VOLUME 2 □ NUMBER 3 MARCH 1965

IEEE spectrum

features

the cover

Visible representations of language are not new, but old signs (hieroglyphs tooled into limestone millennia ago) and new signs (representations used as tools of modern speech research) serve different functions. One offers a prayer, the other produces speech sounds from a machine (see page 115), yet both share a clear elegance. The detail of the Stela of Ma'ety tablet (p. 114), XI Dynasty, courtesy of The Metropolitan Museum of Art, Rogers Fund, 1914.

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE Convention

departments

Spectral lines

On Being a Nonnational Society. Nonnational is one of the significant terms which is used in the IEEE Constitution to describe the character and scope of the Institute. Some of the implications of the fact that the IEEE is not a national, nor an international, but a nonnational society are worth exploring.

A national professional society would presumably confine its membership to citizens or residents in a particular country with some provision made for foreign members. An international society would be an organization whose members were representative of particular nations. The International Electrotechnical Commission (IEC) is an example of such an organization; its members represent national committees from the participating countries.

In contrast the IEEE is an organization whose membership is not organized on national bases. This fact does not mean that there are not geographical divisions and subdivisions within the IEEE organization. Such divisions are needed to provide the necessary framework for Institute activities but the geographical boundaries between these divisions are drawn in such a way as to maximize the service the Institute can provide its members; they may or may not follow national boundaries.

The objectives of the IEEE are such that it is appropriate for it to be organized on a nonnational basis. Scientific and technical contributions are made by particularly able individuals and groups, and as the history ofs cience and technology has amply demonstrated, such groups are found in different countries at different times. Thus a professional society whose objective is the advancement of its technical field should preferably be nonnational.

But how nonnational is the IEEE in practice? It is certainly true that most of its members live in the United States and much of its activity is concentrated there. However, the activities in other countries are at present very sizable as evidenced by the fact that there are now seven Sections in Europe, one in Asia, one in Africa, and five in Latin America. It should also be noted that several new Sections will probably be organized in the near future in these areas. The total membership outside of the United States and Canada is now 13 610. In addition, the 1965 Board of Directors includes Regional Directors who reside in Paris and Tokyo.

While recognizing these significant indications of the global activities of the IEEE, it is also clear that much more needs to be done before the IEEE can be considered a truly nonnational society. This fact was recognized at the first meeting of the 1965 Board of Directors when a motion was unanimously passed requiring the formation of an ad hoc committee to examine the activities of the IEEE in countries other than the United States and to recommend how these activities might be improved.

In this connection, it is particularly important that the publications of the Institute not be parochial in either their content or makeup. Constant attention is needed on the part of the Editorial Board to make sure that the editors of the various IEEE periodicals are aware of the impact of various editorial features and practices on members or subscribers from differing cultural environments.

Committees which carry on the important works of publishing standards and organizing conferences need to be aware of the impact of their activities in countries other than the United States and Canada. Of particular importance is the work of groups such as the Awards Board and the Fellow Committee, which are responsible for recognizing the technical and scientific contributions of IEEE members wherever they live. In many cases, such boards and committees have found it very desirable to include in their membership persons who reside, or have professional acquaintances, outside of the United States and Canada.

The Headquarters staff must learn enough about the postal regulations, customs regulations, and social mores of the various countries in which IEEE members reside so that Institute business can be carried on efficiently and effectively. Individual members must also keep in mind the global scale of the Institute's activities.

The joint AIEE–IRE Committee that drew up the merger agreement which resulted in the formation of the IEEE in 1963 showed both wisdom and foresight in many ways. Perhaps their choice of the term nonnational to describe the scope of the IEEE was one of their wisest decisions.

F. Karl Willenbrock

Machine recognition of human language

Part I—Automatic speech recognition

After many centuries of sporadic interest in the nature of speech, the past 20 years of speech research stand out as being particularly intensive. But despite many illuminating discoveries, the physical realization of automata that will recognize natural speech seems still far away

Nilo Lindgren *Staff Writer*

It should not be necessary to stress, for this audience, the uses, social values, meanings, or mysteries of human speech. Instead, we should like to take up, as quickly as possible, some "practical" questions. Researchers in this century have become seriously intrigued, for one motive or another, with the idea of making automata that can hear and understand what we humans say, and that can speak and make us understand.

There are machines now aplenty that can deal in "artificial" languages, but there are none with which a human can communicate directly in his natural tongue or in natural handwriting. However, the research and engineering aimed at making just such specialized pattern-recognition machines—automatic speech recognizers and automatic handwriting recognizers—has become particularly intensified over the past decade, and it is one objective of this present survey to make an estimate of the state of that research.

In looking into this question, we shall find, perhaps, that our opening premise is not wholly justified, and that those relatively few engineers who have become committed to the natural-language automata have been obliged to "drink deep" of the linguistic mysteries. To get any feeling for the research on speech recognition, we must, as a bare minimum, consider some of the achievements of the phoneticists, the linