, and free, the greater will be the number of interested promoters of scientific genius, and the more will discoveries immediately applicable to productive industry confer gain, fame, and even power, on their authors. For in democracies, the working class take a part in public affairs; and public honors, as well as pecuniary remuneration, may be awarded to those who deserve them. . .Possessing education and freedom, men living in democratic ages cannot fail to improve the industrial part of science; and. . .henceforward all the efforts of the constituted authorities ought to be directed to support the highest branches of learning, and to foster the nobler passion for science itself."

Tocqueville understood, far better than many modern-day prophets, the proper relation between science, education, and other disciplines. Well over a century ago he recognized the growing role of science—and even industrial science. But in his writings he also made it clear that, in spite of its importance, science is just one of several essential elements of our civilization.

Am I suggesting that what we need today is another Tocqueville? Well, the trouble with individual prophetic geniuses is that we tend to believe their predictions more *after* they have come true than before. And in any case, we will be hard put to find many single individuals who can simultaneously grasp all the complexities of today's social, political, economic, and scientific problems. What we need is not so much a *universal genius* as a broad base of *mutual understanding* between *all* physical scientists and *all* social scientists. I think it is evident that education must play the key role in achieving this understanding. I speak, of course, as a physical scientist. For us it is essential to remember that man's future destiny will be achieved not *by* science alone, but *with* science as part of the whole.

A HUMANIST LOOKS AT MAX PLANCK

By JAMES C. O'FLAHERTY

THE following remarks about one of Germany's greatest scientists ought to be prefaced by a personal confession. I should like first of all to state that I was not particularly interested in Max Planck until I learned of his dramatic encounter with Adolf Hitler in the Reich Chancellery in the spring of 1933 over the Jewish question. I had long known that Planck was one of the creators of modern physics, and therefore I respected his name as I would that of Einstein, Rutherford, Heisenberg, Bohr, and others. These are stars that shine brightly in other skies than those I ordinarily scan. But when I learned that Planck had taken the part of a Jewish friend and of the Jews in general before Hitler, I became interested. Suddenly Max Planck sprang into my universe, that is, into that moral universe which concerns all men, no matter what the point of view from which they normally understand reality. For, suddenly, the ivory tower was deserted and the central nerve of Europe's most cruel and systematic tyranny was dangerously touched. True, there was no Luther-like defiance of the dictator, but there was resistance and unambiguous opposition to the spirit of Nazism.

It also seemed to me that it is important for the humanist as well as for the natural scientist to understand Max Planck, the man. It occurred to me that it is both possible and desirable for the humanist and the natural scientist to speak a common language on such subjects as this—all the more in times like the present, when the new emphasis on natural science threatens to widen the gap which exists between the world of the scientist and the world of the non-scientist.

The conflict between Planck and Hitler interested me from still another standpoint. It seemed clear that it marked one of those rare moments in history when two powerful and normally hidden forces of a given culture meet in a visible and significant clash. On the one hand, the turbulent, blind forces of a perverted German romanticism find their epitome in Adolf Hitler, who happened to be at the head of the German state. On the other hand, that remarkable German striving for an abstract understanding of life, for pure theory, finds one of its most impressive embodiments in Max Planck, the theoretical physicist. Thus two sides of the German character come into direct conflict with one another. It is as if the two natures of Goethe's Faust were objectified in German institutions, projecting into the public arena the struggle about which the disillusioned Faust laments. It is not simply the fame of the two men which makes their encounter significant but rather the fact that they were both spokesmen for powerful institutions, namely, the German state and the German university. It is true that the kinds of power represented by the German state and the German university are quite different, but one would make

a grave mistake to suppose that the German university has been a whit less powerful in world affairs than the German state. Its power is, to be sure, of a radically different nature from that of the state. Not the production of cannon but the production of *ideas* has made the German university a power to reckon with.

Consider for a moment the mark which the German professor has made for himself in the world since the Reformation. A professor of theology at Wittenberg, seeking light for his own soul, hammers out a new faith on the anvil of his lecture stand, and we have not yet done with it. A professor of philosophy in the remote town of Königsberg sets definitive bounds for the exercise of pure reason just as effectively as the inhabitants of the Low Countries have set bounds to the ocean that faces them. A professor of philosophy in Jena and Berlin propounds a dialectic from which stem fascism on the right and Communism on the left, and we need no reminder of the dynamic aspects of both systems. A professor in Vienna explores and analyzes the subconscious mind in such a way that, for better or worse, modern man regards himself as considerably lower than the angels. A professor of theoretical physics in Kiel and Berlin, with scarcely more physical equipment than pen, ink, and paper, rethinks the laws of thermodynamics so radically that he almost single-handedly creates the frame of the new world-picture of physics. A professor of physics at Zürich projects a fourth dimension as a mathematical and physical reality, thereby ushering in the atomic and interstellar-space age.

It is true that the achievements of the German scholar have not been universally admired at all times nor universally welcome. Nietzsche, a professor himself for a decade, was never too happy about the abstractness of the German scholar, and could indict him for his mental "rope-dancing." But he readily conceded the intellectual boldness and power of the German scholar of whom he wrote that "all the gods have learnt to be afraid." Whether we rejoice or weep at the advent of German scholarship on the world stage really makes little difference in the end. The fact remains that we must reckon with it in one way or another. Moreover, it is now quite clear that Hitler would have been far more dangerous than he was, had he recognized earlier in the war the virtually unlimited power of destruction which his theoretical physicists held in their hands, a power of which we have become acutely conscious and which figures prominently in our struggle with the Soviet state.

Planck's record is by no means the brightest of those who challenged the Hitler tyranny, for his individualism and his innate conservatism did not allow him to go beyond a certain point in actively resisting Hitler. But it is especially worth noting for two reasons: first, because of his unique position among the scientists, and, secondly, because his case represents the collision of a German academician with the highest political authority, a spectacle all too infrequent in German history, despite the record of the so-called "Göttingen Seven," of Theodor Mommsen, of Rudolf Virchow, and a few others in the nineteenth century. Academic freedom in Germany has meant for the most part freedom from religious pressure. As important as that freedom is, it is certainly no more important in our day than freedom from political pressure. In this latter area the German academic man does not appear so free to American eyes, and collisions between professors and the state are for that reason all the more important and instructive.

The question arises as to why Max Planck became an adamant if reserved opponent of Hitler in a period when many of his contemporaries found it prudent to collaborate or to suppress their genuine convictions. In order to find the answer to that problem it is necessary to consider the main biographical facts of Planck's life and to have some conception of his basic interests and convictions.

Max Planck was born April 23, 1858 in Kiel, Germany, where his father was professor of law at the local university. At the age of nine, Max Planck moved with his family to Munich, where he was privileged to attend the famous Maximilian Gymnasium. His mathematics teacher at the gymnasium, Hermann Müller, inspired the young Planck with a love of physics which never deserted him. Later, Planck studied physics at the universities of Munich and Berlin. At Berlin he heard lectures by the famous physicists Helmholtz and Kirchhoff. In 1879 he completed his doctoral thesis, "Concerning the Second Law of Thermodynamics" without any help or guidance from his professors. In 1885, after several years as *Privatdozent* at Munich, he became associate professor of theoretical physics at Kiel, and in 1889 associate professor in the same field at the University of Berlin. Within three years, he was promoted to a full professorship at that institution. On October 19, 1900, Planck announced his law of radiation to the Berlin Physical Society, and on December 14 of that same year he disclosed his formulation of the quantum theory. It was this theory and its extension which was to revolutionize the science of physics in the twentieth century. In 1901 Planck gained international fame with the publication of his treatise, *The Law of Radiation*. Recognition for the quantum theory came slowly in some quarters, but the proof of its correctness accumulated with the years, and, by 1918, Planck was awarded the Nobel Prize for physics. In 1912 Planck became a Permanent Secretary of the Prussian Academy of Sciences. In 1930 he succeeded the famous church historian, Adolf von Harnack, as president of the Kaiser Wilhelm Society for the Advancement of the Sciences, comprising 35 member institutes largely dedicated to the natural sciences. He remained at the helm of this great research enterprise until 1937, when he resigned because of advanced age. Planck's writings include, among his many scientific publications, essays on the philosophy of science and religion. Planck was a very accomplished pianist, having at one time, as a youth, considered taking up music as a career, perhaps giving it up only because he felt his talent was not sufficient for a career. Nevertheless, even at an advanced age, he enjoyed and drew great refreshment from music. Planck was an enthusiastic mountain climber. At the age of 72 he climbed the Jungfrau, and at the age of 79 the Grossvenediger. He died at Göttingen October 4, 1947, at the age of 89, having moved to that city from war-ravaged Berlin, where his home had been destroyed with most of his possessions. The measure of Planck's eminence in the scientific world today is the fact that the former Kaiser Wilhelm Society is now called the Max Planck Society.

II

After Hitler became Chancellor of Germany on January 30, 1933, he moved promptly to carry out his plans for the persecution of the Jews. Within a few months after his accession to power, he launched a campaign to oust Jewish persons in the employ of the state and to boycott Jewish business and professional men. One of the first victims of the anti-Jewish campaign was the famous chemist Fritz Haber, who had won the Nobel Prize in 1918 for the synthetic production of ammonia. Haber had been the director of the Kaiser Wilhelm Institute of Physical and Electro-Chemistry for twenty-two years. The swiftness of the attack on Haber was apparently brought about by the charge that he had retained too high a percentage of Jewish researchers in the institute of which he was director. As a result of Nazi pressure, Haber resigned his post as director on May 2, 1933. Max Planck was directly concerned with this political interference in the affairs of Haber's institute since that organization was one of the 35 constituent institutes of the Kaiser Wilhelm Society. Moreover, Haber was greatly admired and beloved by his colleagues. There was more than a touch of irony in the fact that Haber had done more than any other scientist in World War I to render German chemical warfare effective.

The Nazi treatment of Haber precipitated a crisis for Planck. In the first place, he was by inclination and training politically conservative. In fact, he had been, until the abdication of Wilhelm II in 1918, a loyal monarchist, and there is evidence that he remained inwardly loyal to his former monarch in exile. During World War I he had been so nationalistically minded that he had signed the statement of the German university professors which sought to justify the German invasion of Belgium as well as German militarism. The loyal monarchist, the son of a professor of law, and the heir to the weighty tradition that the German university professor does not interfere in matters of state, would not readily challenge the Chancellor of the Reich. Furthermore, Planck was naturally reserved and cautious in expressing himself. This fact came out strikingly in the physics colloquia at the University of Berlin, which were attended by such men as Nernst, Einstein, von Laue, and Schroedinger. In contrast to the readiness of most of the others to express themselves freely, Planck maintained a dignified reserve which in later years became, in spite of his quite regular attendance upon the meetings, a dignified silence [1].

The fact that Planck reached the decision to make a personal call upon Hitler to protest against the persecution of Haber and of the Jews in general makes him stand out among his colleagues and in the academic profession at large. As president of the Kaiser Wilhelm Society, Planck would normally have been expected to make a formal call upon the new Chancellor of the Reich. Thus Planck occupied a position of tremendous prestige in the intellectual world. But even so, his decision to call upon Hitler must have been a most difficult one. Planck must have understood something of the Führer's contempt for intellectuals, and he must have had some inkling of the difficulty of dealing with Hitler in personal interviews. At any rate, he was to learn from first hand experience what Hitler's press chief, Otto Dietrich, has to say of the dictator's capacity to fend off the attempt of others to influence him.

> "Hitler would listen only to their first sentences. Then for an hour he would knead and pound the subject with all the rhetoric at his command, would irradiate it with the peculiar light of his own system of thought. In the end his listeners were in a state of intellectual narcosis, incapable of urging their own point of view—even if they had been given the opportunity of doing so . . . if the other persisted in holding his ground, an outburst of hysterical rage would freeze the words in the unlucky man's mouth and the blood in his veins" [2].

For a man as reserved and cautious as Planck to be subjected to the Hitler technique of terror must indeed have been a dreadful ordeal. Small wonder that Planck's daughter-in-law, now living in Cologne, says she still remembers her father-in-law's "horror" when he spoke of his interview with Hitler [3].

The meeting between Planck and Hitler took place in the spring of 1933, probably during the month of May. The only authentic account of the meeting is that of Max Planck, published some fourteen years after the event by a German physical journal at the request of its editors. But this account is quite sketchy, and does not give the date of the meeting. I have dealt with the interview itself elsewhere [4], and therefore shall not recapitulate the known details except to say that Planck spoke first on behalf of Fritz Haber and then on behalf of those Jews who could be considered valuable citizens. Hitler was, of course, not willing to admit that any Jews could be valuable citizens, charging that they were all Communists. Before the end of the interview Hitler had worked himself up into a rage, albeit boasting at the same time that he had

"nerves of steel." There was nothing left for his distinguished visitor to do but to be silent and to withdraw.

In the years following his interview with Hitler, Planck said nothing about it even to his closest associates like Max von Laue and Otto Hahn. One would be in error, I think, to suppose that this silence stemmed from cowardice or intimidation. Quite the contrary. For we now know that Planck remained a staunch champion of the persecuted Jewish scientists, and continued to suffer at the hands of the Reich government as a result of his loyalty. It is true that he did not speak out publicly against Hitler nor engage in overt acts of defiance against the government. He even dissuaded others from what he regarded as futile acts of protest. A few days after Planck's interview with Hitler, Werner Heisenberg, another physicist and Nobel Prize winner, made the trip from Leipzig to Berlin to discuss with Planck the advisability of the simultaneous resignation of the best known German scientists throughout the Reich in protest over the dismissal of their Jewish colleagues. At Leipzig, the matter had been brought to a head by the dismissal of the mathematician Levi. According to Heisenberg's statement, Planck advised against such action on the grounds that it would not have the slightest effect on Hitler and his followers [5]. But Planck did not swerve an inch from the path of duty as he saw it. The fact that he saw it in the light of extreme conservatism was simply the result of his basic outlook.

If Max Planck had done no more than impugn the persecution of the Jews in Hitler's presence, however, he would, in the eyes of many, have earned his laurels as a champion of freedom. But Planck's act in interceding with Hitler on behalf of the Jews was not his last act of moral courage in the face of the Hitler terror. For example, he was one of the principal instigators and planners of the commemorative meeting for Fritz Haber on the first anniversary of his death, January 29, 1935. This meeting was held in open defiance of an order from the Reich government, relayed through the Rector of the University of Berlin, forbidding professors and all civil employees participation in the ceremonies for Haber. By presiding at this meeting, Planck gave notice that, at least in this particular area, he was prepared to act in disobedience to the state. The communication from the office of the rector of the University described the projected meeting as "a challenge to the National Socialist State" [6]. The Nazis were neither to forget nor to forgive that challenge.

We have ample evidence that Planck consistently defended the Jewish scientists whenever he was able to do so. For instance, we now know that the refusal of the Reich Propaganda Ministry to approve the award of the Goethe Prize of the city of Frankfurt am Main to Planck in 1943 was due to what it called his "persistent championing of the Jew Einstein up to the present moment" [7]. In 1937 Planck was forced to give up a so-called permanent secretaryship of the Prussian Academy of Sciences, an honor which he had enjoyed since 1912. On the occasion of his 80th birthday, April 23, 1938, no representative of the Reich government was present, a situation which would have been unthinkable under normal circumstances. This snub was official notice that Planck was still in bad odor in Nazi circles. Although Planck was apparently not connected with the organized conspiracy against Hitler, his son Erwin, to whom he was particularly close, was described by a key conspirator as "the circuit preacher" of the organized resistance, so incessant was his activity and so intense his devotion to the cause. It is very unlikely, however, that Max Planck knew anything of the conspiracy [8]. Erwin Planck was executed January 23, 1945 for his complicity in the plot against Hitler's life which had failed of its purpose on July 20, 1944. The words of Erwin Planck concerning the attempt on Hitler's life point to an objective idealism which is at least analogous to his father's thinking in other areas: "The attempt has to be tried," he said, "solely because of the moral rehabilitation of Germany, no matter whether an improvement of Germany's prospects can be reached thereby" [9].

It is eminently clear, I think, that Max Planck was no friend of the Third Reich, and that in his own way he resisted the encroachment of Hitler's minions upon his own important domain. He had supported the Second Reich wholeheartedly, and had sacrificed a beloved son to the Fatherland before Verdun in 1916. But he could not give the same support to the Third Reich, and another beloved son paid with his life for high treason against Hitler's state. It is true, as we have already noted, that Planck signed a statement along with ninety-one other German notables in 1914 justifying the German invasion of Belgium [10]. Later, however, he publicly repudiated that endorsement. The only statement bearing his name which supports the Hitler regime can, however, be proved to be a fraud.

III

The story of Max Planck's later life is the story of an unwanted conflict with corrupt state power. In him we see the German professor at loggerheads with his own government. But Planck was first and foremost the German university professor devoted to an abstract understanding of the physical universe. Insofar as he was devoted to abstraction he was typically German. Someone has said that, if a German were placed before two doors, one marked "Entrance to the Kingdom of God," the other marked "Entrance to Lectures on the Kingdom of God," he would inevitably choose the latter. This parable epitomizes that tendency to abstraction which is so remarkable a feature of German intellectual life. We may well doubt that Max Planck would have turned his back on the Kingdom of God in favor of pure theory, but we may be sure that, within his own chosen discipline, he preferred theory to empirical knowledge, not because he despised empirical knowledge but because something in his nature responded with unusual affinity to the call of the abstract. Planck possessed to a rare degree what we might call the inwardness of thought. That he was characterized to a somewhat lesser degree by inwardness of feeling we know from his lifelong devotion to music. But his first love was abstraction or inwardness of thought.

Planck's passion for an abstract understanding of physical reality came to light while he was still a youth in the gymnasium at Munich. Normally, Planck's prose style is quite sober, but there is a passage in his autobiographical essay which glows with genuine if subdued enthusiasm, namely, the passage which describes the impact of the law of the conservation of energy upon his youthful mind, as it was unfolded for him by his mathematics teacher at the gymnasium, Hermann Müller.

> "My mind (he wrote in *A Scientific Autobiography*) absorbed avidly, like a revelation, the first law I knew to possess absolute, universal validity, independently from all human agency: The principle of the conservation of energy. I shall never forget the graphic story Müller told us, at his raconteur's best, of the bricklayer lifting with great effort a heavy block of stone to the roof of a house. The work he thus performs does not get lost; it remained stored up, perhaps for many years, undiminished and latent in the block of stone, until one day the block is perhaps loosened and drops on the head of some passerby" [11].

This is an important passage for an understanding of Planck's mind. It should be noted that it was the objective, absolute, universal quality of this law that appealed to the young Planck. Herein lay, I think, the roots of his preference for theoretical over empirical certainty. An experimentally established fact can, to be sure, yield certainty, but the certainty it yields is a limited one. For there is always the possibility that a given explanation of an experimentally established fact, no matter how satisfactory it appears in the present, may be superseded by another explanatory hypothesis in the future. Moreover, a theoretical formulation which possesses inner coherence, that is, which is not self-contradictory, and which agrees with experimentally determinable facts, possesses an absolute quality which the empirically grounded hypothesis never can. This is simply another way of stating the logical superiority of deductive over inductive reasoning. Planck stands here in the great tradition of Leibnitz, and indeed in the specifically German academic tradition.

Planck's devotion to the absolutes of science gave him an inner certainty which was unusual and which enabled him to stand alone when necessary. That it was necessary for him to stand alone on numerous occasions was a result of the bias of the academic mind when he came to maturity. I know of no one whose record of steadfastness in the face of misunderstanding surpasses that of Planck. Because he had chosen an intellectual discipline which had not yet come into its own, he had to face much opposition to his work. To begin with, his teacher of physics at the University of Munich, von Jolly, advised him *not* to go into physics because there were, in his opinion, no new discoveries of consequence to be made! Upon completing his doctoral thesis, the young Planck was naturally quite eager to have the opinion of his most respected teachers, Helmholtz and Kirchhoff. The sad fact is that the former probably never read it, and, although the latter read it, he disapproved of it. Rudolf Clausius, professor of physics at the University of Bonn, whose writings had been so important to Planck in his formative period, could not be reached by letter or in person. The failure to elicit a word of appreciation or at least of stimulation from any one of these three physicists must have indeed been discouraging to the young Planck, for, of all the physicists in Germany, these were precisely the ones most likely to understand what he was attempting to do and to sympathize with his efforts. Later at Munich, where he was *Privatdozent* for several years, Adolf von Baeyer, the great organic chemist, let Planck know that he considered theoretical physics an entirely superfluous and sterile discipline. So stubborn was the opposition to the theoretical approach to physics that, on some occasions, propositions which had been theoretically derived and thereupon experimentally confirmed were denied on the grounds that there was probably some error in the experimental proof, as in the case of Dubois-Reymond's criticism of Planck's paper on the potential difference of the electrolytes in 1890 [12]. Max Planck could not have enjoyed the frustration and opposition of the early years, but his unflinching belief in the validity of what he was doing saw him through without leaving any trace of bitterness. Planck insisted that the intellectual enterprise requires faith. "Science," he said once in an interview, "demands also the believing spirit. Anybody who has been seriously engaged in scientific work of any kind realizes that over the entrance to the gates of the temple of science are written the words: *Ye must have faith*. It is a quality which the scientists cannot dispense with" [13]. Planck cited Kepler as an example of a scientist who triumphed because of his faith. "Compare him," he said, "with Tycho de Brahe. Brahe had the same material under his hands as Kepler, and even better opportunities but he remained only a researcher, because he did not have the same faith in the existence of the eternal laws of creation. Brahe remained only a researcher; but Kepler was the creator of the new astronomy" [14]. The fact that Planck's own career parallels that of Kepler in this regard makes his allusion to the point a telling one.

In the area of religion Planck may be considered an absolute idealist with a basically Christian orientation. His view of God parallels his view of objective physical reality in certain ways but diverges from it at an important point. Planck states in his essay, "Religion and Natural Science," that God "existed before there were human beings on the earth, and that he holds the whole world, believers as well as disbelievers, in

his omnipotent hands, since the beginning of eternity, and that He will continue to rule from his heights inaccessible to human imagination long after the earth and everything on it will have crumbled to dust" [15]. This is as unambiguous a statement of the absolute sovereignty of God as one is likely to find, and contrasts vividly with the conception of a limited deity held by contemporary thinkers like Alexander, Whitehead, and Charles Hartshorne.

Although Planck's conception of God's omnipotence and omnipresence parallels the notion of the complete objectivity and universality of physical reality, the methods of knowing God and knowing physical reality diverge radically. In the case of scientific knowledge, immediate experience and measurements derived from them yield reliable knowledge of the physical reality beyond them, especially when creative imagination inspires abstract thought. Faith in the ultimate success of the whole process is, as we have seen, also necessary. In the case of religious knowledge, however, directly and immediately given data point beyond themselves to God. Indeed, God stands at the beginning and end of religious cognition. Mensuration and abstract formulation are out of place here. Only faith is necessary. Planck held that the quests for scientific and religious knowledge run along parallel lines "and intersect at an endlessly removed common goal." In both quests man is a seeker after objective truth. The major difference between scientific and religious cognition is the need for a higher degree of abstractness in the former and a higher degree of faith in the latter.

Ethical decisions are made, according to Planck, on the basis of religious insight, hence are simple and intuitive in nature. Abstract reasoning cannot advance ethical choices in any significant way. "We stand in the midst of life," he wrote, "and its manifold demands and needs often make it imperative that we reach decisions or translate our mental attitudes into immediate action. Long and tedious reflection cannot enable us to shape our attitudes properly; only that definite and clear instruction can which we gain from a direct inner link to God. This instruction alone is able to give us the inner firmness and lasting peace of mind which must be regarded as the highest boon in life" [16].

Such was the mind of the man who found it in his heart to question Hitler's tyranny. This particular professor had indeed lived up to Nietzsche's description of the German scholar as one who through the nimble use of reason had made the gods of the intellect to tremble. For Planck had dethroned the classical physics and replaced it with a system which, in the words of Heisenberg, is "for scientific comprehensiveness and mathematical simplicity. . .not a whit inferior to the classical scheme of theoretical physics" [17]. Moreover, he achieved this revolution in science almost singlehandedly. One may draw inspiration from the life and work of Planck as a purely academic man. But

Planck is to be venerated on other grounds also. For he drew a line which Hitler could not cross. We might wish that he had done more. But it must be borne in mind that in Nazi Germany *no* opposition to the totalitarian state was tolerated. Planck did not believe with Eugen Fischer, Rector of the University of Berlin in 1934, that, since Germany was in the midst of a revolution, "objectivity was neither possible nor desirable" [18]. Planck believed that objectivity is always possible and that it was precisely in such times that it was most desirable.

The freedom of the German scholar from religious authority and dogma is a bright jewel in the crown of German culture and a welcome boon to civilization. But the world still awaits the day when the German academic man shall rise up against political authority and dogma with a power equal to that which he has exerted against religious authority. When that day comes, Max Planck will be cherished as one of the pioneers of the new freedom as well as a revolutionary thinker in science. It is one of the subtler ironies of history that, twenty-five years after Planck's painful encounter with the head of the German state, the Federal Republic of Germany should place the state's highest seal of approval on Max Planck, the man and the scientist, by having struck in his honor a coin bearing his likeness. Perhaps the new day of freedom has already begun to dawn.

REFERENCES

1. Walter Meissner. "Max Planck, the Man and His Work," *Science, 113*, 80 (1957).
2. Otto Dietrich. *Hitler*, trans. Richard and Clara Winston (Chicago, 1955), pp. 17–18.
3. Nelly Planck. Letter to the author, March 21, 1954.
4. See my article "Max Planck and Adolf Hitler," *Bulletin of the American Association of University Professors, 42*, 437–444 (1956). Reprinted without annotation in *Best Articles and Stories, 2*, 52–54 (April, 1958).
5. Werner Heisenberg. Letter to the author, November 28, 1956.
6. Edgar Y. Hartshorne, Jr. *The German Universities and National Socialism* (Cambridge, 1937), p. 135.
7. Confidential letter from the office of the mayor of Frankfurt am Main, dated October 26, 1943, to the business director of the Kaiser Wilhelm Society in Berlin. A copy of this letter was furnished to the author by Professor Otto Hahn, now president of the Max Planck Society.
8. On this point Frau Dr. Nelly Planck has written to the author: "Mein Mann hat seinen Vater nie mit den politischen Einzelheiten belastet, um ihn nicht zu gefährden, denn dazu war er zu alt. Meinem Mann war der Tag des Attentates auch nicht bekannt. Über die wirklichen Widerstandspläne wurde vollkommenes Stillschweigen bewahrt." Memorandum, September 4, 1958.
9. Franz Reuter, *Der 20. Juli* (Berlin, 1946), p. 31.
10. Professor Max von Laue says that Planck's endorsement of the statement justifying the invasion of Belgium was caused by false propaganda. "Noch während des ersten Weltkrieges hat Planck in einem öffentlichen Brief an H. A. Lorentz diese Unterschrift zurückgenommen; sie war durch falsche Propaganda veranlasst. Auch ich hatte dieses Pamphlet unterschrieben, weil damals niemand unter den Gelehrten etwas ahnte von den Propaganda-methoden des deutschen Generalstabs." Professor von Laue also adds that he considers the propaganda of the Allies at that time equally reprehensible. Memorandum to the author August 26, 1958.
11. Trans. Frank Gaynor (New York, 1949), p. 14.
12. Hans Hartmann, *Schöpfer des neuen Weltbildes* (Bonn, 1952), p. 93.

13. Max Planck, *Where Is Science Going?* trans. James Murphy (New York, 1932), p. 214.
14. *Ibid.*, p. 216.
15. *Scientific Autobiography*, trans. Frank Gaynor (New York, 1949), p. 159.
16. *Ibid.*, p. 185.
17. *Where Is Science Going?*, p. 19.
18. Shepard Stone, "Twilight of the German University," *Current History, 40*, 43 (1934).

MEASURING THE TEMPERATURES OF INCUBATING PENGUIN EGGS[1]

By CARL R. EKLUND and FREDERICK E. CHARLTON

EVER since the Emperor Penguin (*Aptenodytes forsteri*) was first found incubating its eggs in the antarctic winter at temperatures as low as −77°F., students of animal physiology have wondered how it was possible for the embryos to develop. Until 1957, these men could do little more than continue to ponder the question, because of the inaccessibility of the Emperor's breeding rookeries and the lack of a suitable instrument to study the phenomenon. At that time, however, with the establishment of various stations in Antarctica by the United States for the International Geophysical Year, a rare opportunity was presented for carrying out an investigation which might shed light on this question.

Of course, there was still the matter of instrumentation. As a first step toward obtaining the device needed, the senior author, then with the National Research Council, discussed the problem with Dr. Orr Reynolds, then Head of the Biological Sciences Division of the Office of Naval Research. It was pointed out that the instrument would have to be small enough to be placed within a penguin egg, and so designed that it did not need to be linked by wires with an antenna located outside the egg, as would be necessary if a thermocouple were used. The reason for this requirement was that the Emperor has no fixed nest, often moving many times while incubating its egg. This is possible because it carries its egg on the webs of its feet, and thus can move or "sit" at will.

At the time the request was made, ONR was considering plans for a miniature telemetering instrument which could be used in human physiology studies. In fact, a contract for the design and construction of this instrument was then in the process of being drawn up with the American Electronic Laboratories of Philadelphia. Fortunately, several of the instruments were completed in time to be used in our work. These were flown to New Zealand just before our departure for Antarctica in December 1956.

The authors were stationed at Wilkes Station, which is located among the Windmill Islands of Vincennes Bay on the Budd Coast of Wilkes Land at latitude 66° 15′ S., Longitude 110° 31′ E. Although this area is very rich zoologically, no Emperor Penguins were nesting there during our visit. Consequently, we decided to carry out the temperature experiment simultaneously on two summer-nesting species—the Adelie Penguin (*Pygoscelis adeliae*) and the South Polar Skua (*Catharacta maccormicki*). The studies were conducted over a nine-day period from December 11 to 19, 1957, on Clark Peninsula, one-quarter mile from Wilkes Station.

[1] Reprinted, by permission of the Office of Naval Research, from the December 1958 issue of *Research Reviews*, pages 1–6.

The temperature telemetering system consisted of a transmitter small enough to permit insertion in an egg, a receiving loop antenna placed over the nest, a low-frequency radio receiver, and an audio-frequency pulse-counting device.

The transmitter included a transistor oscillator, which was placed, together with a temperature-sensitive element and three mercury-cell batteries (life, about 100 hours), within a multi-turn single-plane directional loop antenna. Of an elliptical shape, and only about 1 by 1½ by 2 inches in size, it fitted snugly within a large skua or penguin egg. The unit was made waterproof by sealing it in plastic.

FIG. 1. F. E. Charlton, Electronics Technician Chief, USN, left, displays temperature telemeter which has been prepared for insertion in penguin egg shell held by Carl R. Eklund, Polar Research Division, U. S. Army Office of Research and Development.

To prepare the eggs for insertion of the transmitter unit, they were first cut in half lengthwise and emptied. The transmitter was then placed between the two halves of the eggshell and the edges cemented together. After the cement had dried, a glass syringe was filled with the appropriate albumen, and the substance injected through a 1/16 inch hole at the top of the egg. The hole was then sealed with cement.

To provide comparative data, transmitters were placed in both an Adelie Penguin egg and a skua egg. As the normal clutch of eggs for these birds is two, only one telemeter egg was placed in each nest, this being substituted for one genuine egg. It was desirable that one normal egg remain in the nest to serve as a control, since, if it hatched, it might indicate that the experimental egg had received a normal temperature. It was also felt that the bird should at all times sit on two eggs—the normal clutch size. It should be noted that upon completion of the study the normal egg hatched in each instance.

When the telemeter egg was returned to the nest, a two-foot diameter loop antenna was placed over it. The antenna was mounted on a six-foot pole held upright by a wooden stand seven feet high. The pole

Fig. 2 (a & b). The radio thermometer and empty penguin egg shell are readied for use. Left, the telemeter is placed in shell. Right, albumen is injected into sealed shell.

could be turned within the stand, the purpose being that if the setting bird rotated the telemeter egg into a no-signal position (the egg was placed in the nest with its internal loop horizontal) the loop antenna could be rotated so as to pick up the signal again.

A transmission line connected the loop antenna and the antenna tuner, which was located along with the receiving instruments in an 8 by 10 foot shelter, or "wannigan," mounted on a sled. This was located 180 feet from a skua nest and 130 feet from a penguin nest in which the telemetering eggs were placed. A window in one side of the shelter en-

abled observers to see both nests. Twenty-four-hour watches were maintained at 8- to 10-hour intervals by the co-authors, aided occasionally by several men from the station.

The system functioned in this way: The telemeter egg radiated a modulated electromagnetic signal, the number of pulses per second being determined by the temperature inside the egg. This signal was intercepted by the loop antenna placed over the nest, relayed by a transmission line to the low-frequency radio receiver located in the "wannigan," and there amplified, detected, and sent to the counter,

Fig. 3. Eklund removes one genuine egg from penguin nest and replaces it with egg containing telemeter.

which recorded the number of pulses. Even though the signals were continuous, a count was taken for each bird only every 15 minutes on a 24-hour basis, alternately on the skua and the penguin. The pulses transmitted by each experimental egg were carefully calibrated against a standard thermometer before the egg was placed in the nest, and they were recalibrated on completion of the study.

As stated previously, studies were conducted over a nine-day period. However, due to technical difficulties, some of the records obtained during the period of observation were discounted. The temperatures shown below are from those sections of the continuous record in which it was felt the instruments were operating normally. The skua records extend only from 2:30 P.M., December 15 to 11:30 P.M., December 17.

FIG. 4. Frederick E. Charlton orients loop antenna over penguin nest as Carl Eklund places egg containing telemeter under bird. Temperature signals picked up by the antenna are carried by a transmission line to a receiver located nearby.

The penguin temperatures represent a longer period.

Results of the study show that the average temperature of the incubating skua egg was 96.6°F., with a maximum of 103.5°F. and a minimum of 87°F.[1] As determined in a published study made previously by the senior author, the average body temperature of this species is 106.1°F., ranging from a maximum of 108.4°F. to a minimum of 103.4°F. The average temperature of the incubating egg was thus 9.5°F. less than the average body temperature.

FIG. 5. Charlton calibrates temperature telemeter. Instruments shown, from left to right, are receiver, counter, and telemeter (in bottle).

The average temperature of the incubating Adelie Penguin egg was 92.7°F., with a maximum of 98.2°F. and a minimum of 84.5°F.[1] The average body temperature of the Adelie is 103.8°F., ranging from a maximum of 106.4°F. to a minimum of 100.3°F. Thus the average temperature of the incubating egg of the Adelie was 11.1°F. less than its average body temperature, and 3.9°F. less than the average temperature of the incubating skua egg.

The greater range in the incubating temperatures of the skua egg is explained by the number of times the parent bird changes incubation duties with its mate. One of the pair of skuas under study was sprayed

[1] These data are part of a Ph.D. thesis being prepared by the senior author for the University of Maryland.

with green paint to provide easy differentiation of both birds on the nest. A continuous 132-hour record of the nest-changing was kept, and during this period the pair relieved each other on the nest 13 times, or an average of 2.36 times every 24 hours. The Adelie Penguin, on the other hand, may not be relieved by its mate for as long as two weeks, as in the case of the male when it first starts the incubation after the egg is laid. This probably accounts for the greater fluctuations in temperatures of the skua egg.

On the basis of studies made of the temperatures of chicken embryos, it has been found that a rise in body temperature of the embryo may be expected after the tenth day of incubation (the incubation period is then about half over), when the heat production of the chick embryo increases rapidly. If similar action takes place in the eggs of the skua and the penguin, the temperatures recorded during our study would be valid only up to about the first 10 to 15 days, after which the embryo would affect the temperatures.

The theory of application and basic design of the temperature telemeter were excellent for the work as carried out, and it is believed the instrument can be used in physiological studies to measure temperatures remotely. Unfortunately, we did not answer the question we had uppermost in mind when we undertook the study—determining the incubation temperatures of eggs of the winter-nesting Emperor Penguin. It is hoped, however, that such research will be carried out in the future. Through it, we will be able to determine if eggs of the Emperor actually do hatch at lower temperatures than those of other birds. If it is found that they do, we will then want to know what there is in the physiological make-up of the bird which allows this to happen. It is at this point that the work may take a turn that will directly benefit man, for the findings could have important medical applications.

SIMPLICITY AND ELEGANCE IN THEORETICAL PHYSICS

By JOHN D. TSILIKIS

SIMPLICITY and elegance are fundamental principles that circumscribe the aesthetic qualities of theoretical physics. These qualities pervade the logical foundations, the physical content and the mathematical construction of almost all results of physical theory.

I

The scientific methods for investigating the structure and the behavior of nature are observation and experiment. In these investigations, the constituent parts and the external factors of the phenomena are studied, and the connections between them are determined. The results of observation and experiment will furnish significant aspects of nature if they are reduced to their essential elements and if the conditions among the elements are explained by means of a few basic ideas. The accomplishment of this task is the primary objective of theoretical physics. Theoretical physics is the science that employs the results of observation and experiment to design a system of thought in terms of which physical problems are understood and dealt with.

Theoretical physics uses the results and the methods of mathematics in its attempts to describe and explain physical problems in quantitative terms. In pursuit of this task, the theoretical physicists invent physical quantities that express discrete attributes of physical objects and agents, represent these quantities by mathematical variables, and then use these variables to construct mathematical formulae that describe the phenomena in terms of relations and relative changes among the physical quantities. The formulae are symbolic constructions that give the cause or the chance of the occurrence and describe the pattern of physical events. The ultimate goal of theoretical physics is the organization of all data of experience into physical theories that bind all cognate phenomena under their common elements and that give a simple and elegant explanation for all physical problems.

Experience, mathematics and theoretical physics are interwoven in many respects. Experience is knowledge of the physical world accumulated by observation and experiment. Mathematics is the science that furnishes the process for resolving the contents of experience to their fundamental elements and the language for expressing these elements and their connections in formulae. Theoretical physics entwines harmoniously experience and mathematics in the construction of physical laws and theories that account for the structure and the workings of nature in terms of a comparatively small number of fundamental concepts.

Theoretical physics is a broad, deep and diverse inquiry. It has earned man's confidence and respect for its intellectual merits and practical uses. However, the methods and the results of theoretical physics are not sacrosanct and are open for question to everyone. At different times, the physicists have been willing to adopt quite different methods of approach and to change or revise former conclusions as experience augments and mathematics is developed. Regardless of purpose and methods, the theoretical physicists always take into consideration certain aesthetic criteria in their attempts to use experience and mathematics in the formulation of physical concepts, laws and theories. These aesthetic criteria are adequately embodied into the principles of simplicity and elegance. Simplicity and elegance are constructive standards which incarnate the proper manner for treating experience and mathematics in the construction of a coherent picture of reality that appeals to human reason and concurs with experience.

Simplicity and elegance are extremely common terms in scientific literature. Physicists and philosophers alike talk enthusiastically about the simplicity and the elegance of this scientific result or that scientific activity. Critical analysis on the meaning and the usage of these concepts in scientific literature suggests the mean mode of using them in theoretical physics. The ideas of simplicity and elegance become clear, distinct and objective if they are stated as principles that postulate the fashion of manipulating physical problems. A physical problem is a large group of physical facts about many cognate phenomena and conceals a pattern which describes the phenomena. Simplicity and elegance may be stated in the following working definitions.

Simplicity is the principle that postulates the resolution of a physical problem to the fewest possible mutually independent elements in terms of which the problem is quantitatively represented into simple mathematical formulae.

Elegance is the principle that postulates the adequate representation of a physical problem in mathematical formulae which bestow unity, symmetry and harmony among the elements of the problem.

Elegance is less restrictive than simplicity. It characterizes mostly the mathematical construction of the physical theories, for it postulates that theories must be adequately formulated so that they bestow unity and order into the vast expanse of accumulated experience. In contradistinction, simplicity often transcends the mathematical construction and originates in the foundations of the physical theories. It postulates that the theories must account for the phenomena in terms of the ultimate elements of nature. Thus, elegance is an essential quality of almost all results of theoretical physics, whilst simplicity predominates in the causal laws and theories.

II

Man desires to understand the physical world in which he finds himself and often tries to accomplish this by simplification. The search for simplicity has helped in identifying the elements and the factors and in clarifying their connections in the phenomena. The successes of this approach have developed the conviction that simplicity is a powerful principle in the hands of the physicist for attacking a variety of physical problems. Truth is usually simple, and the discovery of simple equations for the description of natural phenomena gives some assurance that simple results involve factual truths.

Newton was the first physicist who understood fully the significance of simplicity and used it successfully in the construction of the foundations and the formalism of classical physics. His theory of gravitation was the outcome of sheer application of the principle of simplicity for organizing a wide diversity of known physical facts. Newton formulated the theory on the basis of all scientific knowledge of his age. In the seventeenth century, the physicists knew that all material bodies fall to the ground according to Galileo's laws of falling bodies, that all projectiles follow parabolic paths, and that the planets move around the sun in elliptical orbits according to Kepler's laws of planetary motion. Newton analyzed these phenomena and laws and found that all three cases were actually the same physical problem. Each case contained a system of two bodies of masses m_1 and m_2 separated by a variable distance r_{12}. He assumed that the effects of all external factors on the system are practically negligible and then tried to determine the connections among the elements m_1, m_2 and r_{12}. After rigorous inquiry, he discovered that if he assumed, as a fourth element, a mutual attractive force F_{12}, operating between the two masses along r_{12}, then the connections among the elements of the problem are adequately represented in the simple equation:

$$F_{12} = G\frac{m_1 m_2}{r_{12}^2}$$

in which G is a constant whose numerical value is independent of the system under consideration. This is the equation of universal gravitation and in itself constitutes a theory, for it not only describes all the above phenomena but it also explains their occurrence and features by means of the concept of mass-attraction. From Newton's equation, one can deduce the laws of Galileo, projectiles, Kepler and many other physical results.

It is admitted that the preceding analysis is at most a possible interpretation of Newton's work. However, it is clear that all the information about the phenomena was known and awaited the appearance of Newton who was able to apply the principle of simplicity for organizing the phenomena under their common elements into one extremely simple equation.

In the theories of relativity and of quanta, the principle of simplicity has been extremely useful. These theories often deal with unusual physical systems and processes which cannot be inferred from experience or verified by experiment at the present time. Therefore, the physicists are obliged to resort to the principle of simplicity and assume a simple representation is possible for any phenomenon. Many physicists, among whom are notably Einstein and Bohr, have constructed simple models of physical systems and have invented "thought experiments" for visualizing unusual configurations of systems undergoing extraordinary changes in state. In a "thought experiment" the physicist imagines himself in a perfectly equipped laboratory in which he can devise any apparatus and perform any operation, provided that he does not contradict logic and basic physical principles.[1] Models and "thought experiments" have been very successful concepts in confirming many theoretical results. Striking examples are Bohr's model of the atom as a system consisting of a central nucleus of protons and neutrons surrounded by orbital electrons and Einstein's famous "box experiment" which demonstrated the equivalence of gravitational field and accelerated motion.

The results of theoretical physics lose in simplicity and gain in elegance as the scientific method improves in effectiveness and scientific knowledge broadens in scope. Simplicity is not an indestructible principle which can never be violated. It has its limitations. The assumption that the conclusions of theoretical physics must be simple gives good first approximations but is sometimes arbitrary and tends to divert the physicist from reality. The gradual penetration into the mysteries of the physical world uncovers details which dictate changes or revisions of the physical laws and theories into more complicated versions.

Deviations from simplicity and convergence to elegance are encountered in the transformations that occur in going from classical to relativistic mechanics. Classical laws are simpler in mathematical construction than the equivalent relativistic laws. The relativistic laws are broader and account for the phenomena with elegance. The law of motion of a material particle constitutes a typical example. Newton formulated the first version of this law, and his work is superbly clarified by the principle of simplicity. The phenomenon of motion involves a particle of mass m which is at fixed point P_0 of space when at instant t_0 of time a force F acts on it. According to experience, a particle under such conditions will move along the direction of force occupying the successive positions P_t at later times t_t. In the light of simplicity, the problem contains as elements the mass m, the variable distance r from P_0 to P_t, the variable time interval t from t_0 to t_t, and the force F acting on m in the direction of r. Newton assumed that the phenomenon is

independent of external factors. In relativistic dialectics, Newton's assumption implies that the relative changes of r and t are the same if viewed by an observer stationed at any point of the universe who compares the magnitudes of r and t with meters and clocks fastened to his post. Newton conducted several experiments on the problem of motion and arrived at the conclusion that all connections among the elements may be expressed in the equation:

$$F = \frac{d}{dt}\left(m\frac{dr}{dt}\right) = m\frac{d^2r}{dt^2}$$

in which $\frac{dr}{dt}$ and $\frac{d^2r}{dt^2}$ are the velocity and the acceleration of the particle with respect to the point P_0. Regardless of Newton's cautions, he and posterity used the second form of this equation which is sufficiently accurate for the manipulation of all problems of motion at ordinary velocities. However, the second form of this law is valid only if Newton's assumption is correct, for in this case the irreducible element m and the reducible element $\frac{dr}{dt}$ are mutually independent.

In modern times, theory and experiment have shown that Newton's assumption is valid only if the observer's post moves with exactly the same motion as the particle and remains always equidistant from the stationary point P_0 and the moving particle. In all other cases, the magnitudes of the elements m, r and t of the problem depend on the motion of the observer relative to point P_0. If the observer's post is stationary relative to point P_0, then the mass m of the moving particle depends on its velocity dr/dt and grows to infinitely large magnitude as the velocity of the particle approaches the velocity of propagation of light. The mutual dependence of mass and velocity of the particle is described in the equation:

$$m = m_0\left[1 - \frac{1}{c^2}\left(\frac{dr}{dt}\right)^2\right]^{-1/2}$$

in which m_0 and m represent the mass of the particle at points P_0 and P_t and c is the velocity of light. Lorentz and Einstein incorporated the variation of mass with velocity in Newton's law of motion in order to include in it motions at high velocities. Then, the law takes the form:

$$F = \frac{d}{dt}\left(m\frac{dr}{dt}\right) = m_0\frac{d^2r}{dt^2}\left[1 - \frac{1}{c}2\left(\frac{dr}{dt}\right)^2\right]^{-3/2}$$

The Lorentz-Einstein law of motion is definitely more complicated than is Newton's law, for it contains more physical quantities and involves more mathematical operations. It constitutes a case in which the principle of simplicity is partly violated, for the elements m, r and t are mutually dependent. Nevertheless, the relativistic law is the most accurate and the most general in applicability.

III

Theoretical physics aims at the formulation of theories that approximate reality. It must involve the presence of elegance, for truth and elegance have many things in common. Nearly all physicists are convinced about the presence of elegance in the universe and agree that elegance should also be present in theoretical physics. The sense for elegance is a natural power that springs from the depths of human existence and finds its prerogative for operation in man's curiosity about the physical world. Einstein, who lived and acted in the frame of such ideas, was of the opinion that the main motive of all scientific inquiries is disinterested curiosity for uncovering the "pre-established harmony" in the physical world.[2] A. D'Abro is of the opinion that elegance is a standard of perfection which theoretical results must meet in physical content and in mathematical structure. "In mathematics, as in architecture, certain coordinations are beautiful, others top-heavy or unsymmetrical. When, therefore, physical phenomena are translated into mathematical formulae, the theoretical physicist will always endeavor to obtain beautiful equations rather than awkward ones."[3]

The principle of elegance elucidates the arduous work involved in the design of a physical theory. A theory is a small group of equations which state the basic irreducible properties and conditions in a large number of cognate phenomena. The principle of elegance assists in the adequate selection of the physical content and in the appropriate formulation of the mathematical structure of the theory. The physical content is vested in a general physical problem which includes as special cases all phenomena that the theory is designed to explain. The mathematical structure is a system of equations which represent the problem in quantitative terms. The induction and the confirmation of the theory depend on experience, and the manipulations associated with the theory depend on mathematics. However, the manner of carrying out these operations and the architecture of the theory rest on the principle of elegance. A theory will be widely accepted if it bestows unity, symmetry and harmony among the elements of every relevant physical problem.

A splendid illustration is the electromagnetic theory whose rigor and success are due to the elegance it bestows on the phenomena associated with electricity and magnetism. By the nineteenth century, the physicists had developed numerous experimental laws and empirical rules for determining the relations between electricity and magnetism and the phenomena resulting from such relations. Up to that time the general

trend was to reduce all electromagnetic effects to mechanical, and this approach brought the science to a state of complication and ugliness. Electricity could not be reduced into terms of space, time and matter and the electromagnetic forces could only be described as acting at a distance and as being transmitted instantaneously through any medium, even empty space. Explanations such as these seemed implausible. All scientists believe that, at least in principle, a satisfactory explanation is possible for every phenomenon.

James Clerk Maxwell was the physicist who resolved these complications. He succeeded in devising the proper concepts and in using them to transmute the most pertinent experimental results into a general physical problem which represents adequately every electromagnetic phenomenon. Maxwell admitted electric charge as an irreducible quantity among length, time and mass and replaced the concept of "force action at a distance" by the concept of "field." He found that the essential electromagnetic effects are contained in (1) Coulomb's law of electric force, (2) Ampere's law of magnetic force and (3) Faraday's law of electromagnetic induction. Then, Maxwell used the new concepts and the three experimental laws as premises and induced the electromagnetic theory which is a system of four vector differential equations. Maxwell's equations are the mathematical construction of a general physical problem which involves the electric and the magnetic fields and their relations in space and time. The equations in the usual notation are:

$$\bar{\nabla} \cdot \bar{E} = \frac{4\pi}{c} \rho$$

$$\bar{\nabla} \cdot \bar{H} = o$$

$$\bar{\nabla} \times \bar{E} = -\frac{\mu}{c} \frac{\partial \bar{H}}{\partial t}$$

and

$$\bar{\nabla} \times \bar{H} = \frac{\epsilon}{c} \frac{\partial \bar{E}}{\partial t} + \frac{4\pi}{c} \bar{j}.$$

In these equations, all quantities are scalars except those with the bar sign which are vectors. In the equations, $\bar{E}$ and $\bar{H}$ are the electric and the magnetic fields at a point in space, E and μ are the electric and the magnetic properties of space in the neighborhood of the point, ρ and $\bar{j}$ are the electric charge density and the electric current density at the point, t is the time, and c is the velocity of light. The operator del, $\bar{\nabla}$, gives the space description of the fields in the neighborhood of the point. The operation del-dot-field measures the net flux of the field away from the point, and the operation del-cross-field measures the circulation of the field around the point.

The premises and the method of derivation of a theory are not unique.

By virtue of this fact, Maxwell was able to make the theory a simple and elegant argument in which the premises are acceptable, the equations follow logically from the premises, and the phenomena are explained adequately by the equations. The fields are properly defined in the first two equations and are suitably related in the second two equations. The equations bestow unity among all electric and magnetic quantities associated with any point of space and may be integrated for solving any physical problem associated with a finite region of space. The first and third equations express the electric effects and are symmetrical in all respects to the second and fourth equations that express the magnetic effects. There exist no magnetic quantities that correspond to the electric quantities ρ and $\bar{\jmath}$, and thus there are zero terms in positions of the magnetic equations symmetrical to positions of the electric equations in which are ρ and $\bar{\jmath}$ terms. The equations bestow harmony among the electric and the magnetic quantities, for they adopt adequately all quantities to one another in forming the connected whole of the four equations. The electromagnetic theory elucidates all critical points and misconceptions which were involved in the experimental science of electricity and magnetism. The theory demonstrates that magnetic effects are reducible to electric effects and that the electromagnetic forces are propagated with the velocity of light in the form of transverse waves. To the astonishment of many physicists, the theory showed that light is electromagnetic, wave in nature, and accounted for all results of physical optics.

IV

Thus, thought and inquiry disclose that simplicity and elegance are excellent means for manipulating theoretical physics on the creative level. These qualities both excite scientific curiosity and inhere in scientific creation. These qualities illuminate the physical concepts, laws and theories and render them discernible to the limited vision of the human mind. Finally, these qualities furnish a guide in theoretical research, for they constitute useful working tools for tailor-making an adequate conception of nature to the scale of human reason.

Simplicity and elegance stipulate the fashion in which mathematics and experience must be assembled into theoretical physics. Mathematics is abstract knowledge that represents the mechanism by which the intellect operates quantitatively. Experience is empirical knowledge that consists of acts by means of which the physical world manifests itself to man. Mathematics is founded on experience but is developed independently of experience, for intelligence exists independently of external reality. Theoretical physics is rational knowledge that expresses an exact correspondence between the images of the intellect

and the facts of the external world. The formulation of a physical theory, that is a symbolic system which intercedes between intelligence and reality, is not unique and depends vastly on the scientific outlook of the historic period in which it is invented. Simplicity and elegance postulate a sympathetic treatment of mathematics and experience into physical theories which transcend both in grasping an adequate picture of reality in terms of human intellect.

Simplicity and elegance are two aspects of the broadest idea in theoretical physics. This idea incarnates the central objectives of all physical theories. It implies that in a theory the foundations must summarize experience with parsimony, the formalism must be a well-balanced mathematical argument which employs concepts to induce from the foundations formulae with the least effort, and the formulae must account for the phenomena with economy. A theory which is construed by both simplicity and elegance envelops beauty, for beauty inheres in what is essential in the physical world. Reflective thinking reveals that beauty is the proper name for the idea which represents the harmonious fusion of simplicity and elegance.

An illustrious example is the theory of relativity which presents the physical world with serene simplicity and elegance. The theory rests on foundations which are logically simple and solicits formulae which account for the phenomena with economy. The formalism of the theory is perforce difficult but embroiders formulae with preciseness and conciseness. In the words of Dr. Hermann Weyl, the theory of relativity has swept away the old and familiar conception of the physical universe, "but only to make place for a view of things of a wider scope, and entailing a deeper vision."[4]

An examination of the concepts of space and time in classical and relativistic physics will suffice to bring forth the aesthetic charm of the theory of relativity. In classical physics, space and time are considered as two independent entities. Space is a measure of extension, and its structure is expressed by the three-dimensional Euclidean geometry. Time is a measure of duration and is one-dimensional. In this universe, the space separation and the time separation of two point-events are different for different observers. The heavenly bodies move in complicated curves as a result of being acted upon by gravitational forces. In relativistic physics, extremely elegant results were obtained by considering space and time as being two aspects of a more fundamental entity, the space-time continuum whose structure is expressed by a four-dimensional Riemannian geometry. In this universe, the space-time separation of two point events is the same for all observers. The heavenly bodies move in the space-time continuum in "straight line" paths, properly speaking geodesics. There exist no gravitational forces in the world of relativity, for they are identified as properties of the geometry of the space-time continuum.

Simplicity and elegance are unavoidable and inescapable in theoretical physics, and no physical law or theory which completely rejects them will stand the test of time. The dynamic successes of theoretical physics have led to the belief that its qualities characterize the operation of the intellect, the structure of the universe, or both. A comprehensive treatment of this problem must inevitably include the principles of causality, chance and objectivity, and is beyond the scope of this essay.

Of course, the subject has not been exhausted. It is admitted that the discussion has been restricted with certain prejudice, considering the immense size and scope of theoretical physics. Only a few highlights of physics have been considered and exploited and these have come to support the thesis of the essay. The answer for the inquisitive reader is that these highlights are among the most successful results of theoretical physics and as such constitute part of current scientific truth.

REFERENCES

1. G. Gamow, "The Principle of Uncertainty"; *Scientific American;* Vol. 198, No. 1, January 1958, p. 54.
2. A. Einstein, *Essays in Science*, Philosophical Library, New York, p. 4, 1955.
3. A. D'Abro, *The Evolution of Scientific Thought;* Dover Publications, Inc., p. 441, 1950.
4. H. Weyl, *Space, Time and Matter*, Dover Publications, Inc., p. 2, 1958.

THE WORLD OF FINE PARTICLES*

By JOHN TURKEVICH

THERE is a whole world of fine particles—of particles whose size ranges from those of small molecules to those of ordinary dust and sand visible with an optical microscope. Protein molecules, viruses, synthetic polymers, colloidal particles, cosmic dust, soot and fly ash particles from combustion, fall-out particles from atomic tests—all constitute a domain in which these entities have their own size, shape and properties. Their mode or origin, the nature of their growth and mutual interaction have a sufficiently unique character to justify special attention. There are many indirect methods of studying such particles: Light scattering, the scattering of X-rays, behavior in gravitational field, migration in an electric field and adsorption. No one of these methods gives the direct information that is given by the electron microscope—an instrument which has been conceived, developed and widely used within the last forty years.

The Electron Microscope

The theoretical basis for the electron microscope was laid in 1924 by de Broglie when he postulated that electrons can behave as waves. This idea was substantiated in 1927 by the electron diffraction experiments of Davisson and Germer in the United States and of Thomson in England. Thus, one component of a microscope, an electron beam with wave properties, was available for image construction. In 1926, the second component necessary for a microscope was discovered when Busch developed a magnetic lens for an electron beam. Having these two components it was a short step, in 1931, to the design of a microscope using a beam of electrons instead of light and a complex magnetic lens instead of the usual combination of accurately ground glasses. A noteworthy contribution to science was made by the large electronic companies—Siemens in Germany, Radio Corporation of America and the Philips Company of Holland who soon produced commercial models of the electron microscope. These were made available to scientists all over the world, relieving them of the design, construction and maintenance of these complicated electronic machines. At present, the British, the Japanese and the Soviets are also producing electron microscopes; today, just twenty years from the date of the construction of the first commercial electron microscope, there are over a thousand of these instruments in use in the chemical, biological and medical laboratories of the world.

Microscopes, whether optical or electronic, are very important to the scientist. They extend his most important sense of observation—vision. For, of all the senses man has, vision is supreme. It readily distinguishes

* A Sigma Xi-RESA National Lecture, 1957-58.

the unusual from the ordinary. It visualizes what it senses in spatial relationships. The high signal to noise ratio and the three dimensional integration give the eye its special importance. The reinforcement of this power by the optical microscope uncovered during the last three centuries new phenomena in almost all fields of scientific inquiry. The electron microscope with about the same additional power of observation as that attained by the optical microscope over that of the unaided eye, has, in the last three decades, uncovered many new phenomena and promises, in the next three centuries, to discover still greater worlds for scientific exploration.

Magnification, resolution and contrast are three figures of merit which may be used to evaluate the performance of a microscope. Of these, magnification is not an exacting criterion. Magnification without disclosure of detail—empty magnification—limits the usefulness of the concept. A more valid criterion is useful magnification, magnification which reveals the greatest detail without straining the human eye. For an optical microscope its value is about two thousand fold, while, for the electron microscope, it is about five hundred thousand. These figures mean but little by themselves unless one considers the other two quantities: resolution and contrast. A resolution of n millimeters is defined as the ability to distinguish as two, points separated by a distance of n millimeters. The resolution of an unaided eye is about a tenth of a millimeter or a million Ångströms where one Ångström is 10^{-7} millimeters. In terms of atomic quantities, an eye can distinguish as two, points separated by a million hydrogen atoms. The optical microscope, under favorable conditions, has a resolution of two thousand Ångströms. An eye, aided with such a microscope, can appreciate detail within a limit of two thousand hydrogen atoms separation. A routine electron microscope has a resolution of fifty Ångströms. With careful design of the instrument, favorable specimen preparation, skillful and patient manipulation, a resolution of eight Ångströms or eight hydrogen atoms may be realized. There are very able and imaginative microscopists who, with special preparations and at propitious times, claim a resolution of three Ångströms. It is then that men's desire to see the individual atom seems close to realization. This has been done recently in the brilliant experiment of Menter who was able by means of the moiré effect to obtain a direct visualization of the platinum atoms in a crystal of platinum phthalocyanide.

Contrast is the third important factor in determining the efficiency of microscopic observation of fine particles and the fine structure of matter. If there is no contrast between the sample and the supporting membrane, high magnification and powerful resolution are of no avail. The difficulty is then similar to that which one encounters when one wishes to distinguish the features of a spotless white elephant standing against a white background. In optical microscopy, staining with dyes which

have different affinity for acidic and basic parts of the tissue offers the additional possibility of correlating chemical properties with morphology. In electron microscopy, contrast is based on the increase in scattering power of elements with increased atomic number of the sample. It is, therefore, imperative to have as a supporting membrane a material with as low an electron density as possible. A very important development in increasing the contrast has been the use of evaporated carbon films for mounting the specimens. A carbon film of fifty Ångströms thickness is strong, has no apparent structure and gives little background scattering of electrons. Contrast in the specimen itself can be increased by shadow casting or by staining with compounds containing heavy metals. Shadow casting is a method of enhancing the differences in the geometrical form of the specimen by subjecting it, prior to observation, to an evaporation in a high vacuum by a heavy metal from a point source. The latter is situated at an angle with respect to the plane of the specimen. Under these conditions the atoms of the heavy metal traveling in straight lines only coat that side of the specimen which the source can see. The areas in the shadow of the specimen do not receive the evaporated heavy atoms. The resulting electron micrograph shows an enhanced contrast which reveals the three dimensional geometrical contours of the specimen. Since electrons lack color and the scattering of the electrons by the specimens is not related to their chemical structure, the large collection of dyes which are used so effectively in optical microscopy is lacking in electron microscopy. Successful use has been made of a limited number of electron stains such as phosphotungstic acid, osmic acid and the like. The field of electron stains is one that still remains to be developed. Its importance is great in solving the relation of the appearance in the electron microscope to the chemical structure of the specimen.

During the last two decades many specimens have been examined with the electron microscope using many different techniques of sample preparation. Accounts of these have been given in several books and in two previous Sigma Xi National Lectures. The present article is a report of the work carried out at The Frick Chemical Laboratory at Princeton University, surveying the field of fine particles and studying quantitatively one type of fine particle—colloidal gold.

Fine Particles in Air and Vacuum

Let us first turn our attention to the fine particles present in air—the aerosols. When magnesium is burned in air, as when a photoflash is ignited, a white smoke is produced. This, examined with the electron microscope, is seen to consist of cubes of different sizes attached to each other in filaments. Dr. J. Amick and the author have examined smokes produced by arcing thirty different elements in air and examining them with the electron microscope. It is interesting to note that the oxide smokes so produced had characteristic shapes, magnesium as cubes, zinc as platelets which serve as junctions to unite four needles, molybdenum as eight sided plates, chromium as six sided plates and iron as six sided drums. The form of the particle is often characteristic of its chemical composition and it is quite possible to develop an analytical technique which will identify chemically the particle by the morphology that it assumes after known chemical reactions. When a sufficiently large

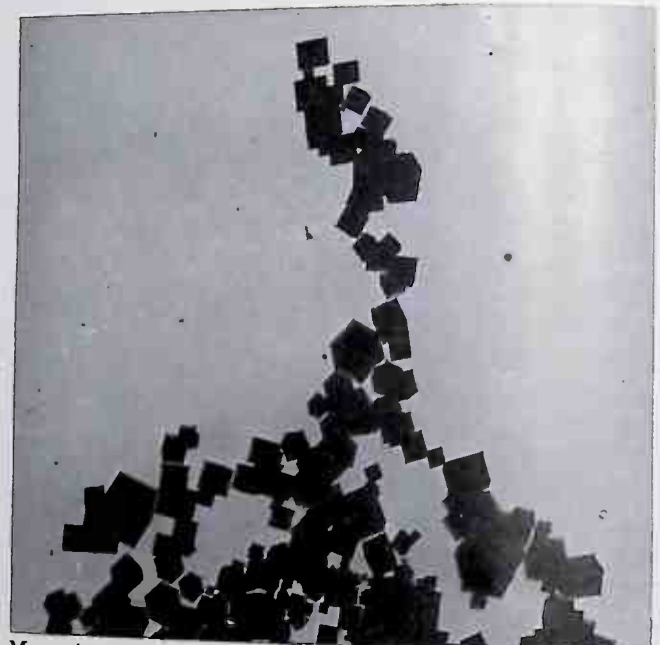


FIG. 1. Magnesium oxide smoke produced by burning magnesium in air. Note the cubic form and the mode of attachment. Magnification 26,000×.

number of smoke particles were collected, they could be characterized by electron diffraction. In all the oxides examined the crystalline structure was found to be the same as that reported for the oxide in bulk of the element arced. The mechanism of formation of metallic smokes was studied by quantitative measurement of the size distribution of metallic smokes. Fourteen different elements were vaporized in a high vacuum and in an inert gas and collected at various distances from the point source. The deposits so obtained were examined in the electron microscope. At high vacuum, the vaporized material traveled in atomic form in a straight line from the point source to the collecting membrane. Whatever fine particles were observed on the membrane were due to those formed by surface migration on the membrane. When an inert gas is introduced between the source and the collecting membrane, a smoke is produced. The mechanism of smoke production involves three

steps, nuclei formation from the vaporized atoms, growth of these nuclei into particles and the coagulation of the particles into filamentous aggregates. Separation of these three steps can be attained by varying the pressure of the inert gas and the distance of the collecting membrane from the source. It was found that the nucleation depends on the super-

FIG. 2. Zinc oxide smoke produced by burning zinc in air. Note the sharp needles originating from plates. Magnification 25,000×.

saturation and the surface tension of the element that is vaporized; that growth takes place very slowly. The coagulation process takes place more rapidly than growth. The result, at short distances from the source, is the formation of small particles and a structureless background of atoms which reached the collecting membrane without nucleating in the gas phase. At large distances from the source and at several millimeters of mercury pressure of the gas, the deposit on the collecting membrane consisted of threads which intertwined to form complex aggregates. The growth of the nuclei may be markedly accelerated by the introduction of impurity atoms. Thus, when magnesium is evaporated in helium many small particles are formed. Nucleation takes place readily but the nuclei do not grow. If, however, some copper vapor is introduced, large platelets are formed. Pure materials do not seem to grow readily in the gas phase and this may be due to the perfection of the nuclei formed from pure materials. Only when fine particles have imperfections, chemical or physical, is there a driving force for the growth. To be effective for growth, the imperfection, as Frank pointed out, must perpetuate itself during growth. A good example of such imperfection is the spiral dislocation shown in the crystal of the hydrocarbon. This spiral dislocation is responsible for the growth of the particle. One thus comes to a very intriguing concept—that the formation of a crystal with its ordered array of atoms from a disordered gaseous state requires, for its success, a definite type of disorder.

Collection and Characterization of Fine Particles in Air

The atmosphere around us is full of fine particles: Carbon particles from the exhaust of automobiles, smoke from industrial plants, fly-ash and soot from home furnaces, pollen from the flowers and, more recently, at certain times and certain places, fall-out particles from atomic tests. Characterization of the inorganic particles is complicated by their high dilution in air and the presence of organic matter contaminating the air around us. Filtration through cellulose or synthetic fiber filters is the most efficient and convenient way to collect particles from air. However, such filters cannot be used directly for observation in the electron microscope. T. Streznewski and the author have developed a method of removing the organic particulate matter and the filter fibers by oxidation at room temperatures with oxygen atoms. The supporting membrane for the filter fibers was a two hundred Ångström film of silicon monoxide and the oxygen atoms were obtained from a discharge tube. Oxidation takes place readily at room temperature. The inorganic particles can be examined in the electron microscope. The method of microincineration can be extended to the study of biological tissue sections before and after staining with electron stains.

Quantitative studies were made of the rate of this oxidation by measuring the change in particle size distribution as a function of oxygen atom exposure. It was shown that the width of the size distribution did not change as the oxidation proceeded indicating that the rate of oxidation was determined by the surface reaction. The latter was studied by measuring optically the rate of erosion of evaporated carbon films of 50 to 600 Å. thickness by oxygen atoms. It was found that the oxygen atom attacks the carbon with a fifty per cent efficiency and that this atomic reaction has no temperature coefficient in the temperature range from

room temperature to that of 120°C. Thus, electron microscope observation can be used to study quantitatively some aspects of the combustion of fine particles of carbon.

Fine Particles in Aqueous Colloidal Solutions

Another type of system of fine particles is that found in aqueous colloidal solutions. An examination of the various types shows spherical particles in colloidal gold, colloidal silver, rubber, polystyrene latex and

FIG. 3. Fibers of chrysotile asbestos. Note the tube through the fiber. Magnification 19,500×.

fresh sols of silica or alumina. Other colloidal systems, such as vanadium pentoxide, aged alumina, or chrysotile asbestos show submicroscopic rods. Chrysotile asbestos and some other minerals have an unusual morphology in that the rods are hollow. Plate-like particles are found among clays, aluminum oxide monohydrate, tungsten oxide and other fine particles. Under appropriate conditions, one can prepare colloidal gold particles as plates, rather than as spheres. The mechanism of the formation of the particles from the atomic and ionic species in solutions is of some interest. Thus, in the preparation of the plates of aluminum oxide monohydrate the following steps can be distinguished by electron microscope examination. The first product of the interaction of the aluminum ion with the hydroxyl is the formation of small granules of aluminum hydroxide. These granules soon aggregate to form flexible chains. These grow in length, developing lateral forces. If the latter are strongly developed

FIG. 4. Fibers of vanadium pentoxide. Magnification 166,667×.

the fibers line up forming a platelike crystal. If, on the other hand, the lateral forces are not strongly developed, a gel will result with a structure similar to that of a pile of randomly oriented match sticks. A similar type of particle formation takes place on aging in the transformation of the short slender rods of freshly prepared vanadium pentoxide into long rigid rods which adhere laterally into plates and finally form three dimensional crystals.

The detailed mechanism of the formation of fine particles in solution still remains to be worked out. One generalization can be made from a survey of a large number of fine particles in solution: The formation of

FIG. 5. Aged aluminum oxide sol. Note fibers in background and crystals that are formed from them. Magnification 166,667×.

the large particles is not a random process but follows certain laws which cannot be readily deduced from the first principles of electronic and atomic dynamics.

Fine Particles in the Biological World

The biological world has its own fine particles—protein viruses, enzymes spherical in shape; muscle fibers, chromosomes, tobacco mosaic virus in form of rods and fibers. The mystery of their formation, the relation of their texture as revealed by the electron microscope to their molecular structure is a fascinating subject. One of the most remarkable electron microphotographs is the one taken by Drs. S. Mudd and J. Hillier of the attack of a bacterium by a bacteriophage. The electron microscope has shown that this virus particle was not a simple sphere but that it had a tail. When the virus attacks the bacterium it plunges its tail into the outer membrane of the bacterium. Material is transferred from the virus to the bacterial cell. Radiochemical evidence using radio-

FIG. 6. Bacteriophage attacking bacteria. Electronmicrograph of J. Hillier and S. Mudd. Magnification 33,333×.

phosphorus and radiosulfur showed that it was a radiophosphorus that was transferred. The presumption is that the nucleoprotein material is transferred. After this attack by the bacteriophage, the bacterium does not divide into new bacteria but explodes into a swarm of new virus particles. Biologically, this process represents infection and death of the bacteria. It is, however, a death through transfiguration. Chemically, it is merely a crystallization process. The nucleoprotein material in the bacterial cell, after nucleation by the virus, becomes a constituent of virus particles. The building blocks for this crystallization and particle for-

mation process, the nucleoproteins, are complex in structure. Let us turn to a much simpler process of particle formation, the formation of small gold particles from the gold ions present in aqueous solution.

Colloidal Gold—Formation

Colloidal gold has been the object of admiration and study since the dawn of chemistry as a science. It was the *aurum potabile* of the alchemist. It was used as an ingredient of the *Goldwasser of Danzig* and its ruby color

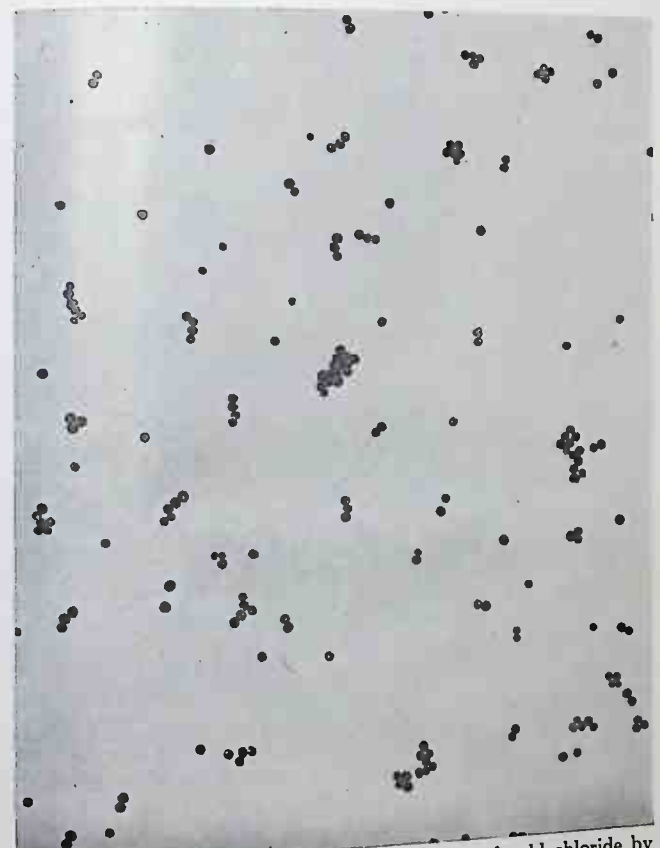


FIG. 7. Colloidal gold particles obtained by reduction of gold chloride by sodium citrate. Magnification 50,000×.

has been admired by all who have an appreciation for beautiful color. Its composition and properties were studied in the middle of the nineteenth century by Michael Faraday and Thomas Graham and, in the beginning of the twentieth century, both by the Göttingen School of chemists and a generation of Swedish scientists. Indirect methods determined its size as being in the 30 to 2000 Ångström diameter range. The development of electron microscopic observation permitted a quantitaelectron microphotographs is the one taken by Drs. S. Mudd and J. Hillier of the attack of a bacterium by a bacteriophage. The electron microscope has shown that this virus particle was not a simple sphere but that it had a tail. When the virus attacks the bacterium it plunges its tail into the outer membrane of the bacterium. Material is transferred from the virus to the bacterial cell. Radiochemical evidence using radio-

FIG. 6. Bacteriophage attacking bacteria. Electronmicrograph of J. Hillier and S. Mudd. Magnification 33,333×.

phosphorus and radiosulfur showed that it was a radiophosphorus that was transferred. The presumption is that the nucleoprotein material is transferred. After this attack by the bacteriophage, the bacterium does not divide into new bacteria but explodes into a swarm of new virus particles. Biologically, this process represents infection and death of the bacteria. It is, however, a death through transfiguration. Chemically, it is merely a crystallization process. The nucleoprotein material in the bacterial cell, after nucleation by the virus, becomes a constituent of virus particles. The building blocks for this crystallization and particle for-

mation process, the nucleoproteins, are complex in structure. Let us turn to a much simpler process of particle formation, the formation of small gold particles from the gold ions present in aqueous solution.

Colloidal Gold—Formation

Colloidal gold has been the object of admiration and study since the dawn of chemistry as a science. It was the *aurum potabile* of the alchemist. It was used as an ingredient of the *Goldwasser of Danzig* and its ruby color

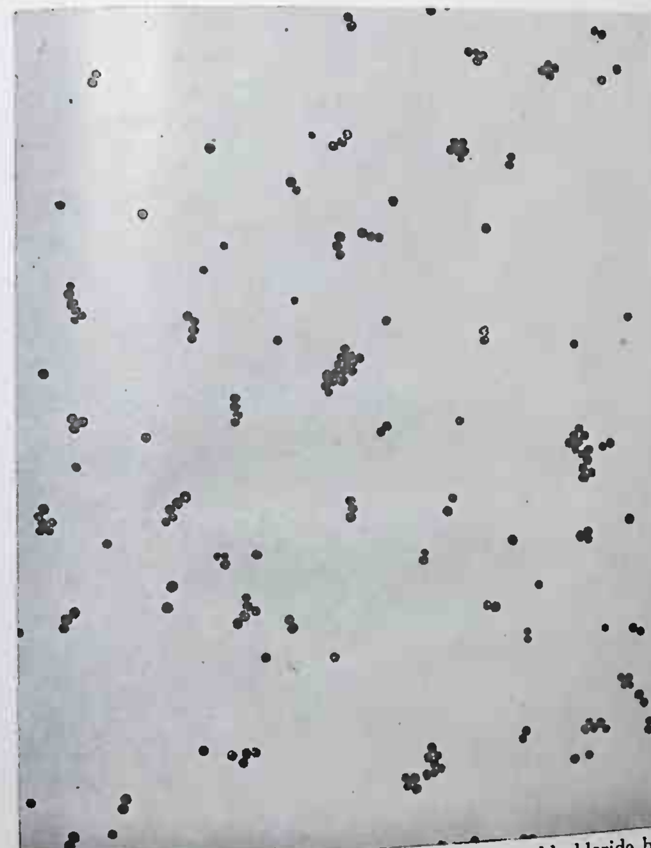


FIG. 7. Colloidal gold particles obtained by reduction of gold chloride by sodium citrate. Magnification 50,000×.

has been admired by all who have an appreciation for beautiful color. Its composition and properties were studied in the middle of the nineteenth century by Michael Faraday and Thomas Graham and, in the beginning of the twentieth century, both by the Göttingen School of chemists and a generation of Swedish scientists. Indirect methods determined its size as being in the 30 to 2000 Ångström diameter range. The development of electron microscopic observation permitted a quantitative and detailed study of colloidal gold systems. Peter C. Stevenson and the author examined various preparations of colloidal gold. This electron microscope survey showed that the gold particles, depending on the type of preparation, had a variety of shapes—spheres, plates and irregularly twisted rods. The preparation made by the reduction of gold chloride with sodium citrate at 100°C. was unusual in that all the particles were uniform spheres of two hundred Ångströms diameter. Measurement of a number of particles permitted a quantitative definition of uniformity; the root mean square deviation was twelve per cent. Further information about this system of "uniform" particles was obtained from the distribution in size curve. It was found that this curve was skew in shape with a slow rise in the number of particles as the diameter is increased until a maximum number was reached, and then a rather rapid drop to the maximum diameter observed. The distribution curve was not an error curve: it had character to it and this character expressed itself in the change of the shape of the curve with changes in the variables of the conditions of preparation—concentration, dilution and temperature. Because of the rather uniform distribution of particle size in this

FIG. 8. Gold plates and spheres obtained by reduction of gold chloride by citric acid 25,000×.

colloidal gold preparation, it was of interest to determine the factors that determined the size and the uniformity of these particles. The process of particle formation can be broken up into the following steps, as was shown in the discussion of the aerosols: nucleation, growth and aggregation. If the nucleation takes place rapidly, stops, and then is followed by growth without aggregation, the particles will be of uniform size. If, on the other hand, the nucleation, growth and aggregation take place simultaneously, there will be in the resultant particles a large distribution in size and a great variety in the shape. To study the complex process of particle formation it was desirable to isolate one of these processes from the other two. A review of the experimental results obtained by the Göttingen school in several decades uncovered a reaction in which growth is the only process. If a gold chloride solution is mixed with one of hydroxylamine hydrochloride in a scrupulously clean glass vessel, no reaction will take place at room temperature for several hours. If one adds to this solution some colloidal gold, reduction of the gold chloride takes place very rapidly. The colloidal gold particles act as nuclei and catalyze the reduction. The gold chloride-hydroxylamine solution, which by itself does not produce gold particles, can be used in several ways to study the formation of fine particles of gold. In the first place, this growth medium offers an experimental test for the presence of a nucleus. Any solution which, when added to the growth medium produces in it colloidal gold, must have nuclei present in it. The second use of the gold chloride-hydroxylamine growth solution is to grow uniform particles of one size to those of a larger size. For, if growth takes place only on gold particles and these are of the same size, one ought to build up larger particles of uniform size from the smaller ones. The diameter D of the larger size is related to the diameter d of the original size by the relation

$$D = d\sqrt[3]{\frac{a+b}{a}}$$

where a is the weight of the gold in the starting particle and b is the weight of the gold in the gold chloride solution. If one wishes to grow a particle double the diameter of the one in hand, one takes one part by weight of the gold in the original gold particles and seven times that amount in the gold chloride. Experimental work showed that one can grow uniformly sized particles from a diameter of 30 to 1600 Å. The availability of a set of monodispersed gold particles permits the study of the variation of properties of fine metallic particles with size. The red color of colloidal gold solution is due to a narrow absorption band at 520 millimicrons in the spectrum. The color changes to violet brown as the diameter increases. Gustav Mie, in 1907, calculated, using the classical Maxwell equations, how the absorption spectrum of colloidal gold varied

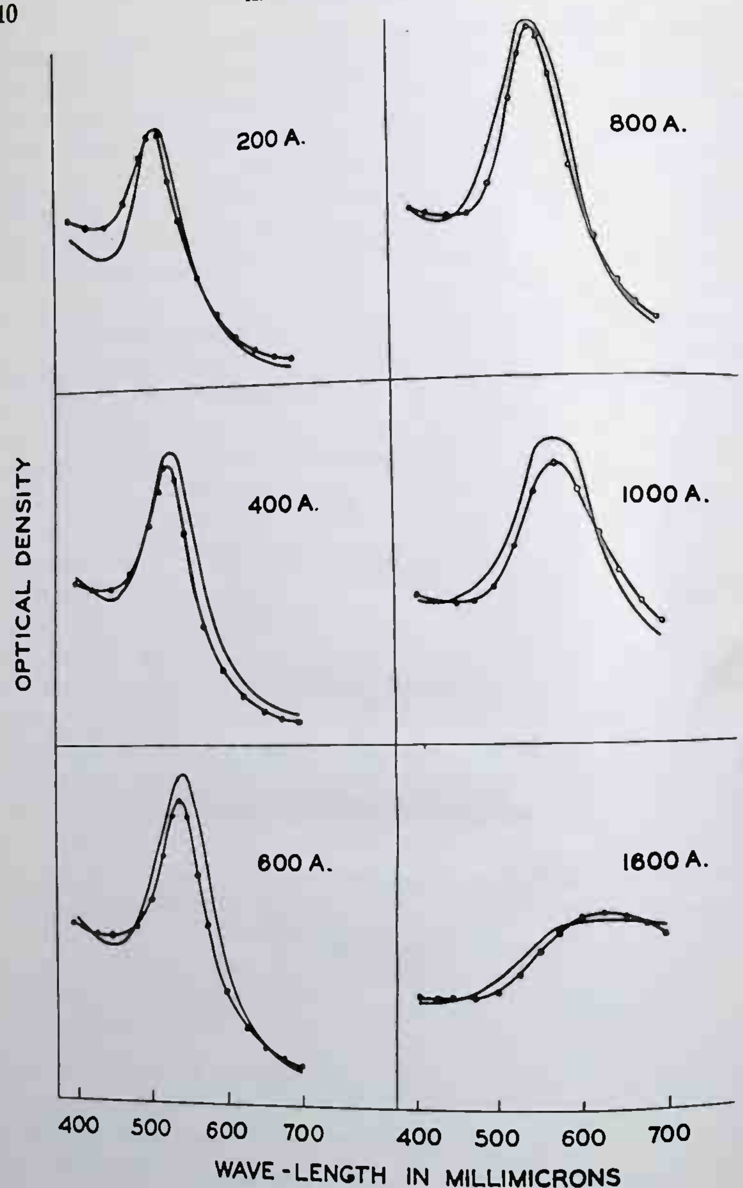


FIG. 9. Absorption spectrum of colloidal gold as a function of the diameter of the particle. The points are experimental while the full curve is the absorption spectrum predicted by G. Mie.

with the diameter of the particle. Garton measured the absorption spectrum of the monodisperse gold sols and found that the experimental results were exactly those that Mie predicted almost fifty years before.

Low angle scattering of X-rays has been used to determine the size of fine particles. Concordance was found between the diameter obtained by this method on the one hand and the diameter obtained by direct electron microscopic observation, using the monodisperse two hundred Ångström sol.

The question may arise as to how monodisperse, or uniform, are the grown particles. Measurement of the size distribution of the various sets of grown particles showed that the root mean square deviation remained the same, at about twelve per cent, while the absolute spread in size increased as one increased the mean diameter. A law of growth that follows from this observation is that the rate of growth of the diameter is proportional to the diameter. The larger the particle the faster it grows. Diffusion of hydroxylamine hydrochloride or of gold chloride to the gold particle cannot be the rate determining process, for then the rate of growth would be inversely proportional to the diameter and the particles would become more narrow in distribution as they grew larger. Surface reaction also cannot be the rate determining step for the process of growth, since the width of the distribution curve would remain unchanged as the particles grew. The following mechanism is proposed to explain the observed law of growth. The rate of the reduction of gold chloride to gold on the surface of the gold particle is limited by the ability of the particle to dissipate the energy liberated in the reaction on the surface to the bulk of the particle. It must be recalled that, because of the metallic character of the particle, there is very little direct coupling of the particle to the aqueous medium in which it is dispersed. The larger the gold particle, the more degrees of freedom are available for absorbing the energy generated at the surface and ultimately dissipating it by means of the lattice vibrations. Thus, the larger the particle the faster will be the surface reaction, which increases the surface by an exothermic process. This concept of the coupling of the surface vibrations and the lattice vibrations may be used to interpret the role of heterogeneous catalysts. In order to have an efficient catalyst, the particle must be so small that little of the energy liberated on the surface by the exothermic steps will be degraded into lattice vibrations. Most of it must be stored on the surface to become available for the endothermic steps in the catalytic process.

Whatever the explanation and implications of this strange law of growth, the fact still remains that the larger the particle the faster it grows. The law of growth so stated has an important corollary: The smaller the particle the slower it grows. It follows that there must be particles so small that to all intents and purposes they do not grow at all. Particles could not be formed from this mechanism of growth. There must be another process whereby particles are formed to such a size that they grow and this process is the process of nucleation.

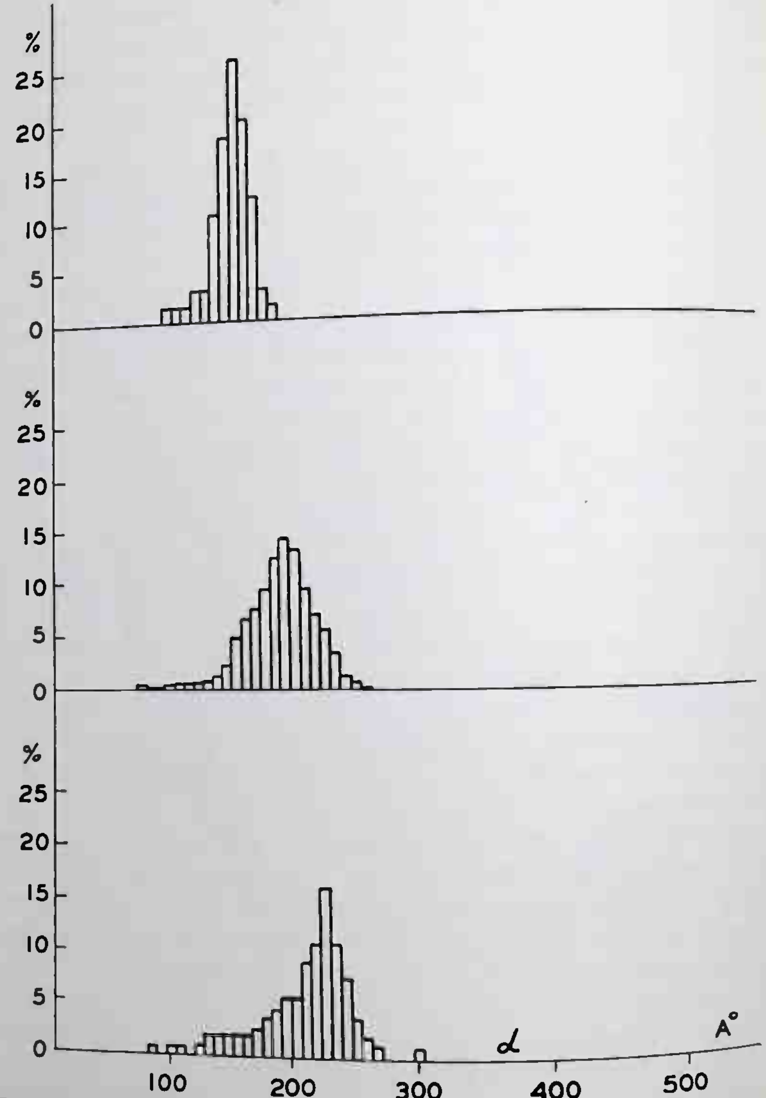


Fig. 10A. Particle size distribution curve of colloidal gold prepared at different concentrations.

Colloidal Gold—Nucleation

Information on the nucleation process was obtained by studying the size distribution curves of the gold particles obtained under different preparative conditions. The largest particle in a given particle size distribution is large because it was the first to nucleate and had a long time to grow before the growth medium, gold chloride, was exhausted. The smallest particle in a given distribution of sizes is small because it was

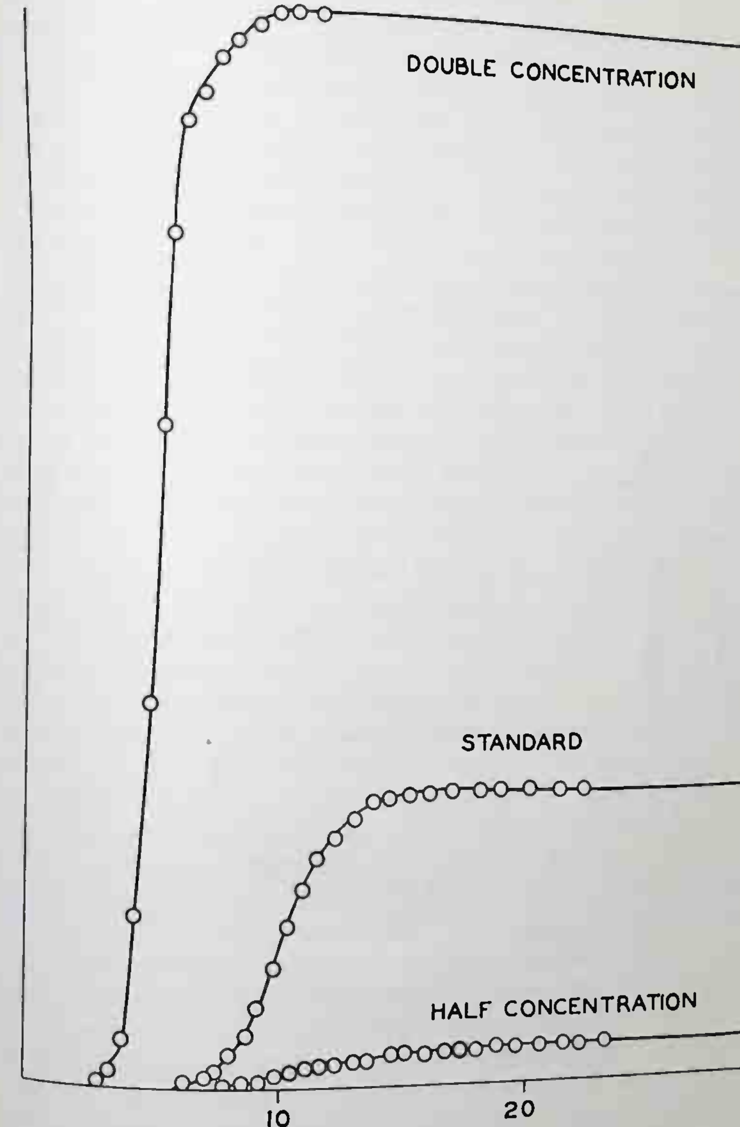


Fig. 10B. Nucleation curve obtained from the distribution curves. (Ordinate number of nuclei, abscissa—time in minutes.)

the last to nucleate, and, because of the law of growth, it could not compete effectively with its "older brethren." Thus, there is a generic relationship between the size distribution curve representing the number of particles as a function of size and the nucleation curve which represents the number of particles as a function of time. The distribution curve can be used to study the kinetics of the nucleation process. Knowing the law of growth, a mathematical procedure was developed for converting a distribution curve into a nucleation curve. In this way, a number of nucleation curves were obtained for various conditions of concentration, dilution and temperature of the two reactants, sodium citrate and gold chloride. These curves, while differing among themselves, have this general appearance—an induction period during which no nuclei formed, an autocatalytic period when the rate of nuclei formation accelerates with time and then a slowing-up period of nuclei formation. We were presented with a baffling situation—the nucleation had stopped, though a large excess of the two reagents that had caused it, was still present. An examination of the kinetics of the induction period showed that it was due to a chemical reaction with an activation energy. Examination of the literature of the oxidation of sodium citrate disclosed the possibility of an intermediate, the sodium salt of acetone dicarboxylic acid. The latter was therefore synthesized and made to react with gold chloride. Colloidal solution was readily formed. The particle size distribution curve was translated into a nucleation curve. The induction period had disappeared. Instead, the data indicated that the formation of nuclei was due to a decomposition of a precursor of the nucleus and that this decomposition followed a unimolecular law. The following mechanism was proposed for the nucleation process using sodium citrate as the reducing agent for gold chloride. The sodium citrate is first oxidized to the sodium salt of acetone dicarboxylic acid. The latter forms a complex with the gold chloride and this complex polymerizes to form a giant polymer. When this polymer reaches a certain critical size, it decomposes, according to a unimolecular law, to form gold nuclei and the carbonate ion. The polymer of the critical size must be such that it contains a sufficient number of gold atoms in it so that, after decomposition, a stable gold particle is formed. The particle must be stable in the sense that its lattice energy is greater than the disruptive energy due to surface tension. The particle must be large enough so that growth will be appreciable. To check this mechanism, sodium citrate-gold chloride solution was passed through an ion exchange resin during its induction state. The resin removed the gold ions. The solution obtained was found to be able to nucleate the growth medium. On examination in the electron microscope, the percursor was shown to consist of diffuse bodies of about two hundred Ångströms diameter which decomposed under the bombardment of the electron beam to form small gold particles.

Colloidal Gold—Aggregation

The last process in the story of the origin, growth and disappearance of fine particles is aggregation. In this process, the fine particles coagulate to form particles of such size that many of the properties, characteristic of matter in the finally divided state, disappear. Coagulation was

again studied with colloidal gold of uniform size. The number of collisions that such gold particles undergo with each other at ordinary dilutions is several thousand per second, yet, in spite of this, the particles do not coagulate. The reason for the stability is the existence of repulsive electrical forces whose strength is greater than the attractive van der Waals forces. The repulsive electrical forces are caused by the diffuse electric double layer that surrounds each gold particle. This double layer is due, on the one hand, to a negative charge firmly fixed to the gold particle and, on the other hand, to the counter-charge of the positive ions. The over-all

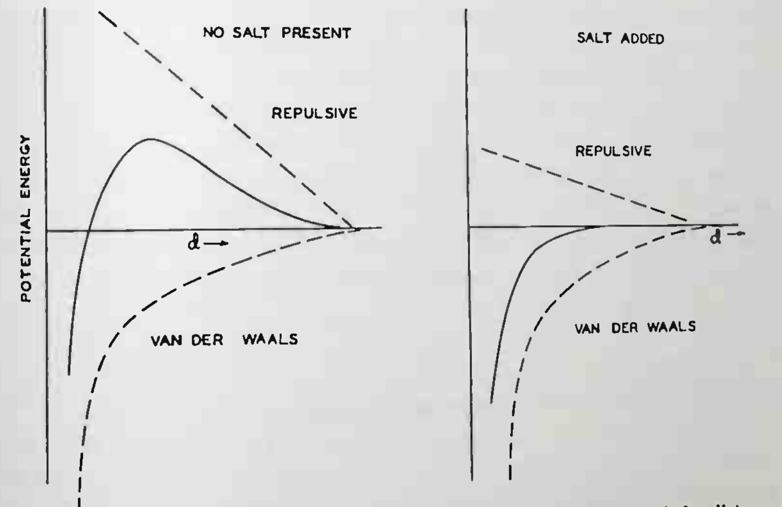


Fig. 11. The potential energy of two gold particles as a function of the distance between them (schematic). The dashed curves represent the two components of this potential—the repulsive of the diffuse double layer and the attractive van der Waals potential.

charge is zero. The ionic atmosphere of the diffuse positive layer expands and contracts as one decreases and increases the salt content of the solution. In very dilute solution, the ion atmosphere has a very large radius and the gold particles cannot come close to each other during collision. On the other hand, in solutions of high salt content the ionic atmosphere contracts sufficiently for the van der Waals forces to come into play at the small separation between the colliding gold particles. The formation of doublets, triplets, etc., then takes place. The situation is presented graphically in terms of the potential energy of the system. The attractive and repulsive potential forces are sketched as a function of separation between the two particles. The van der Waals potential is independent of the salt content of the solution, while the repulsive potential between the two diffuse layers depends strongly on the salt content. The resulting barrier between the two colliding particles of gold can be controlled at will, from very large values at very dilute solution to zero values for solutions containing appropriate amounts of dissolved salts.

The mathematical expression for this potential energy between the two gold particles with their ionic atmosphere has been given by Verwey and Overbeek.

$$E = \frac{64kT\gamma^2 e^{-\kappa d}}{\kappa} - \frac{A}{4\pi d^2}$$

where k is the Boltzmann constant, T the absolute temperature, γ a constant close to 1, e the base of natural logarithms, d the separation between the two particles, and A the van der Waals constant. κ is the measure of the ionic strength of the solution equal to $3 \times 10^7 \Sigma Z_i\sqrt{c_i}$ where c_i is the concentration of the ith species of ion and Z_i its charge. This can be checked by determining the critical concentration of added salt which will just produce coagulation as indicated by change of color of the system from red to violet. According to the Verwey-Overbeek equation, this will occur not only when the potential energy is zero but when its slope is also zero. The expression for the critical concentration, C, of added ions that will just produce coagulation is then given by,

$$C = \frac{8 \times 10^{-22}}{A^2Z^6}$$

where Z is the charge on the ion causing coagulation.

One should note the strong dependence on the charge of the added positive ion. Schultze-Hardy (1882) had noted that the coagulating action depended strongly on the charge of the added positive ion. The Verwey-Overbeek equations have given it a quantitative formulation. Baker and the author studied the critical concentration of the perchlorates of Na^+, Mg^{++}, Al^{+++} and Th^{4+} ions and established experimentally that the critical concentration varies inversely as the sixth power of the charge. They also were able to calculate the van der Waals constant A of gold as 10^{-12}. This compares favorably with the values calculated quantum mechanically and is easier to determine than was done by Overbeek who measured the force between two highly polished plates, or by Deryagin who measured the force between a plane and a spherical surface.

Colloidal Gold—Kinetics of Coagulation

We can now turn to another aspect of the coagulation process—its kinetics. Smoluchowski, in 1916, calculated the rate at which particles coagulate—in other words, how fast the singlets disappear, how fast the doublets, triplets, quadruplets, etc., increase and then decrease in number. Baker and the author were able to verify the Smoluchowski expressions by allowing a dilute solution of uniform 400 Å. to coagulate by adding a sufficiently large concentration of salt to collapse the ionic atmosphere and make every collision lead to aggregation. It was found that the coagulation process could be stopped at any time by the addition of gelatin. Samples could be obtained at different times of coagulation and mounted for electron microscope examination without changing the extent of coagulation in the process of mounting. The results given in

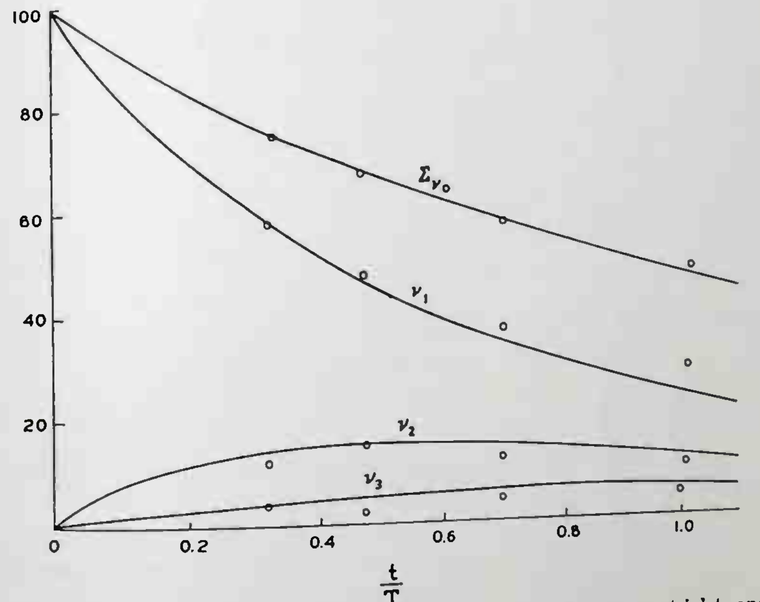


FIG. 12. Variation of the number of cluster: ν_1 singlets, ν_2 doublets, ν_3 triplets and $\Sigma\,\nu_0$ total number of clusters as a function of the reduced time.

Figure 13 present the number of clusters ν_k of size k as a function of reduced time t/τ where

$$\tau = \frac{3\eta}{4KT\nu_0}$$

and η is the viscosity of water and ν_0 the concentration of singlets in the original dispersion.

Other more stringent experimental tests were applied to the Smoluchowski theory. The coagulation rate can be calculated in a rigorous way with no adjustable parameters and these calculations can be checked experimentally.

Thus, the three processes of nucleation, growth and aggregation of fine particles in solution can be placed on a quantitative basis obeying, for the most part, generally accepted physical and chemical concepts and laws.

The account that has been given represents the view point of one group of scientists looking at the world of fine particles. Many beautiful particles were seen when this world was examined with the electron microscope. Strange phenomena were observed and made more understandable. Other scientists in different parts of the civilized earth are also looking at this world of fine particles. We are sure that they are also

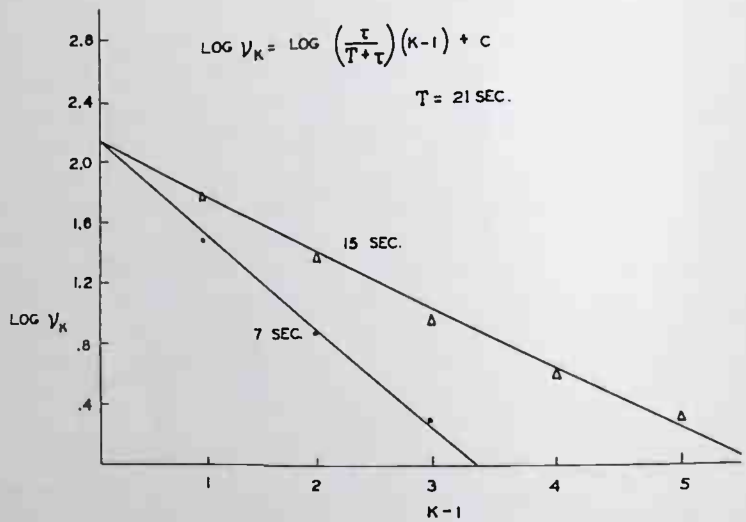


FIG. 13. Experimental verification of Smoulchowski relation between the number of clusters ν_k containing k particles, and the number of k particles in a cluster, for two times during the process of aggregation.

seeing many interesting species and phenomena. To us, as to them, there is no question that there are many more interesting studies to be carried out in the world of fine particles.

Acknowledgment

Part of this investigation has been carried out with the financial support of the U.S. Atomic Energy Commission. The author wishes to express his appreciation to Dr. J. Hillier for his introduction to the field of electron microscopy, and to Dr. Stuart Mudd for permission to use the electronmicrograph on the attack of the bacteria by bacteriophage particles.

BIBLIOGRAPHY

1. MENTER, J. W. *Proc. Roy. Soc. A236*, 119 (1956).
2. VON ARDENNE, M. Uebermikroscopie, Berlin 1940. BURTON, F. F., and KOHL, W. H. The Electron Microscope, 2 ed., Reinhold Publishing Co., New York, 1946. ZWORYKIN, V. K., MORTON, G. A., RAMBERG, E. G., HILLIER, J., and VANCE, A. W. Electron Optics and the Electron Microscope, Wiley and Sons, New York, 1945. HALL, C. E. Introduction to Electron Microscopy, McGraw-

Hill Book Co., New York, 1953. COSSLETT, V. E. Practical Electron Microscopy, Academic Press, New York, 1951. WYKOFF, R. W. G. Electron Microscopy, Technique and Application, Interscicnce Press, New York, 1949. ZWORYKIN, V. K. Image Formation by Electrons, Science in Progress, 3rd series, Yale University Press, New Haven, 1942, 69–107. WYCKOFF, R. W. G. The Macromolecular Structure of Biological Material, in Science in Progress, 7th series, Yale University, New Haven, 1951, 203–235.
3. AMICK, J., and TURKEVICH, J. Electron Microscopy and Diffraction of Aerosols Formed in a Direct Current Arc. U.S. Atomic Energy Commission Report NYO-239. Further Studies in Electron Diffraction of Aerosols Formed by a Direct Current Arc. U.S. Atomic Energy Commission Report NYO-3431.
4. TILL, P. H., and TURKEVICH, J. Metallic Smokes, U.S. Atomic Energy Commission Report NYO-3435.
5. FRANK, F. C. *Discussions of the Faraday Society*, No. 11, Crystal Growth, 48 (1949).
6. STREZNEWSKI, T. J., and TURKEVICH, J. "La reaction de l'oxygène avec des particles de carbon," *Revue de l'Institut Francais du Petrole, 13*, 686 (1958).
7. TURKEVICH, J., and HILLIER, J. Electron Microscopy of Colloidal Systems. *Anal. Chem., 21*, 475–485 (1949).
8. TURKEVICH, J., HILLIER, J., and STEVENSON, P. C. Discussions of the Faraday Society No. 11. The Size and Shape Factor in Colloidal Systems, 55–74 (1951).
9. TURKEVICH, J., GARTON, G., and STEVENSON, P. C. The Color of Colloidal Gold, *J. Colloid Sci.*, Supplement *1*. 1954, 26–35.
10. HUBBELL, H. H., and TURKEVICH, J. Low Angle X-ray Diffraction of Colloidal Gold and Carbon Black, *J. Am. Chem. Soc., 73*, 1 (1951).
11. VERWEY, E. J. W., and OVERBEEK, J. T. G. Theory of Stability of Lyophobic Colloids, Amsterdam, 1948.
12. BAKER, C., and TURKEVICH, J. Coagulation of Colloidal Gold, U.S. Atomic Energy Commission Report NYO-3436.
13. SCHULTZE, H. *J. prakt. chem., 27*, 320 (1883).
HARDY, W. B. *Proc. Roy. Soc. London, 66*, 110 (1900).
14. OVERBEEK, J. T. G. *Disc. Faraday Soc., 18*, 12–24 (1954). DERYAGIN, B. A. *Quart. Rev.* (London) *10:* 295–329 (1956).

THE SOCIETY OF THE SIGMA XI

REPORT OF THE TREASURER FOR 1958

Schedule 1. Balance Sheet as of December 31, 1958, with 1957 Figures for Comparison

Assets

		1958	1957
Bank Balances:			
Union and New Haven Trust Co. (Checking Account)		$ 47,818.85	$ 25,767.24
Connecticut Savings Bank		20,000.00	20,000.00
New Haven Savings Bank		18,069.29	18,069.29
First Federal Savings and Loan		10,000.00	10,000.00
National Savings Bank		10,000.00	10,000.00
Union and New Haven Trust Co. (Berg Fund)		2,841.23	2,481.22
Investments (Schedule 4)		132,343.24	132,343.24
Total Assets		$241,072.61	$218,660.99

Liabilities

		1958	1957
Semi-Centennial Fund		$ 15,050.00	$ 15,050.00
Berg Memorial Fund		2,841.23	2,481.22
William Procter Memorial Fund		22,624.55	22,624.55
Ada P. McCormick Fund:			
January 1, 1958	$ 2,167.53		
Interest	80.00	2,247.53	2,167.53
Research Fund (Schedule 3)		32,705.41	34,466.31
Securities Reserve		741.95	741.95
Reserve for Royalties to Authors:			
Reserve on January 1, 1958	$ 1,823.78		
Received in 1958 for 1957	793.78	2,617.56	1,823.78
Reserve for Prepaid Assessments:			
Reserve on January 1, 1958	$ 4,650.00		
Net Decrease for 1958	1,048.00	3,602.00	4,650.00
Capital Reserve:			
Reserve on January 1, 1958	$84,544.23		
Transferred from Surplus	30,000.00	114,544.23	84,544.23
Surplus:			
Balance January 1, 1958	$50,111.42		
Transferred to Capital Reserve	30,000.00		
	20,111.42		
Operating Gain, 1958	23,986.73	44,098.15	50,111.42
Total Liabilities		$241,072.61	$218,660.99

Schedule 2. Operating Statement, January 1 to December 31, 1958

Receipts

		1958	1957	Budget 1958
Chapter Assessments		$ 64,997.50	$ 61,354.00	$ 64,000.00
Club Assessments		4,683.50	4,296.00	4,000.00
Membership-at-Large	$32,981.58			
From Reserve for Prepayments	1,048.00	34,029.58	23,640.31	20,000.00
Initiation Fees		6,886.50	6,510.22	6,000.00
Charters (Net)		252.00	100.00	...
Sales of Certificates (Net)		442.40	432.78	800.00
Sales of Insignia (Net)		1,862.73	2,700.20	3,000.00
Subscriptions to American Scientist		11,303.26	11,005.04	10,000.00
Carried Forward		124,457.47	110,038.55	107,800.00

	1958	1957	Budget 1958
Brought Forward	124,457.47	110,038.55	107,800.00
From RESA for American Scientist	3,500.00	2,000.00	3,500.00
Advertising (Net)	34,606.64	34,950.79	28,000.00
Sales of *Science in Progress* (Net)	(133.79)	(621.00)	500.00
Royalties, *Science in Progress*	793.78	518.31	...
Honoraria for Lecturers	11,002.88	8,308.00	8,000.00
Interest and Dividends (Net)	7,969.49	7,122.02	6,000.00
Miscellaneous	152.50	49.46	100.00
Total Receipts	$182,348.97	$162,366.13	$153,900.00

Disbursements

	1958	1957	Budget 1958
Headquarters:			
Salaries	$ 26,241.06	$ 24,957.95	$ 25,000.00
Supplies and Expenses	4,082.56	5,240.23	3,000.00
Postage and Mailing Expenses	1,905.15	1,399.21	1,500.00
Rent	2,448.00	2,448.00	2,700.00
American Scientist:			
Salaries	5,750.00	6,750.00	8,000.00
Printing and Mailing	76,097.70	66,657.47	70,000.00
Office Expenses	1,037.77	925.29	1,200.00
Honoraria	50.00	...	1,000.00
Rent	1,260.00	1,200.00	1,200.00
Lecture Program:			
Honoraria	7,290.00	5,750.00	7,000.00
Travel	8,464.37	2,768.13	5,000.00
Clerical	2,880.72	2,124.46	2,200.00
Royalties (To Reserve for Authors)	793.78	518.31	...
Officers' Travel and Expense	3,099.27	2,674.04	3,000.00
Membership Program	5,648.70	4,098.07	4,000.00
Science Service	200.00	200.00	200.00
Bank Charges	178.77	198.64	100.00
Refunds of Overpayments	...	54.00	...
New Equipment	455.65	649.26	500.00
Berg Memorial Fund	300.00	300.00	320.00
Interest to Berg Fund	60.01	20.65	...
Interest to Research Fund	602.00	602.00	600.00
Transferred to Research Fund	5,000.00	5,000.00	5,000.00
Social Security	610.85	550.53	600.00
Manuals of Procedure	1,168.77	584.28	800.00
Initiates' Booklets	1,119.69	1,154.40	1,200.00
IBM Machine Rental	792.00	742.00	750.00
Medical Service	315.80	182.40	200.00
Interest to McCormick Fund	80.00	83.00	...
Reprints	26.34	18.93	...
Miscellaneous	403.28	263.82	500.00
Total Disbursements	$158,362.24	$138,115.07	$145,570.00
Operating Gain	23,986.73	24,251.06	8,330.00
	$182,348.97	$162,366.13	$153,900.00

Schedule 3. Research Fund Operations

Balance, January 1, 1958			$34,466.31
Contributions, Membership-at-Large and Chapters	$17,737.10		
Interest on Semi-centennial Fund	602.00		
From 1958 Operating Account	5,000.00		
	$23,339.10		
Less Grants for Year	25,100.00		
Decrease in Fund for Year			1,760.90
Balance, December 31, 1958			$32,705.41

Schedule 4. Investments

(As of December 31, 1958. Carried at Cost)

Bonds:			
$ 4,000	American Machine and Foundry Co., 5%, Conv.	$4,335.60	
$ 5,000	American Tel. and Tel. Co. 3⅜%, 1973	5,229.38	
$ 5,000	American Tobacco Co., 3¼% 1977	5,079.99	
$ 5,000	Borden Company, 2⅞%, 1981	4,992.09	
$ 5,000	Boston Edison Co., 2¾%, 1970	5,090.23	
$ 2,000	Chicago Union Station, 3⅛%, 1963	2,060.28	
$ 5,000	Consumers Power Co., 3⅛%, 1981	5,114.19	
$ 4,000	General Telephone Co., 4%, 1971	4,171.50	
$ 3,000	Southern Bell T. and T. Co., 3%, 1979	3,273.75	
$ 7,000	U.S. Treasury, 2¼%, 1959–62, June 15	7,000.00	
$ 8,000	U.S. Treasury, 2¼%, 1959–62, December 15	8,187.50	
$ 3,000	U.S. Treasury, 3¼%, 1978–83	3,000.00	
$10,000	U.S. Treasury, 2¾%, September 1961	9,915.10	
	Total Bonds		$ 67,449.61
Stocks:			
50 shares	Allied Chemical & Dye Corp.	$3,813.53	
40 shares	American Airlines, 3½% Pfd.	3,743.95	
100 shares	American Cyanamid Co.	2,503.22	
60 shares	Chesapeake & Ohio R.R.	3,180.00	
102 shares	Commonwealth Edison Co.	4,241.37	
112 shares	Consumers Power Co.	4,158.17	
25 shares	Du Pont de Nemours Co.	2,437.92	
200 shares	El Paso Natural Gas Co.	2,582.87	
100 shares	General Foods Corp.	2,235.10	
100 shares	Marshall Field Co.	3,595.66	
100 shares	Niagara Mohawk Power Co.	2,806.56	
100 shares	Owens-Illinois Glass Co.	3,927.64	
40 shares	Pacific Lighting Co., 4¾% Pfd.	4,286.48	
100 shares	Procter and Gamble Co.	3,125.00	
155 shares	Standard Oil Co. (N.J.)	3,161.35	
102 shares	Texas Company	2,494.47	
30 shares	Texas Eastern Transmission Co., 4½% Pfd.	3,127.50	
100 shares	Tri-Continental, $2.70 Pfd.	5,300.35	
80 shares	Westinghouse Electric Corp.	3,688.49	
	Total Stocks		$ 64,409.63
	Uninvested Cash		484.00
	Total Investments		$132,343.24

Market Value, December 31, 1958, over $175,000.

AUDITOR'S STATEMENT

SIGMA XI

I have examined the books and supporting records of The Society of the Sigma Xi, Incorporated, for the year ended December 31, 1958.

All receipts recorded by the Secretary's and Treasurer's offices and income due from investments were traced to the Cash Account. Expenditures were verified against cancelled checks.

Cash balances were verified by reconcilement of bank statements and by direct confirmation with banks. Assets of the Investment Account were examined.

Payments for awards of Grants-in-Aid of Research were checked for proper authorization.

In my opinion, the records of the Society are properly maintained and in good order. I certify that the schedules of the report of the Treasurer are true to the best of my knowledge.

Respectfully submitted,
C. S. Lockrow
Auditor

New Haven, Connecticut
February 5, 1959

THE SCIENTIFIC RESEARCH SOCIETY OF AMERICA

REPORT OF THE TREASURER FOR 1958

Receipts

	1958	*1957*	*Budget 1958*
Assessments, Branches	$11,340.50	$ 8,558.30	$ 8,000.00
Dues, Membership-at-Large	1,753.50	944.00	800.00
Charter Fees	375.00	300.00	...
Insignia Sales (Net)	51.65	(6.94)	25.00
Dividends and Interest (Net)	5,190.35	5,148.08	4,500.00
Miscellaneous	29.00	40.00	10.00
	$18,740.00	$14,983.44	$13,335.00

Disbursements

Salaries	$ 5,711.92	$ 4,920.81	$ 5,500.00
Postage and Supplies	872.15	599.76	600.00
Rent	312.00	312.00	310.00
Officers' Travel and Expense	711.54	482.31	750.00
New Equipment	120.74	167.67	100.00
Certificates	514.65	374.38	350.00
Charters	330.00	385.00	...
Refunds of Overpayments	116.00	72.00	...
Social Security	89.68	90.80	200.00
Procter Prize	1,025.00	1,025.00	1,025.00
Sigma Xi, for American Scientist	3,500.00	2,000.00	3,500.00
Members' Pamphlet	...	207.90	50.00
Miscellaneous	42.54	47.00	100.00
	$13,346.22	$10,684.63	$12,485.00
Operating Gain	5,393.78	4,298.81	
	$18,740.00	$14,983.44	

Balance Sheet as of December 31, 1958, with 1957 Figures for Comparison

Assets

	1958	*1957*
Cash, Union and New Haven Trust Co.	$ 9,153.68	$ 5,362.20
Securities (at cost)	102,583.16	102,583.16
Total Assets	$111,736.84	$107,945.36

Liabilities

Research Fund:				
Balance, December 31, 1957		$ 199.75		
1958 Contributions		397.70	597.45	199.75
Reserve for Research			94,331.27	94.331.27
Procter Memorial Fund			13,000.00	8,000.00
Capital Reserve				
Surplus, December 31, 1957		$ 5,414.34		
Operating Gain for 1958		5,393.78		
		$10,808.12		
Special Appropriation for Pamphlet	$2,000.00			
Transferred to Capital Reserve	5,000.00	7,000.00		
Surplus, December 31, 1958			3,808.12	5,414.34
Total Liabilities			$111,736.84	$107,945.36

INVESTMENTS AS OF DECEMBER 31, 1958 (AT COST)

Bonds:		
$10,000 Pacific Gas and Electric Co. 1st & Ref. 3⅜% 1985	$10,302.60	
$10,000 Southern N. E. Telephone Co. Deb. 3¼% 1985	10,101.40	
$10,000 U.S. Treasury 2¾% September 1961	9,822.50	
Total Bonds		$ 30,226.50
Stocks:		
54 shares Allied Chemical and Dye Corp.	$ 3,714.55	
200 shares American Cyanamid Co.	5,813.70	
150 shares Associated Spring Co.	3,125.00	
228 shares Connecticut Light and Power Co.	3,532.12	
388 shares El Paso Natural Gas Co.	8,587.69	
102 shares National Lead Co.	3,226.65	
110 shares Ohio Edison Co.	3,853.69	
80 shares Owens-Illinois Glass Co.	3,010.95	
60 shares Charles Pfizer & Co., $4 Pfd.	6,765.65	
200 shares Procter and Gamble Co.	6,275.00	
155 shares Standard Oil Co. (N.J.)	4,208.63	
80 shares Stanley Works	3,180.00	
102 shares Texas Co.	2,904.65	
200 shares Texas Utilities Co.	3,845.54	
100 shares Torrington Co.	3,375.00	
50 shares United Fruit Co.	3,082.65	
100 shares Westinghouse Electric Corp.	3,754.28	
Total Stocks		$ 72,255.75
Uninvested Cash		100.91
		$102,583.16

Market Value, December 31, 1958, over $158,000.

AUDITOR'S STATEMENT

I have examined the books and supporting records of The Scientific Research Society of America for the year ended December 31, 1958.

The cash balance was verified by reconcilement of bank statements and direct confirmation with the bank. Expenditures were verified against paid bank checks. All securities were examined and income receivable was accounted for.

In my opinion, the records of the Society are properly maintained and in good order. I certify that the schedules of the report of the Treasurer are true to the best of my knowledge.

Respectfully submitted,
C. S. LOCKROW
Auditor

New Haven, Connecticut
February 5, 1959

HERE AND THERE

By the Board of Editors and the Membership of the Sigma XI-RESA Societies

The Editors have decided to inaugurate, in this new volume of the AMERICAN SCIENTIST, a feature dedicated to the publication of interesting quotations from both current and older books. In this way we hope to do more for our readers than is possible through the Scientists' Bookshelf, giving them some of the actual flavor of a book that a review cannot easily convey. The existence of this feature will offer to all the membership an opportunity to bring to the notice of fellow-members passages from their reading which they deem of particular interest. The Editors will be glad to receive suggested passages, typed double-spaced, with the name of the book, the author's name, the location of the quote (page number) and the publisher of the book. The Board of Editors will undertake to secure, if possible, from the publishers the necessary permission to publish such excerpts as are selected for use. The Editors submit the following excerpts from recent works as examples of what they have in mind. Reader reaction will be welcome.

Is belief or doubt the better path to discovery? In *Personal Knowledge*, Copyright 1958 by Michael Polanyi, by permission of the University of Chicago Press, is the following on page 277:

"There exists, accordingly, no valid heuristic maxim in natural science which would recommend either belief or doubt as a path to discovery. Some discoveries are prompted by the conviction that something is fundamentally lacking in the existing framework of science, others by the opposite feeling that there is far more implied in it than has yet been realized. The first conviction may be regarded as more sceptical than the second, but it is precisely the first which is more likely to be hampered by doubt—owing to excessive adherence to the existing orthodoxy of science."

On the choice of a research problem, Polanyi analyzes the matter thus (page 124):

"In choosing a problem, the investigator takes a decision fraught with risks. The task may be insoluble or just too difficult. In that case, his effort will be wasted and with it the effort of his collaborators, as well as the money spent on the whole project. But to play safe may be equally wasteful. Mediocre results are no adequate return for the employment of high gifts, and may not even repay the money spent on achieving them. So, the choice of a problem must not only anticipate something that is hidden and yet not inaccessible, but also assess the investigator's own ability (and those of his collaborators) against the anticipated hardness of the task, and make a reasonable guess as to whether the hoped for solution will be worth its price in terms of talent, labor and money. To form such estimates of the approximate feasibility of yet unknown prospective procedures, leading to unknown prospective results, is the day-to-day responsibility of anyone undertaking scientific or technical research. On such grounds as these he must even compare a number of different possible suggestions and select from them the most promising problem. Yet, I believe that experience shows such a performance to be possible and that it can be relied upon to function with a considerable degree of reliability."

Analysis and judgment in the solution of problems is the concern of J. R. Goldstein in a paper on "Scientific Aids to Decision Making," pages 53–62, in a recent book on *Operational Research in Practice*, edited by Max Davies and Michel Verhulst, published in 1958 by the Pergamon Press. From page 55 we read:

"Many problems, or some critical parts of some problems, just do not lend themselves to analytical solution. You have to use judgment. Well, what is judgment? Judgment is just experience. But here you run into a difficulty. Today most of our serious problems lie in areas where the factors involved are so technical and where the problem itself is so broad in scope that no one can honestly claim to have experience.

We must come back to the original idea of analyzing the problem. What does this mean? It means breaking the problem down into its components. There may be many parts. Some of them we can handle by analysis, using various scientific techniques. You may have to make a mathematical model of one part of the problem; you may have to use other devices to examine other parts. But some of the components may still defy analytic treatment. Here, because you have broken it down into narrower fields, it may be possible to find individuals who have direct, sound experience in these specialized fields. When this is the case, you can draw upon their individual judgment.

But be careful: We trust a man's intuition in a field in which he is expert. But in these cases we may be dealing with a field so broad no one can be called expert. No one's unsupported intuitions in such a field can be trusted.

Analysis should be looked upon not as the antithesis of judgment but as a framework that permits the judgment of experts in various subfields to be combined—to yield results which transcend any individual judgment. This is its aim and opportunity.

But we still have the question: Where is the "expert" in the field as a whole with the judgment required to design an analysis and interpret its results? We know there are not any real experts. But we think we can demonstrate that the degree of expertness required to design an analysis is less than the degree of expertness required to intuit a good answer without an analysis."

The following quotations are from *Man's World of Sound*, Copyright c 1958 by John R. Pierce and Edward E. Davis, Jr., reprinted by permission of Doubleday and Company, Inc.

On sounds that we hear, and the necessity for experts on crickets, we read (page 102):

"A number of years ago an acquaintance of mine who works at another industrial research laboratory noticed, as he drove through the New Jersey meadows in the evening, that the crickets all appeared to chirp in unison, and with the same pitch. Could he, he wondered, make an electronic supercricket and influence the chirping of this vast horde of insects?

The idea so intrigued him that he sought the help of two colleagues. He himself made an electronic device which produced, as nearly as he could tell, very authentic-sounding cricket chirps of variable pitch and variable repetition rate. His colleagues constructed a powerful portable amplifier and speaker system. One summer evening they loaded all this electronic gear into a station wagon, took it to a suitable location along the highway, pulled off the road and started to unload the equipment and set it up. A suspicious state trooper soon stopped by and asked what they were doing. The three scientists explained. This was in the early days of the war and the trooper was naturally suspicious. He told them to pack up and get going.

The three did not give up their project, however. By some means they succeeded in getting the gear set up in a favorable location at a favorable time. They twiddled the dials in vain, emitting giant cricket chirps at various rates and pitches. The crickets went happily on at their own rate and pitch, paying no attention to electronic chirps, large or small.

The three scientists, baffled, loaded up their apparatus and went home. The chief investigator then sought out the name of a world authority on crickets and wrote the whole sad story to this English cricket expert. The expert politely informed him that while crickets do emit a sound audible to human ears, there is also an ultrasonic component to their stridulation, a sound too high in pitch for our hearing. It is only this ultrasonic vibration that the crickets themselves hear. The experimenters had produced a tremendous noise to which the crickets were entirely deaf!

The moral of this story would appear to be that we hear only what we can hear, that there may be a great many obvious differences among sounds which must forever escape our ears."

On artificial speech, maybe of use in a dentist's chair, the following passage occurs on page 232:

"A wonderfully ingenious form of artificial speech was demonstrated by that inventive Englishman, Sir Richard Paget, in the 1920's and '30s. His device, which he dubbed the "cheirophone" contained no mechanical parts at all. It consisted merely of his hands and a sound source. His description is as follows: "... if the two hands be clasped at right angles to one another, so as to enclose the largest possible volume and the second, third, and fourth fingers of one hand be withdrawn inside the cavity, but still held side by side, they will divide the cavity into two parts. The double cavity thus formed may be considered analogous to the human mouth and the three fingers to the tongue; movement of the fingers toward the palm of the hand to which they belong is similar to drawing the tongue backward in the mouth. The palm of the other hand becomes the hard palate. An artificial larynx was inserted between the thumb and first finger of the tongue hand, while the thumb and first finger of the palate hand operated as lips. . . .

With this primitive arrangement Paget could produce most of the English speech sounds; only the plosives were somewhat elusive. He could produce various sentences, among them: "Hello, London, are you there?" and "Oh, Leila, I love you." He even developed a special sentence for use when the vocal organs are immobilized in the dentist's chair: "Easy there, you're on the nerve." Paget admitted that considerable practice was required to become proficient with the technique. However, anyone can make a few vowel sounds by experimenting with various tongue positions, supplying the vocal energy by means of a Bronx cheer rendered into the appropriate opening."

The amateur scientist seems still to have a place amid governmental, university and industrial laboratories (page 235):

"I have often heard people say that the day of the garret inventor and the amateur scientist is past. This may be so in general. It certainly *seems* unlikely that a lone worker can match the modern resources of governmental, university, and industrial laboratories. But the work of Paget and Von Kempelen shows that significant research can be done without the elaborate equipment of today. Their truly remarkable understanding and simulation of speech was achieved with only their auditory sense to guide them. Their mechanical tools were crude indeed. Even in Paget's day useful electric devices were practically non-existent. It seems to us that these men suffered under crushing handicaps. And yet recently I saw Svend Smith demonstrate how various voice qualities are related to particular motions of the vocal cords by means of models of the larynx cut from rubber with a scissors, glued, clamped and manually held together, and blown with a vacuum cleaner or his own lungs.

All of this makes it clear that scientific progress is a brain child. Knowledge must arise through individual thinking, and men can think in a garret as well as in a stainless-steel laboratory. Such thought, such science, cannot be produced by mechanical means, either animate or inanimate."

The necessity for hard work in research and the nature of an experiment, as contrasted with experience, are the subjects of two quotes from page 270:

"One thing that continually puzzles me on encountering people who dabble in science is that so many of them want to do big and important things with practically no expenditure of effort. The well-worn professional rapidly gets to the point at which he is extremely happy to find one little thing which is both new and true by exerting himself to the utmost. . . .

I think that non-scientists find it extremely hard to understand what an experiment is and how it differs from mere experience. We all have experiences; life is made up of them. Sometimes they are thrust upon us, and sometimes we ourselves seek them out. We all learn something through experience; that is how we learn to walk and talk and to live in the world about us. Some of us learn more from our experience than others. A few seem to learn very little; they repeat the same mistakes day after day and year after year. Wise and observant men learn a good deal; that is how they become successful parents, or physicians, or bosses.

The knowledge of the world which we gain from experience is very real and very useful, but it is often fallible and uncertain in its application, and it is rarely knowledge which can be formulated accurately, codified neatly, and made useful to another. How well parents and bosses know this! The knowledge that we gain from experiments is simpler, and often less useful. The wonderful thing about a good experiment, however, is that the result can be accurately formulated and it can be conveyed to another, less experienced worker who can repeat the experiment, get the same result, and understand and apply it."

NEWS AND VIEWS

(Continued from page 48A)

Southern Tour:
Professor George C. Kennedy, UCLA
"Recent Studies at High Pressures and Implications Concerning the Crust of the Earth"

Plains Tour:
Professor Ralph Wetmore, Harvard University
"Morphogenesis in Plants—A New Approach"

Pacific Tour:
Dr. R. C. Elderfield, University of Michigan
"Australian Trees and High Blood Pressure"

3. Report on the Lectures for 1959–60: The lecture program for next year is now in the process of being arranged; a tentative list of several lecturers for each tour has been drawn up, largely from suggestions furnished to the Lectureship Committee by Chapters and Clubs of the Society. The details of this program will be worked out during the next few months under the supervision of a new Director of the Committee on National Lectureships.

In connection with the next program, I would like to report that the Executive Committee of the Society has approved two important changes recommended by the Lectureship Committee in the operation of the program. The number of requests for the lectures has continued to be so high that the Lecturers have had to give an average of 25 lectures each on their tours. Since most Lecturers are unwilling (or unable) to give more than a month or six weeks to the program, they have had to make only brief stops, usually an afternoon or evening, with each group on their tour. To remedy this situation the Executive Committee has approved an increase in the number of Lecturers from six to eight, starting with the 1959–60 program. The Sigma Xi Chapters and Clubs, and the RESA groups, have been divided into eight tours in accordance with the expanded program (one of these, by the way, now includes Hawaii and another includes Alaska). It should be possible for the Lecturers to visit somewhat longer with the host institution than has previously been feasible.

Another important change approved by the Executive Committee concerns the honorarium paid to the Lecturers. Ever since the program of Lectureships started, the honorarium has been set at \$50. per lecture. This is no longer a realistic figure in view of present income taxes and the general increase in honoraria and salaries paid in academic institutions and those concerned directly with research in applied science. The Executive Committee has approved the payment of an honorarium of \$75. per lecture, starting with the 1959–60 program; and has voted that the increase of \$25. per lecture will be covered in the budget of the National Society, instead of being an added charge to the local groups.

4. The Director of Lectureship Committee: Since I have presented to the Executive Committee my resignation as Director of the Lectureship Committee, effective at the conclusion of this Convention, the Executive Committee has appointed a new Director. I am pleased to announce that this will be Professor Harvey Neville of Lehigh University. Dr. Neville, who has been on the Lectureship Committee for the past four years, is a physical chemist; in his additional capacity as Provost and Vice-President of Lehigh, he has wide interests and contacts with research activities in many fields of science throughout the country. We are indeed fortunate that Dr. Neville has accepted this appointment."

Upon the conclusion of Dr. Carpenter's report a rising vote of thanks and appreciation for his service as Director of National Lectureships was given by the assembled Convention

Grants-in-Aid of Research: Dr. Harlow Shapley, Chairman of the Grants-in-Aid of Research Committee, was absent—being "strike-stuck in Florida"—and the Secretary read a letter from him which emphasized the need for increased funds to carry on this important activity of the Society. Secretary Holme stated that the full report of awards was made in the September issue of the AMERICAN SCIENTIST and represented the all-time high in both number of grants made (82) and total amount awarded (\$26,225). In 1957 the total amount was \$18,053. The Committee hoped that in 1959 the total might exceed \$32,000, but this, of course, would depend upon the response of the membership.

President Boyd then announced to the Convention that he wished to change the order of business as called for on the agenda and called for,

General Business

Annual Convention: The Executive Secretary read the following motion of the 58th Annual Convention:

It was *VOTED that:*

> Motion for independent Convention be tabled until next Convention which will be held in 1958 jointly with the AAAS. Further, that the Executive Secretary attempt to obtain further expression of opinion by letter ballot prior to next Convention,

and reported that less than 50% of the chapters and clubs had replied but those which had were overwhelmingly in favor of continuing to meet with the AAAS (5 to 1).

There being no further action—the 1958 motion remained *TABLED.*

Criteria for Membership: Dr. Roy Whistler reported to the Convention that his Chapter—Purdue—was concerned with the criteria for election to membership in the Society and felt the need for some over-all national review. Upon Dr. Whistler's motion,

It was *VOTED that:*

> The President of the Society of the Sigma Xi appoint a special committee to review the rules for membership—in particular regard to any need for clearer definition of criteria for full membership and to report its findings and recommendations to the 60th Convention of the Society in 1959 for possible reference back to the National Executive Committee.

Proposed Amendment to Constitution—Rochester: Dr. J. Lowell Orbison of the Rochester Chapter presented to the Convention the following proposed amendment to Article VII, Section 5(c) of the Constitution which had been considered by the Executive Committee at its October 1958 meeting, but not approved.

> In any chapter the classifications of chapter member and chapter associate member may include persons not connected with the institution if such procedure is approved by a vote of the local chapter members who are on the staff of the institution in which the chapter is located.

After considerable discussion the Chair ruled that the proposed amendment was out of order as notification of its consideration had not gone to the chapters sixty days prior to the Convention.

Committee on Nominations: In the absence of Dr. George R. Town of Iowa State College, Chairman of the Committee on Nominations—Dr. Roy L. Whistler presented the following report for that Committee:

The 1958 Nominating Committee, appointed by President George H. Boyd, to make nominations for consideration by the 59th Annual Convention of the Society at Washington, D.C., on December 29, 1958, has the honor to present the following candidates, each of whom has agreed to serve if elected:

As President for two-year term (elected in 1956): Frank M. Carpenter, Harvard University, Cambridge, Massachusetts.

As President-elect for two-year term: Donald B. Prentice, Yale University, New Haven, Connecticut.

As Executive Secretary for five-year

term: Thomas T. Holme, Yale University, New Haven, Connecticut.

As Treasurer for five-year term: Donald B. Prentice, Yale University, New Haven, Connecticut.

As Executive Committee for four-year terms: Harry F. Harlow, University of Wisconsin, Madison, Wisconsin; C. M. Jansky, Jr., Jansky and Bailey Incorporated, Washington, D.C.; *two-year terms:* Waldemar T. Ziegler, Georgia Institute of Technology, Atlanta, Georgia. (Replacing Hugh S. Taylor interim appointment of Executive Committee for Frederick E. Terman—resigned)

As Member of the Committee on Membership-at-Large for four-year term: John M. Clark, E. I. duPont de Nemours & Co., Wilmington, Delaware.

The Nominating Committee also recommends to the Executive Committee that an amendment to the Constitution be prepared for submission to the 1959 Convention under which provision would be made for creation of offices of Assistant Treasurer and Assistant Executive Secretary.

As Nominating Committee: Joseph W. Barker, Research Corporation; R. William Shaw, Cornell University; George R. Town, Iowa State College, Chairman.

It was *VOTED that:*

The report of the Committee on Nominations be accepted. The Chairman then asked for nominations from the floor. No other nominations were presented and on motion,

It was *VOTED that:*

The nominations be closed and the Secretary be instructed to cast a unanimous ballot for the officers and committee members named.

The Secretary reported that the ballot had been cast and the Chairman declared them elected.

Dr. Boyd then introduced President Carpenter who expressed his appreciation and, as his first official act, recommended that the Convention express its gratitude to Dr. George Boyd for his devoted service to the Society of the Sigma Xi both as a member of the Executive Committee and as its President. To this there was a unanimous rising vote.

President Carpenter announced that the evening address would be in cooperation with Phi Beta Kappa and Dr. James R. Killian, Jr., would speak on *Science and Public Policy*.

There being no further business matters to be presented to the Convention, the President declared the Convention adjourned at 5:05 P.M.

Respectfully submitted,
Thomas T. Holme,
Executive Secretary

Report of the Editor-in-Chief

During 1958 the four quarterly issues of the AMERICAN SCIENTIST contained a total of 424 pages of text and 372 pages of double-column material and advertising. The average number of pages per issue, 106 and 93 pages respectively, represented an increase of approximately 10 per cent per issue over the figures for the 5 issues of 1957, which was a special jubilee year. The December 1958 issue represented a 16 per cent increase in size above the March 1958 issue. For the coming year, it is anticipated that every number will be materially larger than 200 pages. The print order for the AMERICAN SCIENTIST has progressively increased during the last few years and now stands at 85,000 copies.

For an audience of that magnitude the Editors must report a negligible reader reaction, suggesting that scientists are among the more inarticulate of the intellectuals. Only occasionally do we get an adverse reaction to an article, and that is frequently matched by a letter commending us for the same publication. If one took the silence to mean consent to the editorial policy or tacit approval thereof, one must record that this would be in conflict with editorial judgment concerning the quarterly.

Frankly, there is an undercurrent of dissatisfaction among the Editors with respect to the AMERICAN SCIENTIST. Judging by manuscripts submitted, one would conclude that the scientists are intensely preoccupied with the relationship of science to education, art and poetry, truth, human behavior, extrasensory perception, and various other subjects which fringe experimental scientific research. Only now and then are the Editors receiving, with great pleasure, articles devoted to experimental science in progress, written not for the narrow specialist but for the 85,000 members who make up the recipients of the magazine. "There is great joy" in the editorial office when one of these articles turns up. It is assured a speedy insertion in the issue then underway. On the contrary, papers dealing with the interrelations of science with other aspects of truth have necessarily to be rationed in a journal devoted to the promotion of scientific research. As a consequence we have to turn back to authors many intensely interesting manuscripts in these fields because we do not have enough dealing with science in progress to keep a balance.

Our campaign a year ago to broaden the number of reviewers of books received must be pronounced a real success. Many more good reviews of books have come in than we could publish, and we are now sending out additional volumes to those who revealed their qualifications for such work.

In the coming year we are planning further opportunities for reader participation in our effort to make the AMERICAN SCIENTIST even more worthy of Sigma Xi-RESA support.—*Hugh Taylor*

Standing Committees

Each year the March issue will contain a complete listing, for purposes of reference, of the standing committees of both societies. For 1959, they are as follows:

NATIONAL OFFICERS—SIGMA XI

President—FRANK M. CARPENTER (1960), Harvard Biological Laboratories, Harvard University, Cambridge 38, Massachusetts
President-Elect—DONALD B. PRENTICE (1960) 56 Hillhouse Avenue, New Haven 11, Connecticut
Executive Secretary—THOMAS T. HOLME (1963), 56 Hillhouse Avenue, New Haven 11, Connecticut
Treasurer—DONALD B. PRENTICE (1963), 56 Hillhouse Avenue, New Haven 11, Connecticut

NATIONAL COMMITTEES—SIGMA XI

Executive Committee

The President, the President-Elect, the Executive Secretary, the Treasurer, the Chairman of the Committee on Membership-at-Large and the following:

JOHN S. NICHOLAS (1959), Yale University, New Haven, Connecticut
THORNTON C. FRY (1959), Remington-Rand, Stamford, Connecticut
WALDEMAR T. ZIEGLER (1960), Georgia Institute of Technology, Atlanta, Georgia
DETLEV BRONK (1960), Rockefeller Institute, New York, New York
HARVEY A. NEVILLE (1961), Lehigh University, Bethlehem, Pennsylvania
ROY L. WHISTLER (1961), Purdue University, Lafayette, Indiana
HARRY F. HARLOW (1962), University of Wisconsin, Madison, Wisconsin
C. M. JANSKY, JR. (1962), Jansky and Bailey Incorporated, Washington, D. C.

Committee on Membership-at-Large—Sigma Xi

JAMES H. MARKS (1959), *Chairman*, 811 Fisher Building, Detroit 2, Michigan
R. E. MCCONNELL (1960), Hobe Sound, Florida
WILLIAM M. MARKER (1961), 6240 Lindenhurst Avenue, Los Angeles 48, California
JOHN M. CLARK (1962), E. I. du Pont de Nemours and Company, Wilmington, Delaware

Committee on National Lectureships

HARVEY A. NEVILLE, Director, Lehigh University, Bethlehem, Pennsylvania
ELMER C. EASTON, Rutgers University, New Brunswick, New Jersey
SERGE KORFF, New York University, New York 53, New York
MARSHALL KAY, Columbia University, New York 27, New York
THOMAS T. HOLME, 56 Hillhouse Avenue, New Haven 11, Connecticut

LETTERS TO THE EDITORS

Dear Sirs:

Alfred Redfield's hypothesis, (American Scientist, Sept., 1958, page 205), concerning phosphorus regulation in sea water is very ingenious. I would like to take issue with one statement, however, that the nutrient level of the sea has been largely unchanged in the past.

During the Jurassic, the predominant type of fish had large teeth and well enameled scales. In addition, reptiles having much heavier teeth and bones than today's sea denizens, other than mammals, were fairly common.

During the Cretaceous, these types declined and by its close had disappeared. This phenomenon strongly suggests that the phosphorus/nitrogen ratio was once much greater than today. It is not likely that organs so extremely useful for some types of large animals as armored scales or large teeth or heavier bones would suffer so uniform and all-inclusive a decline unless some factor in the environment were deficient. This is especially so since weight is a small consideration under water. I suggest this factor was phosphorus. The presence of shellfish at the time rules out calcium, and there is no reason to suspect a vitamin.

While increasing amounts of nitrate, due to the advent of the leguminous plants, or some other factor, were possible, the actual deletion of phosphorus is even more probable. Considerable greensand was laid down, and at increasing rates, during the Cretaceous. These deposits were rather rich in phosphorus, especially at the close of the Cretaceous. While it is true that the ocean itself may not have altered much in macro-composition, and therefore pH, near shore where the greensands were deposited may have been a different matter. The increasing success of the rather acid oaks and pines may well have locally contributed enough acidity to bring the water close to the isoelectric point of ferric phosphate, or pH 5.0. The flocculation of ferric phosphate could, then, have had a considerable effect on the ocean's fertility, for these greensands were not a primary sedimentation.

The adsorption of phosphorus on primary sediments must have caused considerable variation also. The erosion of land increased toward the end of the Cretaceous, at least in the North Atlantic. This would have reinforced the deficiency, especially since increased erosion implies a larger percentage of subsoils being washed down.

If it is true that plants withdraw almost all the nitrogen and phosphorus from the water, it is only natural that the ratio of these elements in the plants would be almost the same as the ratio in the water, and probably was then.

Therefore, I contend that numerous mechanisms exist capable of having an effect on the ocean's fertility, and paleontology gives strong suggestive evidence that they were operative.

Sincerely yours,
Charles E. Weber
Mount Horeb Road
Warren Township
Plainfield, N.J.

Gentlemen:

The article, "On Scientific Russian: Its Study, and Translation," in your September 1958 issue, page 222, contains a plea for graduate schools to come down to earth and establish realistic foreign language requirements, along with an entertaining critique of present day scientific Russian style. One cannot but admire the author's erudition and eloquence.

Nevertheless, as a linguist, I feel a certain dissatisfaction with two points in Professor Wilde's article, one of them a matter of theory, the other of strategy. For the former, it seems to me that it is made insufficiently clear that a language, to paraphrase a great American linguist, Edward Sapir, is never to be identified with its lexicon (*Language;* N.Y.: Harcourt Brace, 1921; p. 234). This has become a commonplace among linguists, and Professor Wilde shows that he is not unaware of it by his criticism of "literal dictionary-guided translations" (222). This does not, however, prevent him from stating (although he goes on to criticize a too narrow knowledge of a language) that a book in a specialized field may call only for "an infinitesimally small familiarity with the language" (223) for translation. He is evidently here considering only vocabulary; it would indeed be difficult to find a book, no matter how specialized or how brief, which did not use the basic patterns of the grammar of the language in which it is written. And grammars are not transferable from language to language, as Professor Wilde shows he knows by his criticism cited above. Therefore, it also invites misunderstanding to say that "The student of science may well forego refinements . . . of grammar" (223). To understand the grammar of a language, how it modifies the meanings of words by inflection and adds meaning to them by the way they are joined, is the very heart of adequate translation.

As for the matter of strategy: Professor Wilde deserves thanks for making so explicit what, inexplicably, is known by so few, that current graduate school language requirements lead more often to inadequate knowledge of two (or more) languages than to a necessary minimum of understanding for any one. He would solve this problem by cutting requirements to a single language. This is a surprising conclusion in view of the fact that there are more languages of scientific importance today than ever before, as evidenced by Professor Wilde's own article, and modern nationalism makes it as certain as it can be that there will be more. How long will it be, for example, before Chinese is of major scientific importance? And I wonder if Professor Wilde fully appreciates how little Americans already study foreign languages: "a recent survey[1] showed that the United States ranks, in years of continuity of foreign language study, thirty-sixth among the nations of the world."

There is admittedly a real problem here. As Professor Wilde points out, the graduate student of today in the sciences must spend more and more time on his speciality, and more language seems just impossible to fit in. How then escape the dilemma? Negatively, it seems clear from the above that it would be a short-sighted approach to reduce requirements. The only course I can see to extricate us from the problem is for the graduate schools *not* to reduce requirements, but rather:

(1) Improve teaching methods. Most students of language agree that the quality of foreign language teaching in American schools today, due mainly to a laxity of requirements for teachers and a failure to take full advantage of what modern linguistic research has shown about the learning of languages, is not what it should be, and

(2) Strenuously insist that the percentage of *pre*-graduate school instruction devoted to foreign languages be increased. Linguists are convinced that foreign languages should be introduced earlier into the curriculum: the rote memory ability required for their learning is at its height during childhood and adolescence.

Because the problem is difficult does not mean it should not be attacked directly. Surely the answer to the American's amazing ignorance of foreign languages, in both depth and breadth, is not to reduce the number of languages studied. Americans have no less foreign language aptitude than anyone else, but they seem to show a remarkable inclination not to use what they have.

Sincerely,
KARL V. TEETER
1211 Dwight Way
Berkeley 2, California

[1] *The National Interest and Foreign Languages*, U.S. National Commission for UNESCO, revised edition, 1957. See the chart on p. 50.

Author's Note: I have no disagreement with Dr. Teeter, except that my suggestion to concentrate on a single language was intended for narrow specialists who have not sufficient time at their disposal to master a knowledge of two or more languages. If we discount rare exceptions, I believe my suggestion is a sound one, and I did not have in mind to extend my statement to professional linguists.

On the other hand, I have little background to argue about the technique of teaching languages: this is obviously a task of trained specialists.—S. A. Wilde

DEAR SIRS:

You anticipated that many scientists would question points raised in Dr. Kartman's article, AMERICAN SCIENTIST, Dec. 1958, page 282A, so I hope you will permit me to express a few differences of opinion.

No matter how brisk the business or how great the volume of Christmas sales, this feast can never be subverted, because the fact of birth of Our Savior cannot be altered.

Neither true religion nor true science is a product of human evolution. The truth has always existed and man has taken two different approaches to find it. There should be no internecine struggle because one section of truth cannot contradict another.

Science in its development has not attempted to destroy all metaphysical systems. If some scientists were so disposed, they certainly were not representative.

Religion is simply the relationship of man to his Creator. There is nothing inherently religious about researching, philosophical reflection or creative writing, even if in modern usage we speak of pursuing such things "religiously."

Religious practice certainly makes one more self-critical and better acquainted with his own limitations. An act of faith that would make one more egotistical and less humble would certainly be a strange one.

It is not "a distinct possibility that man can be as great and as good as Christ." There is no possibility that man with his "paltry plumbet may sound the depths of the Infinite." Certain mysteries will be clear to us only when we have attained eternal glory.

Nuclear scientists cannot be responsible for the misuses of atomic energy any more than the chemist who discovered potassium cyanide is responsible for all deaths due to this very effective poison. Scientists like others are responsible only for keeping their own houses in order. A scientist's relations with his fellowmen are important only in so far as they affect the order in his own house.

Yours very truly,
DESMOND D. DOLAN
186 LaFayette Ave.
Geneva, N.Y.

AND FROM THE AUTHOR WE HAVE:

My article, "Science and the Christmas Spirit," may be classified as a humanist position on the importance of the human spirit as the ground on which diverse views may unite. Essentially, my statement was a very personal one and it was offered as nothing more than a spontaneous expression. I certainly had no intention of being dogmatic or of implying any sense of finality to my interpretation of the meaning of Christmas. On the contrary, my views were given as a working proposition for myself and for anyone who might be interested. This came about because I cannot be satisfied with the mystical and final interpretations of the "faithful," and because I believe that ideas of what the truth is in regard to Christ, or any other subject, are the result of a dynamic process in the continual evolution of human thought.

The criticism of Dr. Dolan is very welcome, since, as I stated in the article, my concern is not with biblical history or with the formal facets of Christian tenets. I must confess that the form and content of the criticism tempts one to enter into a polemic on each point. Nevertheless, in spite of the fun this could afford, both in the exposition of biblical fact and in the exercise of logic, I do not think it would serve a useful purpose here.

The large number of appreciative letters which I have received from persons of various faiths and professional interests seem to me enough justification for having placed my feelings in the public view. I must admit that I did not anticipate so much active interest and that I was pleased by the many favorable comments. On the other hand, I not only expected disagreement, but welcomed it as a catalyst for the development of new and more adequate beliefs.

LEO KARTMAN

GENTLEMEN:

I note with interest and considerable approval the article in Perspectives in the December 1958 issue on "The Naturalistic Conception of Life" by Kendon Smith (page 413). But, as do most writers of letters to the editor, I have a bone to pick with Professor Smith. That bone carries the flesh (or is it ectoplasm?) of parapsychology.

However, it is not the question of the existence or nonexistence of parapsychological phenomena which I wish to debate. This has been ably discussed by R. A. McConnell in The AMERICAN SCIENTIST in March 1957. Neither am I concerned with Professor Smith's rather obvious and possibly unscientific *desire* that ESP be an experimental mistake rather than a fact. Rather, the opinion with which I take issue is the tacit assumption that ESP, if it exists, is somehow "supernatural" and vitalistic.

The supernatural of any era is simply the body of observations apparently unexplained by or unrelated to the science of that era. The supernatural of any era often becomes the commonplace of the next era. The phenomena of nuclear physics were supernatural before the Curies and the Rutherfords and the Einsteins. After examining the evidence, I have come to the tentative conclusion that ESP is real. However this in no way impels me to abandon the naturalistic point of view which I share wholeheartedly with Professor Smith. Instead, these experiments in ESP seem to indicate the existence of an entirely new body of phenomena to be studied, classified, theorized about, and added to our existing knowledge. Here possibly is an area which is potentially as exciting and important as nuclear physics. And if there is such a thing as ESP, I am confident that we will be able to account for it in physical and biochemical terms without any recourse to vitalism.

Charles L. Brown
School of Mechanical Engineering
Purdue University
Lafayette, Indiana

Gentlemen:

Various statements made by Kendon Smith [1] do not seem to constitute a dispassionate assessment of the current status of parapsychology. First, he states that, "At the very best, however, the evidence favoring the existence of psychic phenomena is flimsy in the extreme." As one reviewer [2] said of Soal and Bateman's [3] book, "Rather than argue the question at a philosophical level—a level at which there is little agreement concerning definitions and ideals—Soal and Bateman's data, like other scientific evidence, should be evaluated on their empirical merits. And at this level, it seems, most critics are willing to concede that the findings were obtained according to the same rules of procedure we use in orthodox inquiries. The real question, then, is not whether the phenomenon (or phenomena) uncovered by Soal and Bateman is a 'fact'—for by current empirical standards it is a 'fact'—but what kind of a fact is it? What significance does this 'fact' have for the rest of science?" Also, Schmeidler and McConnell [4] have fulfilled, in essence, the following desideratum of Boring [5]. "In a good experiment you would turn telepathy on, and note the number of hits. Then you would turn it off—the control experiment—and note the number. If the difference were large enough to show that you are probably not in the two series dealing with the same populations of guesses, then you have ESP and also an indication of how surely you have it." (Schmeidler and McConnell studied clairvoyance, not telepathy, and they did not turn the phenomenon off and on with a button, but they did show highly significant differences in the extent to which clairvoyance did or did not operate in two groups between which one would have expected to find such differences according to a reasonable, if surprising, hypothesis.)

Second, Smith asserts, "Even those favorably disposed toward the psychic cause will concede that a great part of its literature is worthless; and the remainder of it is far from impeccable in terms of experimental design and analysis." As a professional biometrician, I was most impressed by the Schmeidler-McConnell book; and Ramsey [6] stated in a review of the Soal-Bateman book, "In their endeavor to eliminate deception, fraud and artifact, they have imposed extreme and, at times, absurd precautions."

Smith speaks of "the scientific naïveté which pervades the realm of 'parapsychology.' " Hyman [2] notes that "Soal and Bateman, anticipating negative reactions to their work, have fortified these odds with sophisticated arguments and evidence to the effect that their positive results are not due to recording errors, optional stopping, improper selection of data, ad hoc tests of hypotheses, wrong statistical model, inadequate randomization of target symbols, or deliberate signalling between agent and percipient." This writer might add that Schmeidler and McConnell have been very cautious in drawing conclusions from their data. The statistical procedures they have chosen make certain highly improbable events (in terms of nonparametric models which they did not use) seem to be accountable by chance, using their probability level.

This writer does not know or care where parapsychology is headed. However, he does feel that parapsychological phenomena are legitimate objects of scientific inquiry, and that results of such investigation may be of enormous import for the rest of science. Furthermore, he feels most strongly that other scientists should evaluate results in this field by the same criteria they would apply to results in their own fields, even if the results contradict precepts from emotion-charged areas of our belief.

Respectfully yours,

Kenneth E. F. Watt
1441 Woodroffe Ave., Ottawa 5
Ontario, Canada

1. Kendon Smith. *American Scientist*, *46*, 413 (December 1958).
2. R. Hyman. American Statistical Association, *Journal*, *52*, 607 (1957).
3. S. G. Soal and F. Bateman. Modern Experiments in Telepathy, Second Edition, New Haven, Conn.: Yale University Press, 1954.
4. G. R. Schmeidler and R. A. McConnell. ESP and Personality Patterns, New Haven, Conn.: Yale University Press, 1958.
5. E. G. Boring. *American Scientist*, *43*, 108 (1955).
6. W. O. Ramsey. *Science*, *121*, 361 (1955).

The Editors of the *American Scientist* have kindly allowed me to read the comments of Mr. Brown and Mr. Watt in advance of publication, and have suggested that I respond to these comments. I shall attempt to do so briefly.

As Brown indicates, there seems to be a rather general tacit assumption that, if "ESP" (in the operational sense of "communication not mediated by known physical processes") does exist, its existence must imply supernatural causation. This tacit assumption does ignore a second possibility, that such ESP might be mediated by *unknown* physical processes. Although this latter possibility is always with us, it is evidently not seriously regarded; and I am inclined to doubt that it should be. While no one can be certain of tomorrow's discoveries, it does not now seem very likely that they will include the requisite, radically new, physical process. I am grateful for Brown's strong support of my general position, and sympathetic with his faith in an ultimate naturalistic explanation for even an hypothetical phenomenon; but I am not prepared to concede, as he seems to be, that ESP is real.

Whatever motivation may prompt my conservatism is, of course, completely beside the point scientifically. A view is either evidentially defensible or it is not, and that is the end of the matter. I wish to repeat that the evidence for ESP is not the least bit compelling and that conservatism is justified; perhaps I may elaborate this position a bit in terms of Watt's comments.

With some slight trepidation, I am willing to agree with Watt's contention that the statistical procedures employed in the better parapsychological experiments are appropriate to the ostensible nature of these experiments. I would not agree, however, that the ostensible nature of these experiments is necessarily their true nature—that they are empirically adequate. In the main, the "critics" who have endorsed the procedures employed in these studies have not been qualified to do so: they have not been persons trained and skilled in behavioral research.

To competent psychologists, even Soal and Bateman's book is not likely to be reassuring. As nearly as one can determine from its somewhat unsystematic account, the research with the subjects Basil Shackleton and Mrs. Stewart was carried on in a distinctly casual fashion. Work with Shackleton proceeded in his own basement studio-lodgings (pp. 132, 134), presumably in the evening (pp. 121, 134–166); usually, at least four other persons were present, and there was a constant turnover in personnel among them—during one period, for instance, there was "an influx of fresh visitors" (pp. 134–166, esp. p. 149); Shackleton "was best at this kind of thing . . . after he had had a drink or two," and upon at least one occasion the experimenters joined him (p.121); signalling was vocal (p. 139); when a session was over, Shackleton customarily left his rooms immediately and alone, while the experimenters finally "adjourned to a restaurant, where scores were re-checked" (pp. 141, 144). Mrs. Stewart was tested in her own home, "in the late evening" (p. 200); she was seated in the kitchen, with an aide who was expected to "talk to her and amuse her while she was making her guesses," and there was often "a continual stream of amusing banter" or "conversation . . . mixed with loud laughter" (pp. 200–203); again, signalling was vocal (p. 202); again, too, there was considerable turnover in experimental personnel (p. 204), and on one occasion a "complete stranger" (p. 207) became a responsible member of the party; Mrs. Stewart, more sociable than Mr. Shackleton, prepared tea for the group at the conclusion of each session (p. 206). It will be recalled that, although Mrs. Stewart was successful in calling the target-card itself (p. 209), Shackleton's results showed, *not* a significant rate of *hitting* the target-card, but rather a significant pattern of *missing*—a circumstance whose peculiarity obviously taxes even Soal and Bateman's explanatory facility (pp. 148–175).

These simply are not adequate experiments; and, recalling that they are regarded in parapsychological circles as monuments of methodological rigor, I do not think I was extravagant in my original characterization of the general state of affairs in these circles. There is further evidence, too, to support this characterization: psychical researchers do not even agree among themselves upon the acceptability of the various studies in their literature (*cf.* Nicol, J. F., pp. 24 ff. in: Wolstenholme, G. E. W., and Millar, E. C. P. [eds.]; *Ciba Foundation Symposium on Extrasensory Perception;* Boston: Little, Brown and Company, 1956); more specifically, Soal has expressed misgivings about recent American work (Soal, S. G.; The next ten years in England; *Journal of Parapsychology*, 1948, *12*, 32–36), and one of the best known of the American investigators seems almost dubious about her own results (Schmeidler, G. R.; Personality correlates of ESP as shown by Rorschach studies; *Journal of Parapsychology*, 1949, *13*, 23–31).

One cannot help admiring and respecting the energy and integrity with which much psychical research is carried out. When experimental psychologists "evaluate results in this field by the same criteria they would apply to results in their own fields," however, as Watt feels they should, they rarely find this research and its results acceptable.

KENDON SMITH
The Woman's College
University of North Carolina
Greensboro, N.C.

GENTLEMEN:

In the December issue of the AMERICAN SCIENTIST there were two articles dealing with the moral position of science, Kartman on "Science and the Christmas Spirit" and Smith on "The Naturalistic Conception of Life."

It is hard to justify Smith's casting of most of his paper in a strongly deterministic mold, when his basic resolution of the problem of free will and determinism looks so much like the logical (Is it, really?) attempt to incorporate what is to all intents and purposes free will as one of the causal factors in a deterministic system. Using terms like "volition" and "choice" can hardly prevent, without much more discussion on the part of the author, the impression that this is basically a compromise which weakens the foundations of the typical deterministic position, and should lead to a less partisan examination of it.

Kartman's essay contained many excellent points, some of which could well serve as a starting point for the above examination.

Very truly yours,

DONALD VIVES
Alabama Polytechnic Institute
Auburn, Alabama

GENTLEMEN:

Being a member of Sigma Xi and a reader of AMERICAN SCIENTIST for the past ten years, I hope I have the prerogative to criticize a certain policy of the magazine. I refer to the abundance in the magazine of advertisements from companies lately sprung up whose sole existence in this world would seem to be making of weapons deadly enough to blow up all of us ten times over.

As I see the purpose of scientific endeavor it is not only the seeking of truth but the seeking of it so that the fruits of the search will be the service of the common good of all man. Now I am old fashioned enough to explode when I see that in the magazine which is the spokesman for scientific research in this country there are pages and pages devoted to the exaltations of the glories of this research when it is put to the blasphemous use of murdering most of mankind. Of course I am being moral in this matter, but I do believe that it is about time that we scientists be mature and responsible enough to look about us and observe the world outside the laboratory.

I do hope you print this, for I think the response of many other readers might surprise you.

Sincerely yours,

PHILIP SIEKEVITZ
Rockefeller Institute
New York, N.Y.

Editors' Note: The moral judgment as to where to draw a just line between pacifism and self-defense is the prerogative, even the duty, of every citizen. The Editors have no mandate to draw this line for our readers. We are of the opinion that not more than 5% of the advertisements published fall in the category of "whose sole existence in this world would seem to be the making of weapons deadly enough to blow up all of us ten times over."

SIR:

Hammond's paper, "Smoking and Death Rates—A Riddle in Cause and Effect," *American Scientist 46*, 331–354, will probably help many to understand why the available data on smoking and lung cancer are logically incompetent. Unfortunately, there are two points to which criticism must be directed.

One is the statement (p. 349) that ". . . one can imagine that the use of tobacco in all forms is highly associated with some 'third factor' which causes cancer . . . ; but no one has yet suggested any 'third factor' which seems to meet these specifications." Unless Hammond is hiding behind the word "highly," the suggested possibility of genetic involvement constitutes such a third factor.

Second, and much more serious, is the statement (p. 350) that "It has been suggested that there may be some hereditary factor which results in both lung cancer and a strong desire to smoke cigarettes. This is an ingenious idea. However, if it is true, one must assume that a genetic factor of this sort appeared and became widely spread throughout the populations of many countries during the last fifty years." This statement is wrong or misleading in a variety of ways.

There is no need whatever to assume the recent appearance and rapid spread of such a genetic factor. Instead, it may have been widespread in human populations for eons. In particular, a genetic tendency to lung cancer would be under very slight selection pressure so long as it rarely affects individuals of reproductive age. This must have been almost always the case until recent times, and would still be true usually. Also, it should be remembered that some entirely different effect of the same hypothetical genetic factor may cause it to be favored by natural selection among young people. Nor does it seem wholly ingenious to label the genetic hypothesis an "ingenious" one.

Furthermore, it is well established that many diseases are genetically controlled, or genetically influenced in likelihood of occurrence, age of onset, severity, and so on. With respect to genetic causation of cancer, much of the evidence is of the same character, and open to the same types of criticism, as that relating to smoking. Therefore *it is just as good*, and cannot be logically rejected by Hammond.

Finally, it is a pity that this statement should have been published in December, 1958, four months after the appearance in *Nature* of two letters-to-the-editor on this subject, written by Sir Ronald Fisher, with no footnote to or editorial comment on Hammond's paper to inform the reader that there is now a showing, beyond reasonable doubt, that the genotype has an important influence on choice of smoking habit.

In particular, Fisher reported 33 pairs concordant in smoking habit among 51 monozygous pairs, but only 11 concordant among 31 dizygous pairs, when classified into four classes, as nonsmokers or as smokers of cigars, cigarettes, or pipes. A later sample showed concordance in 44 of 53 monozygous, and only 9 of 18 dizygous, pairs. It might be thought that the concordance among monozygous twins was not a truly genetic effect, but similar to the strong tendency of infant twins to dress alike. Therefore it is of great interest that the monozygous pairs of the second sample were dissected into 27 pairs separated at birth and 26 reared together. The pairs concordant in smoking habit numbered 23 and 21, respectively, showing that the genotype, not the presence of a like twin nor the choice reached by a like twin, is the controlling factor.

In conclusion, I make a plea for integrity on the part of those affirming that smoking causes lung cancer. Let their various papers and public statements include (1) a frank acknowledgment that any such affirmative conclusion is a mere *opinion*, by no means fully substantiated by available data (this Hammond did) and (2) a familiar standard of comparison, choosing something over which we have, or ought to have some voluntary control, such as the death rate associated with the use of the automobile.

Sincerely,

Horace W. Norton
Professor of Statistical Design and Analysis
University of Illinois
College of Agriculture
Department of Animal Science
Urbana, Illinois

Dear Sirs:

Mr. Norton strongly disagrees with my comments concerning the hypothesis that some hereditary factor causes lung cancer and that the same hereditary factor (or a genetically linked factor) also causes a strong desire to smoke cigarettes.

I said that if this hypothesis is correct, "one must assume that a genetic factor of this sort appeared and became widely spread throughout the populations of many countries during the last fifty years." My reasoning is as follows: Suppose that a particular hereditary factor is both necessary and sufficient for the occurrence of some disease. Then the appearance of the disease in a population group would be roughly proportional to the prevalence of the hereditary factor in that population.

There is evidence that age-specific lung cancer incidence rates, particularly in males, have increased enormously during the last fifty years in many different countries. Very few people who have studied the evidence doubt that this is true. Therefore, if the hypothesis is correct, the proportion of people with the lung-cancer-causing hereditary factor must have increased to a similar degree in the populations of these countries during the last fifty years.

Aside from this, there is very strong evidence that certain environmental factors, particularly long and heavy exposure to dusts containing uranium, chromates, or nickel, result in the development of lung cancer in persons so exposed. This argues against the assumption that the hypothetical hereditary factor is sufficient as well as necessary for the production of lung cancer unless it is also assumed that heredity is only one of many different factors which can produce lung cancer.

Personally, I think it not unlikely that some hereditary factor (or some combination of hereditary factors) makes some people more susceptible than others to the lung cancer producing effects of exposure to certain environmental factors. For example, a hereditary defect in function of cilia in the bronchial tubes might allow the accumulation of carcinogenic materials on the bronchial epithelium provided that the individual happened to inhale such material.

I also think it not unlikely that some hereditary factor(s) has some influence on the probability that an individual will smoke cigarettes heavily as opposed to smoking cigarettes moderately, or smoking a pipe, or smoking cigars, or not smoking.

Assuming that susceptibility to lung cancer is hereditary and assuming that heredity has an influence on the tendency to be an excessive cigarette smoker, the question is whether there is a linkage between these two hereditary factors. So far as I know, there is nothing to suggest that this is the case. If we assume that it is the case, this raises several other questions. I will discuss these briefly.

Heredity is certainly not the only factor influencing the probability that a person will smoke cigarettes. For example, some people such as Seventh Day Adventists do not smoke because of their religion and before 1900 relatively few people smoked cigarettes because manufactured cigarettes were unavailable. (It has been shown that lung cancer is rare today among Seventh Day Adventists and lung cancer was rare in the general population before 1900.) If a multiplicity of factors other than heredity influence the probability that a person will smoke cigarettes, then there could not be a perfect correlation between heavy cigarette smoking and inherited susceptibility to lung cancer. Now, if A and B are associated only because both are associated with C, then the degree of association between A and B will be no greater than the association between A and C and no greater than the association between B and C. It was for this reason that I said that to account for the observed high degree of association between cigarette smoking and cancer by the assumption that some hereditary factor causes both, then one must further assume that this hereditary factor is *highly* associated with both cigarette smoking and the occurrence of lung cancer. I am surprised that Mr. Norton objected to my use of the word "highly" in this connection. My point was: 1) if a fairly large number of people with an inherited susceptibility to cancer and an inherited tendency to smoke cigarettes fail to smoke cigarettes for religious or other reasons, and 2) if a fairly large number of people with no inherited susceptibility to lung cancer and no inherited tendency to smoke cigarettes actually do smoke cigarettes heavily because of other influencing factors, then 3) one would not expect to find such a high degree of association between lung cancer and cigarette smoking as has been reported by a great many independent investigators.

Whether or not heredity has an influence on the development of lung cancer and the tendency to smoke cigarettes heavily, I think it extremely unlikely that heredity alone could account for the great increase in lung cancer in many different countries during the last fifty years. Therefore, some other factors must account for this increase. In my opinion, the evidence strongly supports the theory that cigarette smoking is the major factor responsible for this rise although some other factors such as general air pollution or occupational exposure may have contributed to it. This does not preclude the possibility of a heredity linkage between susceptibility to lung cancer and the tendency to smoke cigarettes when cigarettes are available. It would indeed be unfortunate if it turned out that the very people who have a hereditary urge to smoke cigarettes also are those who by heredity are most susceptible to carcinogenic effects of foreign material inhaled into the lungs.

Sincerely,

E. Cuyler Hammond

THE PHI BETA KAPPA AWARD IN SCIENCE

One Thousand Dollar Annual Phi Beta Kappa Prize

The first Phi Beta Kappa Award for outstanding contributions to the literature of science will be given in December 1959, to the best book on science or the interpretation of science published between July 1, 1958, and June 30, 1959.

The purpose of the award is to stress the need for more literate and scholarly interpretations of the physical and biological sciences and mathematics. Such books as Paul Sears' *Deserts on the March*, Rachel Carson's *The Sea Around Us*, Loren Eiseley's *The Immense Journey*, R. W. Gerard's *Unresting Cells*, Eric Bell's *Mathematics, Queen and Servant of Science*, and George Gamow's *Mr. Tompkins Explores the Atom* are examples of the kind of literary scientific scholarship the award is intended to cover. They symbolize the importance of science as a part of our humanistic studies and remind us that the search for wisdom is still a single enterprise.

The award will be offered annually under the following conditions:

1. Entries must be original publications. This stipulation does not exclude books that contain chapters or sections previously published as articles in magazines, newspapers, or learned journals.
2. It is assumed that entries will ordinarily be the work of a single scientist. An exception might be entries written by an integrated "team" of scholars.
3. If a book has been published abroad before the date of its American publication, it is ineligible unless foreign publication is by arrangement with the American publisher.
4. It is re-emphasized that entries should not be of a technical or highly specialized character. In general, reports on research as such are not eligible.

BOOKS RECEIVED

(Continued from page 4A)

Proceedings of the Conference on the Physical Chemistry of Iron and Steelmaking, Dedham, Massachusetts, 1956; edited by John F. Elliott; 257 pages; $15.00; The Technology Press of M.I.T. and John Wiley & Sons, 1958.

Nuclear Engineering Handbook, edited by Harold Etherington; 1872 pages; $25.00; McGraw-Hill Book Co., 1958.

Cold Injury. Transactions of the 5th Conference, Alaska, 1957, edited by M. Irene Ferrer; 341 pages; $5.95; The Josiah Macy, Jr. Foundation, 1958.

Basic Organic Chemistry by Louis and Mary Fieser; 369 pages; $6.00; D.C. Heath & Co., 1958.

Vapor-Liquid Equilibrium by Eduard Hála, Jiri Pick, Vojtech Fried, and Otakar Vilím, translated from the Czech by G. Standart; 402 pages; $14.00; Pergamon Press Inc., 1958.

The Physics of Elementary Particles by J. D. Jackson; 135 pages; $4.50; Princeton University Press, 1958.

Patients, Physicians and Illness, edited by E. Gartly Jaco; 600 pages; $7.50; The Free Press, 1958.

Theories of Figures of Celestial Bodies by Wenceslas S. Jardetzky; 186 pages; $6.50; Interscience Publishers, 1958.

Ordinary Differential Equations by Wilfred Kaplan; 534 pages; $8.50; Addison-Wesley Publishing Co., Inc., 1958.

Human Factors Journal Vol. I, No. I, September 1958, Stanley Lippert, Editor-in-Chief; $10.00 per year from Human Factors Society of America, Box 24032, Los Angeles 25, California; Pergamon Press Inc., 1958.

Manual of Scientific Russian by Thomas F. Magner; 101 pages; $4.60; Burgess Publishing Co., 1958.

A Primer in Statistical Survey Method: Statistics in the Making by Mary Louise Mark; 436 pages; $5.00; Bureau of Business Research, College of Commerce and Administration, The Ohio State University, 1958.

Elementary Mathematical Programming by Robert W. Metzger; 246 pages; $5.95; John Wiley & Sons, 1958.

Theory of Structural Analysis and Design by James Michalos; 552 pages; $12.00; The Ronald Press Co., 1958.

A Modern Introduction to Ethics, edited by Milton K. Munitz; 657 pages; $7.50; The Free Press, 1958.

Physics of Meteor Flight in the Atmosphere by Ernst J. Öpik; 174 pages; $1.95 (paper); Interscience Publishers, Inc., 1958.

Fine Particle Measurement: Size, Surface, and Pore Volume by Clyde Orr, Jr., and J. M. Dallavalle; 353 pages; $10.50; The Macmillan Co., 1959.

Proceedings of the International Symposium on Transport Processes in Statistical Mechanics, Brussels, 1956, edited by I. Prigogine; 436 pages; $10.00; Interscience Publishers, Inc., 1958.

Microsomal Particles and Protein Synthesis, edited by Richard B. Roberts; 168 pages; $5.00; Pergamon Press Inc., 1958.

Internal Conversion Coefficients by M. E. Rose; 173 pages; $6.25; Interscience Publishers, Inc., 1958.

Occupations and Values by Morris Rosenberg with the assistance of Edward A. Suchman and Rose K. Goldsen; 158 pages; $4.00; The Free Press, 1958.

The Geology of Uranium (Supplement No. 6 of the *Soviet Journal of Atomic Energy*), translated from the Russian; 128 pages; $6.00; Consultants Bureau, Inc., 1958.

Annual Review of Nuclear Science, Vol. VIII, edited by E. Segrè; 417 pages; $7.00; Annual Reviews, Inc., Palo Alto, Calif., 1958.

The Chemical Prevention of Cardiac Necroses by Hans Selye; 236 pages; $7.50; The Ronald Press Co., 1958.

Some Problems of Chemical Kinetics and Reactivity, Vol. I by N. N. Semenov, translated by J. E. S. Bradley; 305 pages; $7.50; Pergamon Press Inc., 1958.

Polysaccharides in Biology. Transactions of the 3rd Conference, Princeton, New Jersey, May 1957, edited by Georg Springer; 249 pages; $4.75; The Josiah Macy, Jr. Foundation, 1958.

Polar Atmosphere Symposium. Part I: Meteorology, edited by R. C. Sutcliffe; 341 pages; $12.00; Pergamon Press Inc., 1958.

Information Storage and Retrieval by Mortimer Taube and Harold Wooster; 228 pages; $6.00; Columbia University Press, 1958.

Climatology and Microclimatology, Arid Zone Research. Proceedings of the Canberra Symposium (in French and English); Unesco; 355 pages; $11.00; Columbia University Press, 1958.

Polar Atmosphere Symposium. Part II: Ionospheric, edited by K. Weeks; 212 pages; $10.50; Pergamon Press Inc., 1958.

Aviation Medicine: Selected Reviews, edited by C. S. White, W. R. Lovelace II, and F. G. Hirsch; 305 pages; $12.00; Pergamon Press Inc., 1958.

Nonlinear Problems in Random Theory by Norbert Wiener; 131 pages; $4.50; The Technology Press, M.I.T. and John Wiley & Sons, 1958.

The Biological Way of Thought by Morton Beckner; 200 pages; $6.00; Columbia University Press, 1959.

Sea Shells of Tropical West America by A. Myra Keen; 624 pages; $12.50; Stanford University Press, 1958.

Biographical Memoirs, Vol. XXXII, National Academy of Sciences; 457 pages; $5.00; Columbia University Press, 1958.

Biological and Biochemical Bases for Behavior, edited by H. F. Harlow and C. N. Woolsey; 476 pages; $8.00; University of Wisconsin Press, 1958.

Organic Syntheses, Vol. XXXVIII, John C. Sheehan, Editor-in-Chief; 120 pages; $4.00; John Wiley & Sons, 1958.

Research and Development, Its Impact on the Economy. Proceedings of a Conference, May 1958; National Science Foundation; 223 pages; $1.25; U. S. Government Printing Office, 1958.

Education and Freedom by H. G. Rickover; 256 pages; $3.50; E. P. Dutton & Co., Inc., 1958.

The Astronomer's Universe by Bart J. Bok; 107 pages; $3.75; Melbourne University Press; New York: Cambridge
versity Press, 1959.
University Press, 1959.

Studies in Linear and Non-Linear Pro-
gramming by Kenneth J. Arrow,
Leonid Hurwicz, and Hirofumi-
uzawa; 229 pages; $7.50; Stanford Uni-
versity Press, 1958.

Venture to the Arctic, edited by R. A. Hamilton; 283 pages; $0.95; Penguin Books Inc., 1959.

Surface Phenomena in Chemistry and Biol-
ogy, edited by J. F. Danielli, K. G. A.
Pankhurst, and A. C. Riddiford; 330
pages; $10.00; Pergamon Press Inc.,
1959.

The Crossing of Antarctica by Sir Vivian
Fuchs and Sir Edmund Hillary; 328
pages; $7.50; Little, Brown & Co., 1959.

The Theory of the Properties of Metals and Alloys by N. F. Mott and H. Jones;

326 pages; $1.85 (paper); Dover Publications, Inc., 1959.

Elasticity, Plasticity and Structure of Matter by R. Houwink; 368 pages; $2.45 (paper); Dover Publications, Inc., 1959.

Theory of Beams: The Application of the Laplace Transformation Method to Engineering Problems by T. Iwinski, translated by E. P. Bernat; 85 pages; $3.50; Pergamon Press Inc., 1958.

Studies on the Structure and Development of Vertebrates, Vols. I and II by Edwin S. Goodrich; 837 pages; $2.50 per volume (paper); Dover Publications, Inc., 1959.

Electronic Apparatus for Biological Research by P. E. K. Donaldson; 718 pages; $20.00; Academic Press, 1959.

Nuclear Spectroscopy Tables by G. J. Nijgh, A. H. Wapstra and R. Van Lieshout; 135 pages; $8.90; Interscience Publishers, 1959.

Nautilus by William R. Anderson with Clay Blair, Jr.; 251 pages; $3.95; World Publishing Co., 1959.

Extensive Air Showers by William Galbraith; 211 pages; $7.50; Academic Press, 1958.

Careers and Opportunities in Engineering by Philip Pollack; 140 pages; $3.50; E. P. Dutton & Co., Inc., 1958.

Atomic Energy: Glossary of Technical Terms; United Nations; 215 pages; $4.00 (paper); Columbia University Press, 1958.

The Theory of the Potential (Theoretical Mechanics) by William Duncan MacMillan; 469 pages; $2.95 (paper); Dover Publications, Inc., 1959.

An Introduction to the Geometry of N Dimensions by D. M. Y. Sommerville: 196 pages; $1.50 (paper); Dover Publications, Inc., 1959.

The Elements of Non-Euclidean Geometry by D. M. Y. Sommerville; 274 pages; $1.50 (paper); Dover Publications, Inc., 1959.

Scientific Manpower in Europe by Edward McCrensky; 188 pages; $6.50; Pergamon Press Inc., 1958.

Elliptic Integrals by Harris Hancock; 101 pages; $1.25 (paper); Dover Publications, Inc., 1959.

Lectures on the Theory of Elliptic Functions by Harris Hancock; 498 pages; $2.55 (paper); Dover Publications, Inc., 1959.

The Foundations of Euclidean Geometry by Henry George Forder; 349 pages; $2.00 (paper); Dover Publications, Inc., 1959.

Biological Laboratory Data by L. J. Hale; 132 pages $2.75; John Wiley & Sons, 1958.

Bibliography of Papers published by L. Zechmeister and Co-Authors in the Fields of Chemistry and Biochemistry, 1913–1958; 22 pages; California Institute of Technology, Pasadena, 1958.

Mental Subnormality (Biological, Psychological, and Cultural Factors) by Richard L. Masland, Seymour B. Sarason, and Thomas Gladwin; 442 pages; $6.75; Basic Books, Inc., 1959.

A Symposium on the Chemical Basis of Development, edited by William D. McElroy and Bentley Glass; 932 pages; $15.00; The Johns Hopkins Press, 1958.

Physical Laws and Effects by Frank C. Hix, Jr., and Robert P. Alley; 291 pages; $7.95; John Wiley & Sons, 1958.

Nuclear Reactions, edited by P. M. Endt and M. Demeur; 502 pages; $12.50; Interscience Publishers, Inc., 1959.

Liquid Scintillation Counting, edited by Carlos G. Bell, Jr., and F. Newton Hayes; 292 pages; $10.00; Pergamon Press Inc., 1958.

The Pulse of Radar, The Autobiography of Robert Watson-Watt by Sir Robert Watson-Watt; 438 pages; $6.00; The Dial Press, Inc., 1959.

Darwin, Wallace, and the Theory of Natural Selection Including the Linnean Society Papers by Bert James Loewenberg; 97 pages; $5.00; Arlington Books, Inc. Distributed by Taplinger Publishing Co. Inc., 1959.

Operations Research for Industrial Management by P. E. Chorfas; 303 pages; $8.75; Reinhold Publishing Corp., 1958.

Catalysis, Vol. VI: Hydrocarbon Catalysis, edited by Paul H. Emmett; 706 pages; $19.50; Reinhold Publishing Corp., 1958.

The Science of High Explosives by Melvin A. Cook; 440 pages; $22.50; Reinhold Publishing Corp., 1958.

Rocket Propellants by Francis A. Warren; 218 pages; $6.50; Reinhold Publishing Corp., 1958.

Introduction to Geophysics by Benjamin F. Howell, Jr., 399 pages; $9.00; McGraw-Hill Book Co., 1959.

History and Philosophy of Science, introduction by L. W. Hull; 340 pages; $5.00; Longmans, Green & Co. Inc., 1959.

Drinking and Intoxication: Selected Readings in Social Attitudes and Controls, edited by Raymond G. McCarthy; 455 pages; $7.50; The Free Press, 1959.

Statistics: An Introduction by Donald A. S. Fraser; 398 pages; $6.75; John Wiley & Sons, 1958.

Expansion Machines for Low Temperature Processes by S. C. Collins and R. L. Cannaday; 124 pages; $3.00 (paper); Oxford University Press, 1958.

Psychology of the Child: Personal, Social and Disturbed Child Development by Robert I. Watson; 662 pages; $6.95; John Wiley & Sons, 1959.

Soviet Research in Fused Salts, 1956. Part I: Systems (Binary, Ternary: Quaternary Reciprocal); 147 pages; $30.00; *Part II: Electrochemistry; Aluminum & Magnesium, Corrosion, Theoretical; Thermodynamics; Slags & Mattes;* 268 pages; $10.00; Complete Set: $40.00; Consultants Bureau, Inc., 1958.

Advances in Chemical Physics, Vol. I, edited by I Prigogine; 414 pages; $11.50; Interscience Publishers, 1958.

Introduction to Neutron Physics by L. F. Curtiss; 380 pages; $9.75; D. Van Nostrand Co., 1959.

Zoogeography, edited by Carl L. Hubbs; 509 pages; $12.00; American Association for the Advancement of Science, Washington 5, D. C., 1959.

The Rorschach and the Epileptic Personality by J. Delay, P. Pichot, T. Lempérière, and J. Perse, translated by Rita and Arthur L. Benton; 265 pages; $6.00; Logos Press, 1959.

Tree Fruit Production by James S. Shoemaker and Benjamin J. E. Teskey; 456 pages; $6.95; John Wiley & Sons, 1959.

The Essentials of Chemistry by R. P. Graham and L. H. Cragg; 578 pages; $6.50; Rinehart & Co., Inc., 1959.

The Angers of Spring, a Novel by Joseph Whitehill: 373 pages; $4.50; Boston: Atlantic-Little, Brown, 1959.

The Potential Theory of Unsteady Supersonic Flow by J. W. Miles; 220 pages; $8.50; New York: Cambridge University Press, 1959.

Organic Chemistry, A Short Text by George K. Estok; 275 pages; $5.50; W. B. Saunders Co., 1959.

Dynamics of Flight: Stability and Control by Bernard Etkin; 519 pages; $15.00; John Wiley & Sons, 1959.

Electromagnetic Phenomena in Cosmical Physics, edited by B. Lehnert; 545 pages; $10.00; New York: Cambridge University Press, 1959.

The Evolution of Living Things by H. Graham Cannon; 180 pages; $3.50; Charles C Thomas, Publishers, 1958.

Statistical Estimates and Transformed Beta-Variables by Gunnar Blom; 176 pages; $5.00; John Wiley & Sons, 1959.

The Central Nervous System and Behavior, Transactions of the First Conference, 1958, edited by Mary A. B. Brazier; 450 pages; $5.25; Josiah Macy, Jr. Foundation, 1959.

Matter and Gravity by Eigil Rafmussen; 75 pages; $3.50; Exposition Press, 1958.

Liquid Helium by K. R. Atkins; 312 pages; $11.00; New York: Cambridge University Press, 1959.

The Simplicity of Science by Stanley D. Beck; 212 pages; $3.75; Doubleday & Company, Inc., 1959.

Cobalt, No. I, December 1958. A quarterly in French and English issued at no charge. Cobalt Information Center, c/o Battelle Memorial Institute, Columbus, O.

Introduction to Symbolic Logic and Its Applications by Rudolf Carnap; 241 pages; $1.85 (paper); Dover Publications, Inc., 1959.

Basic Geology for Science and Engineering by Edward C. Dapples; 609 pages; $9.50; John Wiley & Sons, 1959.

Psychopharmacology Frontiers (Proceedings of the Psychopharmacology Symposium, 2nd International Congress of Psychiatry), edited by Nathan S. Kline; 533 pages; $10.00; Little, Brown & Co., 1959.

Linear Groups with an Exposition of the Galois Field Theory by Leonard E. Dickson; 312 pages; $1.95 (paper); Dover Publications, Inc., 1959.

Celestial Mechanics by E. Finlay-Freundlich; 150 pages; $7.50; Pergamon Press Inc., 1959.

Electrical Measurement Analysis by Ernest Frank; 443 pages; $8.75; McGraw-Hill Book Company, Inc., 1959.

An Introduction to Fourier Methods and the Laplace Transformation by Philip Franklin; 289 pages; $1.75 (paper); Dover Publications, Inc., 1959.

Big Molecules by Sir Harry Melville; 180 pages; $3.95; The Macmillan Co., 1958.

The Practical Use of the Microscope Including Photomicrography by George Herbert Needham; 493 pages; $15.50; Charles C Thomas Publisher, 1958.

The Value of Science by Henri Poincaré; 147 pages; $1.35 (paper); Dover Publications Inc., 1959.

Charles Darwin: Evolution and Natural Selection, edited by Bert James Loewenberg; 438 pages; $5.75; The Beacon Press, 1959.

THE SCIENTIST'S BOOKSHELF

By Hugh Taylor and Guest Reviewers

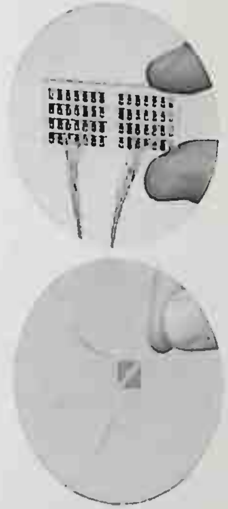
Personal Knowledge: Towards a Post-Critical Philosophy by MICHAEL POLANYI; 428 pages; $6.75; The University of Chicago Press, 1958.

In his earlier novel, *The Search*, recently republished, C. P. Snow makes one of his characters say of the scientists: "The first collection of people in the world who've been trained to be honest and detached about the things they see. They've vowed honesty and detachment, and that's something staggeringly new." Those to whom this remark comes with approving thoughts should address themselves to the much more difficultly readable book by Michael Polanyi which he entitles "Personal Knowledge." This book represents more than ten years of intense study and thought since its author forsook his brilliant career in physical chemistry to become a professor of social studies in Manchester University. The author's Gifford Lectures in 1951–52 in the University of Aberdeen, lectures with a most distinguished tradition, have been embodied in the present volume. The book owes much to his contacts with the Committee on Social Thought in the University of Chicago. These contacts resulted in an invitation to become a professor in the University of Chicago, an invitation which was accepted. How governmental agencies in this country frustrated this arrangement is none too admirable a chapter in recent history.

Polanyi's principal theme in this book is directly contradictory to that of the friend of Arthur Miles, C. P. Snow's hero, already expressed. Polanyi denies that science is dispassionate, completely detached and that scientific knowledge is impersonal. He believes that the scientist's personal participation in his knowledge, in its discovery and in its validation is an indispensable part of science itself. "Personal knowledge in science is not made but discovered . . . . It commits us, passionately and far beyond our comprehension, to a vision of reality. Of this responsibility we cannot divest ourselves by setting up objective criteria of verifiability . . . . For we live in it as in the garment of our own skin. Like love to which it is akin, this commitment is a 'shirt of flame' blazing with passion and, also like love, consumed by devotion to a universal demand. Such is the true sense of objectivity in science."

The book divides itself into four parts, the art of knowing, the tacit component, the justification of personal knowledge and, finally, knowing and being. In the first part objectivity, probability, order and skills are discussed. The second section includes articulation, intellectual passions and conviviality. The logic of affirmation, the critique of doubt and, next, commitment, belong in Part 3, while the final chapters cover the logic of achievement, knowing life, and one on the rise of man.

A friend of mine has written me, in reply to an expression of my enthusiasm for the book, that, while it seems to him an important one, he finds it difficult to understand, not so much in any particular place but the general drift of the argument and the relevance of each part of the whole. This is no easy book to read and its author is aware of the hostilities which his thoughts arouse. "I have watched" he writes, "many a university audience listening to my account of intuitive discoveries silently, with sullen distaste." Fortunately for those unwilling to tackle this monumental volume there is the possibility of sampling its contents in minor doses since there have been articles, corresponding to chapters in this book, published by the author in the years from 1951 to 1957 in such different journals as



The British Journal for the Philosophy of Science, Dialectica, The Lancet, Encounter, with an article in *Science* (*125*, March, 1957) which is a brief survey of the main argument. Entitled "Scientific Outlook: Its Sickness and Cure," the reading of it may well induce a reader to tackle this fine fruit of ten year detachment from experimental science, a condition obtaining also in the case of C. P. Snow.

Biographical Memoirs of Fellows of the Royal Society. Vol. IV; 359 pages; 30s; London: The Royal Society, 1958.

The annual volume contains, this year, twenty-seven biographies of deceased Fellows of the Royal Society. For American readers the memoir on Langmuir will, perhaps, be the first to attract attention. That by Sir George Thomson on Frederick Lindemann, Viscount Cherwell, should be of interest to chemists of the Nernst school, to physicists generally, and to all interested in the interactions of a scientist with men of political and public affairs. Sir George reveals how Lindemann was one power behind the Churchill throne, brought there by an incisive attitude on a whole spectrum of life-and-death questions to the British nation, in the uneasy days before World War II and in the midst of that strife. David Chapman, Douglas Hartree, John Graham Kerr, Franz Simon, Viscount Waverley, better known as John Anderson, and Heinrich Wieland are among those also reported upon with the usual Royal Society standards of excellence.

American Agriculture: Geography, Resources, Conservation by EDWARD HIGBEE; 399 pages; $7.95; John Wiley & Sons, 1958.

Efficiency in Government Through Systems Analysis with Emphasis on Water Resource Development by ROLAND N. MCKEAN; 336 pages; $8.00; John Wiley & Sons, 1958.

California Lands, Ownership, Use, and Management by SAMUEL T. DANA and MYRON KRUEGER; 308 pages; $4.50; The American Forestry Assoc., Washington, D. C., 1958.

Statistics in the Making: A Primer in Statistical Survey Method by MARY LOUISE MARK; 436 pages; $5.00; Bureau of Business Research, College of Commerce and Administration, The Ohio State University, 1958.

Each of the four books listed above merits critical review and would, in my judgment, stand up under it. Yet, despite the range of titles, it seems to me more important to group them at this time. For each, in its way, presents a facet of a large and fundamental problem. We are accustomed by now to the practice of precision in the experimental sciences. We are far less given to making full use of knowledge to minimize the areas of uncertainty in what might loosely be called public affairs. For assuredly the ecology of man, under our system of government, must be dealt with as a matter of public concern. Drift, impulse and intuition are no longer safe, or in large measure even necessary, even though they cannot ever be wholly eliminated.

Edward Higbee, Professor of Geography and Agricultural Economics at Delaware, has already demonstrated his sound knowledge of land use in the United States in his *American Oasis*. Here he presents a more systematic and detailed geography of our agriculture. Starting with a general analysis of the factors influencing land use, he proceeds to discuss typical regions, using well-chosen specimen farm layouts by way of example.

Two emeritus professors, Samuel Dana of Michigan and Myron Krueger of California, sponsored by the American Forestry Association, analyze land use problems in one of the most rapidly growing and heavily urbanized states of the Union. While their chief concern is with wild lands—forest and range—they give due attention to the impact of competing uses. They take into account not only the complex capabilities caused by climatic and edaphic factors, but the difficulties caused by equally complex ownership patterns and the lack of clear objectives. Stressing the necessity for unremitting research, they suggest the need of a strong Natural Resources

Council to insure study and coordination.

Roland N. McKean, a staff member of the renowned study group known as the Rand Corporation, has performed the very great service of discussing systems analysis in a form highly acceptable to the many who should know more about it. He makes clear that such analysis does not eliminate uncertainty, but can do much to reduce it. Although he uses two water projects as specimens, his approach can, as he shows in his concluding table, be widely applied in government as well as business. But he serves us the wry reminder that public officials are not under the same economic pressures for efficiency that obtain outside.

Mary Louise Mark, Emeritus Professor of Sociology and Social Administration at Ohio State, draws upon her rich experience to show in a very practical way how data, especially on social matters, should be gathered and organized. With a sufficiency of books that tell us how to treat data mathematically after they are in, it is good to learn more about the earlier, essential steps. Written with great clarity, the book is entertaining as well as most informative.

It is high time in this so-called scientific age, that the discussion of public affairs take full account of the possibilities for enlightenment that are available. Together, these books form an interesting pattern to justify this statement.—*Paul B. Sears*

The Physical Theory of Neutron Chain Reactors by ALVIN M. WEINBERG and EUGENE P. WIGNER; 800 pages; $15.00; The University of Chicago Press, 1958.

From the realization, in the early part of 1939, that neutrons could cause fission of the uranium nucleus to the first self-sustaining neutron chain reaction, in December 1942, barely four years elapsed. Yet, within this short time, a newly born branch of theoretical physics almost reached maturity. One of the main contributors to this development was Eugene P. Wigner, whose efforts have been recognized by the Atomic Energy Commission's Enrico Fermi Award for 1958, and among his able collaborators was Alvin M. Weinberg, now Director of the Oak Ridge National Laboratory. These two scientists have now combined to make available, in the book under review, their extensive knowledge and experience of the theory of nuclear chain reactors. It is well known that this book has been in preparation for several years, but the delay has not been without advantage. The authors state in their preface: "We have been able to sharpen many of our own ideas . . . and to add much from the independent thinking of reactor theorists throughout the world." The inclusion of references to papers published as late as 1958 shows that this is indeed the case.

The book is divided into twenty-two chapters falling into four sections. The first, entitled "The Nuclear Physics of Chain Reactions," covers the basic concepts of reactor theory, nuclear reactions and cross sections, the fission process, and the shell structure of nuclei. In the second section, on "Transport Theory of Neutrons," the general problems of neutron chain reactions and the diffusion and thermalization of neutrons are discussed. The next section, called "General Reactor Theory," deals with the theory of bare and reflected homogeneous reactors, including reactor kinetics. Finally, there is a section on "Heterogeneous Reactors" that takes up in detail the special theory of such reactors which, as the authors say, was their "first love."

As was to be expected, this book is a classic in the field of reactor theory. Its inclusion in the set of volumes presented by the AEC to the delegates at the 1958 Geneva Conference, although it was not prepared for this purpose, indicates official recognition of the work as a significant U.S. contribution to nuclear science. In the circumstances, therefore, this reviewer feels reluctant to introduce a slightly "sour note." With the development of a new science, nomenclature is always a problem, and an effort has been made to achieve some standardization in the terms and symbols employed in nuclear reactor theory. It is regrettable, therefore, that this book runs counter to the common usage in some respects, although, it must be admitted, the authors provide justification for their choice. Perhaps the most important is the attempt to introduce the term "fermi" in place of the "barn." It is true, in retrospect, that the fermi would have been a better name for the unit of cross section, but the word barn has been used for some sixteen years and is now well established. It is a curious (and unfortunate) fact, to judge from the report of the barn's originators, that, although the names of several scientists were considered, Fermi, whose contributions were so outstanding, was not among them. Apart from this minor criticism, the book is one that can be strongly recommended to those interested in the physics of neutron chain reactors. It is probably not a book for beginners, but those having some knowledge of the field will find this a work that they can read and study with advantage.—*Samuel Glasstone*

Mind and Matter by ERWIN SCHRÖDINGER; 104 pages; $2.75; New York: Cambridge University Press, 1958.

"Who are we really? Where have I come from and where am I going?" are questions Schrödinger names as baffling; in fact they are admitted today to be the most challenging questions not only of philosophy but also of science itself. Schrödinger, in this small book of essays, originally the Tarner Lectures delivered at Trinity College, Cambridge, in 1956, discusses them from the physicist's point of view.

He conceives that "consciousness is associated with the *learning* of the living substance; its *knowing how* . . . is unconscious." Strongly reliant on Sherrington, especially Sherrington as author of *Man on his Nature*, he adopts his dualism: "Sensations and thought do not belong to the 'world of energy,' they cannot produce any change in this world of energy as we know from Spinoza and Sir Charles Sherrington." Know, or believe? Schrödinger's own argument speaks for dualism principally from his considerations of the sense organs as related to subject and object: "'Registering as sound' . . . simply is not contained in our scientific picture, but is only in the mind of the person whose ear and brain we are speaking of."

He maintains that modern theoretical physics has much to contribute to the understanding of mind, particularly as it has studied time and as it has increasingly idealized it. Physical theory, he firmly believes, "in its present stage strongly suggests the indestructibility of Mind by Time." Whether dualism or its opposite will prove to be the more popular position in the immediate future, there is no doubt that through its preoccupation with time modern theoretical physics will be called upon to draw us beyond borderlines to the center of the most provocative problems facing modern science.—*Jane Oppenheimer*

Petroleum Refinery Engineering (4th Edition) by W. L. NELSON; 960 pages; $15.00; McGraw-Hill Book Co., 1958.

In technical literature there are works which deserve to be called classics. *Petroleum Refinery Engineering* is such a classic. This is the fourth edition of a book that has been a standard reference in the petroleum and allied fields since 1936. It reflects a lifetime of experience and specialization by the author.

Technological advances in petroleum engineering have increased exponentially in the past decade. Consequently, it is becoming very difficult to present a complete coverage of this subject in one volume. This is effectively accomplished by Dr. Nelson, however.

Each of the twenty-four chapters is a complete treatment of a phase of petroleum technology. The chapters are too numerous for even paraphrased tabulation, but they describe petroleum composition, characterization, inspection, processing, and plant design and economics. The descriptive sections are liberally supplemented by tabular and graphical data, chemical formulary and process flow sheets.

New material is contained in the chapters on extraction, cracking, reforming and others. A most valuable addition is the discussion dealing with oil evaluation and prediction of product properties from given crudes. This is supported by a completely new appendix which summarizes physical properties of crude and products for every principal domestic and foreign crude source.

In early editions Dr. Nelson demonstrated the application of chemical engineering principles to the petroleum industry. He has continued this presentation. Basic concepts of unit operations such as heat transfer, extraction distillation and others are reviewed. Theoretical bases are developed into pertinent examples. Realistic problems are lucidly outlined and solved. For example, equilibrium vaporization-condensation theory leads into description of both the McCabe-Thiele and Ponchon procedures for solving fractional distillation systems.

Cross-referencing is excellent and bibliographies of pertinent references follow most chapters.

This authoritative volume contains many excellent examples of not only factual information but notes misconceptions often harbored by even those active in the field. One such is the clarification of the definition of "sour crude." Too often, any high sulfur crude is called "sour" but actually the hydrogen sulfide content determines this classification.

Petroleum Refinery Engineering should be a valuable asset to persons in many technical disciplines and indispensable to chemical and petroleum engineers.—*Robert A. Baker*

Introduction to the Theory of Sound Transmission by C. B. OFFICER; 284 pages; $10.00; McGraw-Hill Book Co., 1958.

Sound Pulses by F. G. FRIEDLANDER; 202 pages; $7.50; Cambridge University Press, 1958.

These two books are concerned with several aspects of the theory of the transmission of sound waves, and to a large extent are complementary to one another. Both books deal primarily with the simplified cases where there is no energy loss in the medium being traversed by the sound wave, and where the amplitudes are small enough that only linear terms need be considered in the wave equations.

The book by Officer is the more general. Transmission through solids, liquids, and gases is considered, with interest centering on the two problems of sound waves in the ocean, with both shallow and deep water, and in layered geologic structures. Analysis is carried out by the alternate procedures of tracing the paths of rays and considering the summation of normal modes. Transmission of pulses and of simple harmonic waves is discussed. Only a small amount of space is devoted to scattering and diffraction problems. While the treatment is largely mathematical, considerable experimental data are also given.

The book by Friedlander is limited to a consideration of fluid media and to the transmission of pulses with well-defined wave fronts. Problems considered include the scattering and diffraction of a pulse having a specified wave front produced when the pulse strikes an obstacle of specified shape. The treatment is entirely mathematical, and is based on a theory of distributions and "weak" solutions for the wave equation.

The book by Officer contains relatively many more illustrative figures than does that by Friedlander, but unfortunately most of these figures carry no descriptive captions.—*W. J. Cunningham*

Foundations of Embryology by BRADLEY M. PATTEN; 578 pages; $10.50; McGraw-Hill Book Co., 1958.

Foundations of Embryology is a comprehensive textbook of vertebrate development in which the author has incorporated the essence of his well-known works on chick, pig, and human. To say more is almost an impertinence, so thoroughly have these works demonstrated their value. A generation of students can by now testify to Bradley Patten's skill in writing about, and illustrating, the patterns of development.

The new book follows the straight and narrow path of normal development, beginning with the gametes and proceeding to advanced phases of formation of the fetus. There are no excursions into side issues. The major emphasis is on birds and mammals, but cleavage and gastrulation in amphibians is clearly described, with good diagrams. Later phases of amphibian development are entirely omitted. This lack—which some teachers may regard as a defect—is more than counterbalanced by the presentation of the events of organ development, which is carried to the point of emergence of the gross anatomical features of the adult. This part of the book, in which the author's skill is most clearly reflected, manages to be thorough without the overload of detail that makes organogenesis such a forbidding subject in some of the weightier embryology texts.

The book does however have undeniable deficiencies. Though the work is completely up-to-date, including references to the most recent papers in the bibliography, the text itself all but ignores the experimental approach to embryology. The challenging problems of embryogeny are consequently not brought to the students' attention in any effective way. Abnormal development is overlooked too, and this is a strange omission in the face of the author's statement, in the introductory chapter, that the study of embryology has become of increasing importance because "congenital defects have now moved up among the top ten causes of death in the United States."

Foundations of Embryology will not be acceptable to those who share the curious preoccupation of many embryology teachers with the frog tadpole, nor will it please those who feel that embryology ceases to be worth following beyond the 72-hour stage of the chick embryo. Among those who are concerned to give their students a firm grounding in the events of development of the higher vertebrates, however, Bradley Patten's new work will no doubt win warm approval. In the hand of a teacher prepared to supplement it adequately, the book will provide the basis of a really excellent course. In my opionion, *Foundations of Embryology* is, despite its shortcomings, the most satisfactory textbook of embryology now available.—*Florence Moog*

Microbiology by MICHAEL J. PELCZAR, JR. and ROGER D. REID; 564 pages; $8.00; McGraw-Hill Book Co., 1958.

This addition to the growing number of first texts in the field of microbiology presents to the student an almost complete perspective of the science. Thus, it encompasses all areas in both basic and applied microbiology and should enable beginners to grasp the scope of the science and to be initiated into the phase of his particular interest. Since the range of the text is so inclusive, by necessity, it suffers from lack of detail. However, the fundamental modern theses of microbiology are included and presented in a very simple and understandable fashion. This ease in readability, together with the generous number of illustrations, should tend to stimulate the students' interests.

As is necessary for such a beginner's text, the book is remarkably free of errors. This is an accomplishment for a first edition. Another noteworthy contribution is the inclusion of appendices covering a brief outline of bacterial classification, a compilation of common reagents, media and procedures and a concise glossary.

The simplicity and inclusiveness of the book make it of little value for reference, but of great value as an introduction to microbiology. It is a fascinating discussion of a fascinating science.—*P. F. Smith*

Progress in Biophysics and Biophysical Chemistry, Vol. VIII, edited by J. A. V. BUTLER and B. KATZ, 409 pages; $17.50; Pergamon Press, 1957.

The excellence of this series should be well known by now, and the present volume is up to the standards of its predecessors. There are nine articles chosen from "the wide range of subjects which come under the heading of biophysical research," and as usual each is a critical and authoritative analysis of the present state of a restricted area.

These are not the current literature reviews, listing every reference, but, rather, present a selected and generally lucid discussion of the work the author feels is important. The specialist will find valuable syntheses of recent ideas, while the more general reader can often follow with ease the development of work in a field outside his own.

The articles in Volume 8 range over sensory physiology (hearing, color vision), surfaces and membranes (electrochemistry of the bacterial surface, cytology of the axon surface, ionic permeability of the red cell), molecular biology (physical chemistry of deoxyribonucleic acid (DNA), biosynthesis of connective tissue components, protein biosynthesis) and radiation biology (effect of radiation on DNA synthesis in mammalian cells).

It is unnecessary to discuss each in detail—besides the ones which the reviewer read for his own work, he found the article on color vision particularly interesting. Here, Brindley analyzes the three-channel hypotheses for trichromatic vision and suggests some critical experiments remaining to be done. For those who do not follow closely the current work in molecular biology, the article by Loftfield on protein synthesis is recommended not only for the exciting results it reviews, but also for the variety of techniques which have been involved in arriving at the present state of knowledge.

There is no hesitation in saying that this volume should be in all research libraries, but the price will be steep for many individuals. Technically, the editing and printing are well done and all concerned are to be congratulated again.—*W. R. Guild*

Symposium on Protein Structure, edited by ALBERT NEUBERGER; 351 pages; $7.75; John Wiley & Sons, 1958.

This book is a carefully edited and attractively printed record of the Symposium on Protein Structure which took place in Paris in July 1957 as part of the meetings of the International Union of Pure and Applied Chemistry.

The papers treated under General Problems and Methods include a review of the quantum mechanics of the peptide bond by L. Pauling, a thoroughgoing account of deuterium exchange with the hydrogens of proteins and of synthetic polypeptides by K. Linderstrøm-Lang, a review of zone electrophoresis by A. Tiselius, and a report by L. C. Craig and his co-workers on the escape rates of smaller proteins through cellophane dialysis bags of selected porosities.

The papers treated under Specific Proteins are divided into those on hemoglobin, proteolytic enzymes, ribonuclease, tobacco mosaic virus, and "other proteins" (insulin, pituitary hormones, etc.). The Sanger technique for the determination of the sequence of amino acids is successfully brought to bear on ribonuclease by C. B. Anfinsen and by S. Moore and his co-workers. Anfinsen also attempts to explain enzymatic activity in terms of the amino acid sequence. The papers on the structure of tobacco mosaic virus by H. Fraenkel-Conrat, by G. Schramm, and by the late Rosalind Franklin are also very exciting.

All in all this book successfully brings the reader to the forefront of present knowledge on the structure of proteins. The papers are uniformly good and are frequently highly original in content.—*Gerald Oster*

Population and World Politics, edited by PHILIP M. HAUSER; 297 pages; $6.00; The Free Press, 1958.

Apart from the introductory section by the editor, this volume is comprised of eleven selected public lectures and papers presented at the Thirtieth Institute of the Norman Wait Harris Memorial Foundation.

The topic of the 1954 Institute, "Population and World Politics," has lost none of its importance, nor have the lectures and papers become outdated by events in the interim prior to publication of this volume. Since 1954, unrest in Poland and Hungary have once more demonstrated the ruthless effectiveness of the Communist world in dealing with disaffections while still attracting weaker nations closer to its own orbit. On the other hand, events in the Near East, South Africa and in our own South have again verified that some disunities and tensions in the free world are associated with population growths, changes in population composition, and political instabilities found in underdeveloped areas.

The papers of the eleven contributing scholars are divided into three parts. Parts I and II: (World Population and Resources; Population, Levels of Living, and Economic Development); are introductions to such topics as world and regional population trends, distribution of resources with respect to population, world and regional economic trends, internal migration, and population as it affects or is affected by economic development. Part III: Population Policy and Politics contains four papers dealing with specific relationships between populations and power, political instabilities in underdeveloped areas, United States foreign policy, and population policy in the Soviet Union.

The contributors uniformly agree that population poses not only a problem of numbers but a threat to peace as a source of tensions. Unfortunately, the implementation of their proposed humanitarian solutions appears as remote as the positive check of war seems imminent.

Persons not intimately acquainted with the field of demography will find the papers highly informative factually and helpful to the understanding of a social problem that is an important facet of the political unrest plaguing the world today.—*Philip C. Sagi*

A Textbook of Psychology by DONALD O. HEBB; 276 pages; $4.50; W. B. Saunders Co., 1958.

Interest in the correlation of behavioral properties with physiological and electro-neural measures has increased notably since the publication of Hebb's *Organization of Behavior* in 1949. This elementary textbook includes Hebb's current analysis of the relation of neuro-anatomy to behavior, an area of interest long neglected by many psychologists because of their belief that neuro-physiology was not sufficiently advanced to help them to understand gross behavioral adjustments to the environment. Hebb disagreed with the "empty organism" approach in his 1949 book, and by 1958 has been successful in convincing an ever-increasing number of psychologists of the utility of explicit "neurologizing."

Hebb clearly states his position on problems which are frequently bypassed in elementary texts; such as the differences between the biological and the physical sciences, the scientific significance of the unique event, scientific reductionism, the function of theory, and the nature of evidence in psychology.

The instructor who, like Hebb, wishes to present psychology as a biological science rather than a social science or a set of techniques for solving practical problems, will appreciate the selection and presentation of topics. The brevity of the book will permit the instructor to introduce the student to many of the basic references listed at the end of each chapter. Unfortunately, there is neither a glossary nor a separate student workbook.—*George S. Leavitt*

Analog Simulation: Solution of Field Problems by WALTER J. KARPLUS; 434 pages; $10.00; McGraw-Hill Book Co., 1958.

The same mathematical relations underlie many physical phenomena. Here is a book in which this fact is not only pointed out, but is exploited in setting up analogs for obtaining solutions to engineering problems expressible in terms of partial differential equations.

A brief mathematical section describes the origin of such equations as those of Laplace and Poisson, of diffusion, wave motion, elasticity, and vibration. There is included a discussion of the techniques of transformation of variables and approximation in terms of finite differences.

A second section contains a description of electric models making use of conduction through either a solid or liquid continuous medium, of models based on networks of lumped resistors or disaffections while still attracting weaker nations closer to its own orbit. On the other hand, events in the Near East, South Africa and in our own South have again verified that some disunities and tensions in the free world are associated with population growths, changes in population composition, and political instabilities found in underdeveloped areas.

The papers of the eleven contributing scholars are divided into three parts. Parts I and II: (World Population and Resources; Population, Levels of Living, and Economic Development); are introductions to such topics as world and regional population trends, distribution of resources with respect to population, world and regional economic trends, internal migration, and population as it affects or is affected by economic development. Part III: Population Policy and Politics contains four papers dealing with specific relationships between populations and power, political instabilities in underdeveloped areas, United States foreign policy, and population policy in the Soviet Union.

The contributors uniformly agree that population poses not only a problem of numbers but a threat to peace as a source of tensions. Unfortunately, the implementation of their proposed humanitarian solutions appears as remote as the positive check of war seems imminent.

Persons not intimately acquainted with the field of demography will find the papers highly informative factually and helpful to the understanding of a social problem that is an important facet of the political unrest plaguing the world today.—*Philip C. Sagi*

A Textbook of Psychology by DONALD O. HEBB; 276 pages; $4.50; W. B. Saunders Co., 1958.

Interest in the correlation of behavioral properties with physiological and electro-neural measures has increased notably since the publication of Hebb's *Organization of Behavior* in 1949. This elementary textbook includes Hebb's current analysis of the relation of neuro-anatomy to behavior, an area of interest long neglected by many psychologists because of their belief that neuro-physiology was not sufficiently advanced to help them to understand gross behavioral adjustments to the environment. Hebb disagreed with the "empty organism" approach in his 1949 book, and by 1958 has been successful in convincing an ever-increasing number of psychologists of the utility of explicit "neurologizing."

Hebb clearly states his position on problems which are frequently bypassed in elementary texts; such as the differences between the biological and the physical sciences, the scientific significance of the unique event, scientific reductionism, the function of theory, and the nature of evidence in psychology.

The instructor who, like Hebb, wishes to present psychology as a biological science rather than a social science or a set of techniques for solving practical problems, will appreciate the selection and presentation of topics. The brevity of the book will permit the instructor to introduce the student to many of the basic references listed at the end of each chapter. Unfortunately, there is neither a glossary nor a separate student workbook.—*George S. Leavitt*

Analog Simulation: Solution of Field Problems by WALTER J. KARPLUS; 434 pages; $10.00; McGraw-Hill Book Co., 1958.

The same mathematical relations underlie many physical phenomena. Here is a book in which this fact is not only pointed out, but is exploited in setting up analogs for obtaining solutions to engineering problems expressible in terms of partial differential equations.

A brief mathematical section describes the origin of such equations as those of Laplace and Poisson, of diffusion, wave motion, elasticity, and vibration. There is included a discussion of the techniques of transformation of variables and approximation in terms of finite differences.

A second section contains a description of electric models making use of conduction through either a solid or liquid continuous medium, of models based on networks of lumped resistors or resistors and reactors, and of electronic analog computers. There is also a chapter on nonelectric analogs employing such diverse physical phenomena as hydrodynamics, elasticity, ion diffusion, and flow of heat or electromagnetic flux. In most cases a discussion of the sources of error is given, together with information about the degree of accuracy that can be obtained.

In a final section, specific examples are given showing the application of such analogs as those just described to specific engineering problems. These examples range from the design of electrode structures to achieve desired paths in electron tubes, through the study of underground fluid flow near an oil well, fluid flow around turbine blades, heat flow in a nonlinear conduction system, vibration of membranes having boundaries not of simple geometric shape, to the vibration of elastic beams. A comprehensive bibliography is given.

In addition to presenting techniques for attacking particular problems not easily approached in other ways, this book is interesting because it points out so strongly the basic unity of many phenomena of physics.—*W. J. Cunningham*

Analysis and Control of Nonlinear Systems: Nonlinear Vibrations and Oscillations in Physical Systems by Y. H. Ku: 360 pages; $10.00; The Ronald Press Co., 1958.

The field of nonlinear systems analysis, which had its start just before the turn of the century, remained dormant in the United States until World War II. Since then interest in nonlinear analysis has greatly increased. As in all new areas of interest the pioneering papers and books came first. Now the textbooks which contain systematic presentation of various methods and their application to specific problems are starting to appear.

This book, which is one of the first of the textbook class, deals with the analysis of mechanical and electrical systems. The early chapters deal with equations written in both mechanical and electrical notation. There is enough text material to acquaint the mechanical engineer with the electrical engineers notation and vice versa. Therefore, the text should appeal to both audiences. This bilingual treatment is nice but it should be emphasized that the first two-thirds of the book is a mathematical treatment of the fundamental methods of nonlinear analysis. The problems treated in this portion are: systems with a nonlinear restoring force (with and without damping), autonomous systems with nonlinear damping, forced nonlinear systems, and subharmonics. The latter one-third of the book deals with feedback systems, higher order systems, and multi-degree of freedom systems. As an added feature, there is an extensive bibliography listing the major work in nonlinear analysis from 1860 to 1957.

It should be pointed out that the graphical methods of solution are the ones most emphasized. Whenever possible, the phase-plane and phase-space methods are used either to solve the problem or picture its solution. The classic problems (e.g., the simple pendulum oscillating with large amplitude and the van der Pol equation) are well treated. Classical approaches, such as the perturbation method, are also discussed. Some of the analytical methods which are not thoroughly treated are at least well referenced. Professor Ku's book will be a welcome addition to the limited number of textbooks available in the nonlinear analysis field.—*Phil R. Cobb*

Particulate Clouds: Dusts, Smokes, and Mists by H. L. GREEN and W. R. LANE; 425 pages; $11.25; D. Van Nostrand Co., 1957.

This book is a broad survey of the field of aerosols, written in an interesting and informative manner, with emphasis on fundamental principles. Part I deals with the physics and physical chemistry of small particle systems with chapters on formation, general physical characteristics, optical properties, coagulation, deposition and filtration, sampling and measurement, and the diffusion of clouds in the atmosphere. Part II is concerned with industrial and environmental aspects in chapters on collection, health aspects, air pollution, natural aerosols, and contains a brief chapter on some uses of particulate clouds.

The book is well organized and illustrated, and contains many important and useful mathematical results. It has the unity of a textbook and provides valuable background for the experimenter. The presentation of the theory is generally descriptive rather than analytical; the critical reader and the research worker will find the lists of references at the ends of the chapters of value for a more exhaustive consideration of any of the many topics discussed.

The cumbersome title reflects the authors' reluctance to accept current usage of the word *aerosol* because this term was originally intended to apply only to relatively stable systems, in analogy to *hydrosol*. But, to this reviewer, the word seems appropriate for any system with sufficient stability to permit observations in time intervals of interest to the observer—as an extreme example, the rapid transient behavior of the dust cloud generated by a nuclear explosion. In fact, the reviewer would be inclined to waive the literal requirement that the suspending medium be air, as it is inconvenient to require another term for other gases.

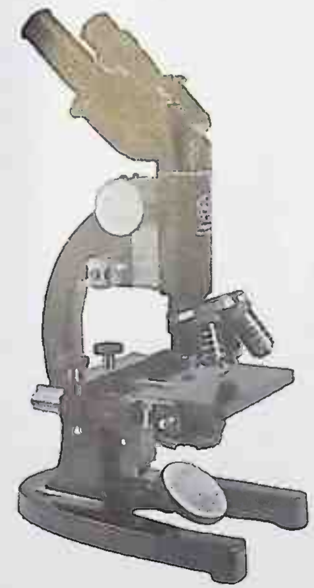
The book constitutes a helpful introduction and guide to the literature for scientists who encounter various aspects of fine particle systems. It should be most valuable to the fine particle technologist and the air pollution researcher. —*C. T. O'Konski*

Manual on Rockets and Satellites (Vol. VI of the Annals of the International Geophysical Year), edited by L. V. BERKNER; 508 pages; $25.00; Pergamon Press Inc., 1958.

This volume of reports and technical papers on rockets and satellites from the files of the International Geophysical Year contains little information that has not, by this time, appeared in various scientific and technical journals, or even in the popular press. However, the selection of some of the more pertinent material and its orderly presentation in this one volume, together with appropriate editorial comment, should be of value to many readers who lack the time or inclination to search through all of the voluminous literature on the subject.

The book consists primarily of a rather detailed series of reports on the plans and status of the I.G.Y. rocket and satellite programs as of mid-summer 1957, with additions and corrections which bring it up to around October 1, 1957. By this time a small number of high-altitude rocket firings had been made under the aegis of the I.G.Y., but little if any reduced data were available. As we now know, the U.S.S.R. was then on the eve of launching the first I.G.Y. earth satellite. An attempt is made by the editors to bring the compilation up to date by including two annexes and a number of footnotes. The first annex describes the changes in the I.G.Y. rocket and satellite program of the U.S.A. as of March 1958, thus covering details for ionospheric studies hastily put into place after the first two U.S.S.R. satellites, and also the introduction of the Jupiter C launching vehicle into the U.S.A. program. It also includes the bare announcement of the successful launching of the first (and so far the only) Vanguard satellite 1958β on March 17, 1958. The second annex is a report by the U.S.S.R., as of about February 1, 1958, on the first two Soviet satellites, 1957α and 1957β. This report is of a rather general nature and contains few details.

The reader who is looking for information about rocket and satellite launching vehicles will be disappointed, especially as regards information from the U.S.S.R.



The only specific information to be found regarding Soviet rocket technology is a not-very-detailed description of the U.S.S.R. meteorological rocket developed around 1950.

However, if the reader is interested in geophysical and astrophysical research that can be conducted through the use of rockets and satellites, or in the detailed technological background of radio and optical observations of satellites and the computations of their orbits, or in the history and organization of this unprecedented scientific operation and in the arrangements for international cooperation which were established, he will find Volume VI of the Annals well worth his study.—*Richard W. Porter*

Theoretical Physics: Thermodynamics, Electromagnetism, Waves, and Particles by F. Woodbridge Constant; 364 pages; $7.50; Addison-Wesley Publishing Co., 1958.

This text is written at the advanced undergraduate level and serves as a companion volume to Professor Constant's previous book, "Theoretical Physics-Mechanics." It is presumed the student is somewhat familiar with theoretical mechanics, vector differentiation and integration, and the differential equations of oscillations, elastic waves, and fluid dynamics.

The book is divided into three parts. Part I, Thermodynamics, also includes kinetic theory, and classical and quantum statistics. The treatment of these subjects is straightforward and competently handled. In particular the presentation of the classical and quantum statistics is excellent and should serve students well as a guide to this important subject. Part II, Electromagnetism, is the main part of the book and covers classical electrical and magnetic theory. This reviewer particularly liked the introductory chapter to this subject which brings the reader to the fundamental concepts of electromagnetism and gives a very readable, and, even more important, understandable introduction to systems of units. The rationalized mks-coulomb system of units is used throughout the book. Part III, Waves, Particles and Relativity, includes physical optics, the special theory of relativity, and wave mechanics. It is gratifying to see Schrödinger's equation introduced in an advanced undergraduate text, and solved for such elementary cases as the simple harmonic oscillator, and penetration of a potential barrier.

Professor Constant writes in a very clear, concise manner that allows him to get right to the heart of the problem without wasting space and time on unimportant details. Examples at frequent intervals throughout the text serve a useful purpose in making the student pause and think about what he has read. Well-chosen problems are at the end of each chapter, with answers given to all odd-numbered problems.

This book should be well received by students. It would serve as an excellent starting point for more advanced and specialized study. Also, since the book is well unified, it could serve, with deletion of some material, as a valuable introduction to the concepts and methods of physics to those students not intending to pursue more advanced study.—*G. D. Magnuson*

Introductory Physics: An Historical Approach by HERBERT PRIESTLEY; 515 pages; $7.50; Allyn and Bacon, Inc., 1958.

The purpose of "*Introductory Physics: An Historical Approach*" as stated in its preface "is to give the reader a sympathetic understanding of science and its contributions to the intellectual, spiritual and physical aspects of our lives, and of the way scientific work is done."

The book is intended for the nonscientific student in the last years of high school, or first undergraduate years in university. It should be interesting reading for laymen of all ages.

The contents consist of a nonmathematical survey of physics, the material covered being roughly the same as in most elementary texts of "College Physics"; mechanics, sound, thermodynamics, electricity, light, modern physics. The emphasis is on the development of ideas, and extensive use is made of quotations from scientific works from the time of Aristotle to AEC reports on nuclear fusion. In addition, an introductory chapter outlines very briefly the philosophy of science, while the last chapter attempts to assess the impact of science on society, present day politics and religion.

The historical approach indicates clearly the sense of continual growth of physical science, illuminating the many and varied persons who have contributed to present day knowledge in this field. As might be expected, the approach is perhaps somewhat more successful in those branches of physics whose most active development lies some distance in the past, approaching more usual treatments in more recent events.

So this volume fills a definite need. It should prove interesting to those members of Sigma Xi whose duties include acquainting young nonscientific students with the ideas and philosophy of science and the development of one of its branches, physics.—*E. R. Walker*

Observation and Interpretation, A Symposium of Philosophers and Physicists. (Proceedings of the 9th Symposium of the Colston Research Society held in the University of Bristol, April 1957), edited by S. Körner, in collaboration with M. H. L. Pryce; 218 pages; $8.00; Academic Press, 1958.

Quantum mechanics, in its current

mathematical formulation, is characterized by a most interesting, although somewhat peculiar, state of affairs. Although it is almost universally accepted as the correct theory to describe non-relativistic atomic and molecular phenomena and there is general agreement on how to extract experimentally verifiable deductions from the theory, there exists, however, no such agreement among the practitioners of the art concerning its interpretation. Einstein, for one, never accepted the statistical interpretation because such an interpretation for him denied the possibility of analyzing the world in terms of distinct, objective elements, existing independently of whether or not we observe them. More recently, other physicists, including Schrödinger and de Broglie, have attacked the widely accepted interpretation of quantum mechanics based on the Bohr principle of complementarity. In the interpretation by the Copenhagen school (Bohr, Heisenberg, Dirac), as well as in the recent criticisms of this interpretation, the act of measurement plays a central role. In particular, the discontinuous and irreversible change of the wave function describing the system which is brought about by the act of measurement has always raised questions concerning the interpretation of the theory; namely, is it the interaction between the measuring apparatus (usually a macroscopic system which is assumed to be describable in classical terms) and the microsystem under observation which brings about this discontinuous change, or is this collapse (reduction) of the wave function accomplished by the observer reading the pointer on the measuring apparatus who thus selects from the possible occurrences the actual occurrence. In a possible extrapolation of this latter view, an event such as the decay of a radioactive nucleus in a photographic plate would remain indeterminate until the plate is developed and scanned by an observer. The question as to whether the decay has occurred or not thus becomes a purely subjective matter. An understanding of these aspects of the quantum theory is of importance, not only for the further development of the theory in the relativistic domain, but also for a clarification of the limitations inherent in a quantum description of the world. These questions in the quantum theory of measurements are thus also questions belonging to the philosophy of science and in particular to epistemology.

The book under review, a record of the proceedings of the Ninth Symposium of the Colston Research Society which brought together eminent philosophers and physicists to discuss various facets of quantum mechanics and its philosophic implications, is a valuable document to anyone interested in these fundamental questions. In their talks, the main speakers each discuss an important aspect of quantum mechanics: Braithwaite and Ayer the conception of probabilities as logical relations, Kneale certain semantic questions concerning perception, Gallie and Ryle prediction and inference, Fierz, Körner, and Polanyi the philosophy and methodology of science, Rosenfeld and Popper the foundation and interpretation of the quantum theory, Fayerabend, Sussman and Groenewold the quantum theory of measurement; Bohm, Vigier and Bopp alternative interpretations of the formalism of quantum mechanics. It should be added that in almost every case this is done in an extremely clear fashion. To the reviewer it is, however, the record of the discussion at the end of each session (which takes up as much space in the book as the addresses) which proved most interesting. For, it is in the course of these discussions between the participants that the differences between various positions are often emphasized and brought out in relief (e.g., the differences between the Copenhagen interpretation and positivism, as emphasized by Rosenfeld) and the main problems brought into sharper focus.

The book is highly recommended to the practicing quantum mechanician steeped in the Copenhagen spirit for a glimpse at a different possible interpretation, to the philosopher of science, and to everyone interested in the epistemological implications of the existence of the quantum of action.—*S. S. Schweber*

The Medical World of the Eighteenth Century by LESTER S. KING; 346 pages; $5.75; The University of Chicago Press, 1958.

On the dust cover this book is stated to be a "sprightly journey back to the eighteenth century"—a very apt description. While packed with information, it is interestingly and in places amusingly written. The author says that the 18th century represents perhaps the adolescence of present day medicine, and this is why he has chosen it as his study. The topics discussed show that during this era medical science was emerging from the trammels of political and religious dominance, unhampered by any external pressure.

Although the author has presented more the medical scene than the individual, it has been impossible not to mention some of the actors in this period. Boerhaave, the most prominent physician of the century, has two chapters to himself. The dispute between apothecary and physician, the existence and practice of quacks, the rise of homeopathy, the classification of disease, medical ethics, the rise of modern pathology, featuring especially Morgagni, John Hunter and Bichat, are among the subjects discussed. A final chapter is entitled "The Practice of Medicine." Treatment was, of course, in most cases useless, but the case histories make interesting reading and show that these old doctors, if poor on treatment, certainly were excellent in observation. A case of phthisis is accurately described, as is another which was almost certainly thrombocytopenic purpura. This case was "cured" with "rhabarbarate purges, easy stomachic chalybeates, elixir of vitriol, Pyromont water with proper diuretics." Apparently the human frame was as hardy then as now.

Your reviewer has said enough to indicate that this is an informative study, and a must for all who are interested in medical history.—*J. T. Irving*

Dr. W. C. Röentgen (2nd edition) by Otto Glasser; 169 pages; $4.50; Charles C Thomas, 1958.

This little book deals mainly with Röentgen's discovery of x-rays and the years following. There is a brief but completely adequate history of the first 50 years of Röentgen's life. We learn that he was already a scientist of some repute before he made his accidental discovery of x-rays. The reader will be quite impressed with Röentgen's experimental ability and logical mind, particularly in his three papers on x-rays (December 28, 1895, March 9, 1896 and March 10, 1897) included in their entirety. The x-ray expert and the lay reader both will be amazed at the wide scope, completeness, carefulness and significance of his measurements. He practically "cleaned out the field," there being very few further important advances in x-rays until 1912 when x-ray diffraction was discovered. He was not famous by a chance discovery, he was a superior scientist. Not only was Röentgen great in science but he was great in humility for he refused to accept nobility.

The book is well written and can be read with profit by the nonscientist, although only a physicist can perceive how brilliant were the Röentgen papers on x-rays. There are appropriate quotations from newspapers, letters and documents. The publicity and excitement following the discovery are made clear; no prior discovery had ever aroused such widespread interest.

The names of the x-ray apparatus are not familiar to the modern x-ray physicist but this is not really essential information except to a historian. The lay reader will miss a few more points and he may also be fooled by some of the inaccuracies in the newspaper quotations. But the book will provide a genuine insight into W. C. Röentgen, man and scientist, for the general reader.

Useful appendices include a list of Röentgen's scientific papers, a chronology of events, and name and subject indices.—*Leonard Muldawer*

Geophysics and the IGY: Proceedings of the Symposium at the Opening of the International Geophysical Year, Monograph No. 2, edited by HUGH ODISHAW and STANLEY RUTTENBERG; 210 pages; $6.75; American Geophysical Union, Washington, D.C., 1958.

This volume assembles a group of papers presented at a symposium held just before the opening of the International Geophysical Year (IGY) in 1957. The material covered ranges from "whistlers" (audio frequency radio waves arising from lightning flashes) to cosmic ray studies. The papers are grouped under the general headings of "Upper Atmosphere Physics," (ionosphere, magnetic field, cosmic rays, aurora, and satellite programs); "Lower Atmosphere and the Earth" (meteorology, oceanography, seismology and gravity); and "Polar Regions."

In view of this sampling of disciplines it is inevitable that certain overlaps have occurred, especially in the areas of ionosphere, aurora, and geomagnetism. However, in general each paper makes its contribution, and the reader interested in a particular field will find a useful historical introduction (brief) that will make the findings of the IGY more meaningful to him. One might wish that a sentence or two had been devoted to indicating the pertinent experience of the author of each paper, since this must vary among the contributors.

The outstanding characteristic of the IGY—simultaneous observations over wide areas of the earth—is of course emphasized throughout the collection. One of the most striking examples of this is the coordination of observations in the Arctic and Antarctic in order to test theories of the effects of nuclear particles from the sun on aurora, earth magnetism and weather. In other areas, observations in one discipline (for example, solar flares) will be related to wide-spread observations in another (primary cosmic ray particles), and the results applied in an attempt to remove long-standing uncertainties introduced by isolated observations of the past.

By virtue of the organization of the book, the considerable historical introductions given by several of the authors, and the widespread interest in the IGY among scientists of many disciplines, will prove this to be an interesting and useful collection of papers.—*George T. Reynolds*

The Prentice-Hall World Atlas, edited by JOSEPH E. WILLIAMS; 122 pages; $9.00; Prentice-Hall, 1958.

This atlas is a dandy! It is small (8½ by 12½ inches), light, and convenient to use but with amazing coverage. It is designed for courses in geography; especially economic geography, but will be found useful for all general purposes.

The book is planned in three sections. First is the systematic geography of the world with Polar projections, Pacific ocean projection, world political, world population, religion, climate, soil, geology, vegetation, and agriculture. There follow maps of economic resources and products and of the continents, with maps of their parts such as eastern U.S., western U.S. etc. There is a good map of Great Britain and western Europe, the Baltic lands, southeastern Europe, U.S.S.R., China, India, ending, of course, with Australia and New Zealand.

Now for some general observations. The merged-color-shaded relief technique is highly successful to give a three-dimensional effect. These maps were prepared in Germany under the direction of Dr. Hugo Eckelt. The press work shows no overlapping of plates. The economic maps show actual locations of such things as mines or oil fields as well as per cent of production by countries.

Since a reviewer is supposed to pick out a couple of little things to "crab" about, I would have liked the type of projection shown on each map (projections are mentioned in the preface where of course no student would ever see it); caption and legend of U.S. geology page 26; index more specific as to locations, for I had a hard time to find a couple of places which I picked at random.

It is a mighty fine little school atlas and should have wide use, even though the price is a bit steep for the average student.—*Paul MacClintock*

The Measurement of Colour by W. D. WRIGHT; 263 pages; $10.75; The Macmillan Co., 1958.

Progress in the quantitation of color since 1944 is shown by the modernization of instruments and broader application, notably in color reproduction in printing, photography and television.

A review of light sources leads into a good summary chapter on vision with a mention of color perception and color vision defects. The history of the 1931 C.I.E. system of color measurement recalls how it just grew up and was put together from information then available, as well as how useful it has proved for many applications. Wright stresses the advantages of dominant wave length and purity for the specification of color. Other experts will point out an inconsistency in this.

Colorimeters and spectrophotometers are discussed under additive, subtractive, photoelectric and visual instruments. Instruments more nearly imitate the behavior of the eye and are now quite adequate for some applications. The problems of color spacing and the making of color atlases and solids are stated, followed by applications of color mixture data to color reproduction and colored products. Appendices include tables of: energy distribution of C.I.E. Standard Illuminants, chromaticity of the spectrum, the distribution of coefficients, and for the selected ordinate method for calculating tristimulus values.

Recent criticism and the question of a revision of the Standard Observer place emphasis on Wright's discussion, for he was one of the contributors to its development. A revision could bring it nearer to the visual sensitivity of the human eye, but any revision would also require change in industrial use. Wright's discussion of these problems merits very careful consideration.

The book is nearly the same size as the first edition and limited to a clear presentation of color measurement. The preface states that other material is available in the several recent books on color. Yet one regrets that so outstanding an authority did not discuss second order theories for color vision and the elaborate, advanced systems of color mathematics that have appeared recently.

Wright's book is recommended to all people concerned with seeing and with the reproduction of color. His style is clear, interesting, easily readable and should be required reading for beginners in the field of color.—*Oscar W. Richards*

Combination of Observations by W. M. SMART; 253 pages; $6.50; Cambridge University Press, 1958.

This book is full of information and will be useful to research workers primarily interested in the techniques of the traditional theory of errors. It does not touch at all seriously on the question of when such techniques are appropriate.

After disposing in Chapter 1 of various preliminaries from descriptive statistics, and evaluating certain integrals, the author comes in Chapter 2 to his central theme. This is to give an "acceptable precept" by which to determine the value best representing a given set of observations, and to ascribe some "ascertainable degree of precision" to the given value. Chapters 2–4 then provide various arguments to show that these purposes are answered in some circumstances by the sample mean as representing value and the root-mean-square error as measure of degree of precision. In Chapter 5 the author considers the effect of assigning different weights to the observations, while the next chapter deals with solving approximate linear equations ("equations of condition") for the "most probable" values of the unknowns.

The last three chapters branch out from these roots. Chapter 7 treats general curve fitting, with Gram-Charlier series and certain solutions to Pearson's differential equation; the next chapter brings in the problem of making corrections in observed data; and the last takes up correlation and the normal distribution for two and three variables.

It is plainly no part of Professor Smart's purpose to consider the decisions that might depend on his estimations, nor to discuss how one might assess various estimation procedures in terms of a "loss function," for example. That he did not even provide a context in which such questions can properly be put forward is clear in the meaningless introduction he gives to his discussion of probability (which, of course, lies beneath all such problems). Here he offers in turn two "definitions": (a) the probability of an event is m/n if the event "is expected to occur m times in n trials," and (b) (in case you hadn't let that one slip by) its probability is p if it occurs pN times in "a large number, N, of trials."—*Robert L. Davis*

Naven: A Survey of the Problems Suggested by a Composite Picture of the Culture of a New Guinea Tribe drawn from Three Points of View (2nd edition) by Gregory Bateson; 312 pages; 28 plates; $6.00; Stanford University Press, 1958.

As an intellectual exercise examining the ramifications within a New Guinea society of a highly formalized relationship of persons and their sister's children this has long been an outstanding work. It has highly deserved republication. Unfortunately, portions remain confusing and unnecessarily difficult to understand.

It is interesting to find that the author, although trained in psychiatry, and handling material in which sexual behavior is important, does not take a Freudian "approach" in the later edition. Less understandably he has made no effort to bring the ethnographic data up-to-date. Anthropologists will not discover here how the *naven* relationship has been altered, if it has, in the last twenty years. They will wish that Dr. Bateson had followed the example of his former collaborator, Dr. Margaret Mead, who did this for her report on the people of Manus. There are few other serious criticisms, however. The editing has been excellent and the illustrations are clear, interesting and well-chosen. This is an important book but one, definitely, for the specialist.—*Robin A. Drews*

Passive Network Synthesis by James E. Storer; 319 pages; $8.50; McGraw-Hill Book Co., 1957.

It is difficult to compress a subject such as passive network synthesis into a mere 319 pages. It is not surprising therefore to find Dr. Storer's compact book suffering from certain deficiencies which perhaps could have been avoided if the size of the book had been increased.

The book under review is primarily a collection of numerical examples illustrating some of the modern synthesis techniques as well as some of the image-parameter methods. These numerical examples are clearly presented and carefully selected and represent the chief asset of the book; in addition, the final fifth of the book is devoted to a reasonably adequate discussion of the rational-fraction approximation problem including time-domain synthesis. Unfortunately, the numerical examples are connected by an inadequate theoretical discussion which is occasionally equivocal and sometimes false. In particular, the definition of a minimum-phase network varies throughout the book. The notation used is quite cryptic and frequently inconsistent. Some of these deficiencies are, of course, a consequence of the severe space limitations, but many of the inconsistencies and ambiguities could have been detected and eliminated by a careful editing and revision.

Apparently the author is cognizant of the theoretical limitations of his book since he is careful to describe it in the preface as merely a survey; however, the reviewer considers this term a regrettably popular euphemism. The book should be more accurately described as a collection of numerical techniques suitable for the circuit designer who does not insist upon a thorough understanding of the basic principles underlying these techniques. As such, the book is well-conceived and executed and should meet the long-standing need for a compact, practical exposition of modern passive synthesis techniques.—*Paul Slepian*

Principles and Applications of Random Noise Theory by Julius S. Bendat; 431 pages; $11.00; John Wiley and Sons, 1958.

During the past ten years, there has been considerable activity in the field of random noise theory. Dr. Bendat's book is one of several texts dealing with random signals that have been published recently.

The book contains ten chapters, the first three of which treat basic random noise theory. Other texts in the field also cover the same basic material, but two differences in presentation are notable in Dr. Bendat's work. First of all, time and frequency domain descriptions of random processes are introduced before probability theory is covered. Second, ensemble averages are employed wherever possible in order to make the results applicable to nonergodic processes, as well as stationary, ergodic processes. It is thought that the treatment of basic set theory and operations with sets was rather short in the chapter concerning probability theory.

The methods of Wiener, Bode, Shannon, and Floyd for optimum linear prediction and filtering are treated in chapter four. Filter realizability is also examined. Chapter Five contains several physical examples of the occurrence of exponential-cosine autocorrelation functions and combinations thereof. Analog computer techniques for predicting rms ensemble errors of constant parameter and time varying linear systems are presented in Chapter Six.

The last four chapters contain advanced work, some of which had not previously been published. Errors in the measurement of autocorrelation functions, as well as the classical problems associated with the zero crossings of a random noise, are presented. The linear filter theory in Chapter Four is extended to cover time variable filters. Problems in envelope detection and envelope correlation are also treated.

In general, the book is well organized, well written, and fairly easy to follow. Each chapter is followed by a short list of references and a complete list is included at the end of the book. The documentation within the text is also excellent. Usable engineering tables are also given in several instances. One drawback as a course text, however, is the absence of problems for the student to solve. However, there are many solved examples within the text. Noise from electron devices, such as vacuum tubes, is not treated in detail because the author's theme is the basic mathematical theory behind noise problems. The book should be a good reference text for the graduate student and the practicing engineer.—*Theron Usher, Jr.*

Living Resources of the Sea: Research and Expansion by Lionel A. Walford; 321 pages; $6.00; Ronald Press Co., 1958.

The author, as Chief of the Branch of Fishery Biology of the U.S. Fish and Wildlife Service, treats the problem of more intelligent utilization of marine biological resources from a broad background of work with commercial fisheries. He has the experience and the insight to see many of the difficulties which hinder the practical exploitation of the world's fisheries and he is actively aware of the biological unknowns. His background, coupled with a sympathetic interest and understanding of the problems facing the fisherman, the fisheries biologist and the director of fishery research, adds importance and wide appeal to this book.

The writing is never dull and the statistical material has been selected with care and restraint. The many figures give a world-wide picture of the fisheries and of conditions which affect fishing. A short list of references includes titles of especial significance to the theoretical and practical exploitation of the biological resources of the sea.

The author emphasizes our lack of knowledge of the environment and behavior of marine animals. Attention is paid to the need for careful scientific experimentation with the design of fishing boats and gear. The practicality of harvesting plankton, farming brackish waters, the role of disease and poisonous marine animals are also examined with interest. He does not attempt to make detailed recommendations for any one fishery. Supplying this specific information will require much more research which is strongly recommended. Walford concludes, "If we really desire to exploit the sea fully, if it is knowledge that we need to accomplish that purpose—and that is the theme of this book—then we had better make the necessary costly investment and put full effort into the job of acquiring that knowledge."—*E. Lowe Pierce*

Marine Ecology by Hilary B. Moore; 493 pages; $9.50; John Wiley and Sons, 1958.

The scope of the book is wide, as suits the science which spans the oceans. The author draws many references from England, Bermuda and America, which is a reflection of his own background. The contents cover the physical, chemical and biological environmental factors, habitats and organisms.

Few principles are presented and the treatment tends to level out many of the salient features which might have been given more emphasis. A chapter once read was difficult to summarize. But, in this large and expanding field, it is no small task to gather together the material for the first textbook of marine ecology. It is to Moore's credit that he has had the initiative to do this. In his closing paragraph he writes, "The understanding of the forces and processes involved in the system constitutes ecology; it is still in its infancy." The present text is a commendable contribution towards our developing knowledge of the sea.—*E. Lowe Pierce*

Enzymes by Malcolm Dixon and Edwin C. Webb; 782 pages; $16.00; Academic Press, 1958.

Dixon and Webb are Reader in Enzyme Biochemistry and Lecturer in Biochemistry respectively at the University of Cambridge. The senior author was among the pioneers at the University who helped to usher in the modern era of enzymology. Recent books in this field have tended to be bare outlines for beginning students or four-volume encyclopedias designed for reference alone. However, these authors have prepared a book of intermediate size which not only discusses general principles of enzymology but is also sufficiently comprehensive and detailed to be valuable to research workers. The subjects covered are those of enzyme isolation, reactions, kinetics, specificity, mechanism of action, inhibitors, and a variety of more biological problems, such as enzyme formation, the integration of multienzyme systems, and enzyme distribution in cells and tissues. Problems of enzyme kinetics and mechanism of enzyme action are accorded extensive discussion.

An interesting and novel contribution is a list of the 659 enzymes recorded by mid-1957. Each enzyme is indicated with important biological source or sources, the reaction catalysed, and prosthetic groups and essential cofactors. A spot check of ten enzymes so listed, revealed a high order of accuracy and relevance in the information afforded. The authors have evidently sifted an enormous literature most successfully to have produced the useful bibliography of 2353 references.

While the largest part of these references are directed to primary sources, the authors occasionally are compelled to refer to less useful reviews or abstracts. This may be a dangerous practice. For example, in recording data on the so-called pyruvate dehydrogenase and ketoglutarate dehydrogenase which are described as producing free acyl lipoates, reference can only be given to a secondary source. The listed description of the reactions was only assumed and not proved by early workers and indeed it is now believed that if acyl lipoates are formed at all they occur only on the enzyme surface, perhaps as lipoamides. The authors' treatment of these α-keto acid oxidases in effect eliminates them from consideration as teams of catalysts, structurally organized to fulfill a reaction sequence, when they are probably an outstanding example of such integration. The authors' interpretation possibly bears on their hesitancy in later sections in accepting the concept of nutochondria or other cell particulates as organized mosaics of enzymes. They suggest alternatively that such particulates may only be enzyme concentrates in which the transit times of reactants between enzymes have been minimized by relatively nonspecific devices other than structural integration.

The book could also be useful as a text in an advanced course in biochemistry, if the price does not in fact eliminate such use. The various subjects are developed clearly without assuming advanced training, and the detail so important for the specialists is most frequently organized into tables or sections which can be by-passed temporarily. However, one wonders if the organization of the book could not have been improved somewhat with respect to its possible use as a text. Thus, the authors hope to assist research workers and students in reaping the harvest which is augured by the present development of enzymology, a harvest which is already being obtained at the level of investigation of enzyme activity within biological units and at the level of the physicochemical study of the interactions of enzymes with substrates, cofactors etc. Nevertheless, we find a relatively limited and belated treatment of an enzyme as one or more folded polypeptide chains whose properties such as binding sites and active centers are becoming one major focus of enzymology. The biology of enzymology is treated more fully, but is also discussed at the end of the book. One might ask, for example, whether the recent knowledge and method of study of enzyme distribution in cells described in Chapter XIII should not be useful in enzyme isolation, which is described in Chapter III entirely as an empirical operation. It is my feeling therefore that the organization of the book is more characteristic of the history of enzymology than of its modern status and of its perspectives. Nevertheless, since the main facts and problems of enzymology are presented clearly and thoughtfully, the book should contribute to the training of advanced students, as well as to the activities of the research biochemist. —*Seymour S. Cohen*

Introduction to Nonlinear Analysis by W. J. Cunningham; 349 pages; $9.50; McGraw-Hill Book Co., 1958.

This book presents many of the techniques which are available for the solution of nonlinear differential equations. The equations considered are those whose analytical solutions in closed form may be easily found, and those which are amenable to hand calculation or graphical analysis. A chapter is included on nonlinear differential-difference equations, and also a chapter on linear differential equations with varying coefficients.

Basing this volume on a course for electrical engineers which he has given at Yale University, the author assumes only a knowledge of electrical circuit theory and mechanical vibrations on the part of the reader. The emphasis is placed on the use of the mathematical techniques as a tool for solving engineering problems, rather than on presenting the techniques with their full mathematical rigor. An extensive bibliography is included, along with a discussion to guide the reader to those references which pertain to particular methods in each chapter.

The material includes: numerical methods amenable to hand calculation; graphical methods for differential equations of first and second order; equations with known exact solutions; analytical methods such as perturbation, reversion, variation of parameters, and harmonic balance; iteration and perturbation procedures for forced oscillating systems; and a discussion of stability of nonlinear systems. Attention is given to the feedback configuration with and without delays in the loop.

The book is very well written, and the individual chapters can essentially stand on their own. The book should be quite useful for instructional purposes, and provides a welcome summary for the great amount of material now available on nonlinear methods.—*Preston R. Clement*

Index to Advertisers

Please mention the AMERICAN SCIENTIST when you are writing to an advertiser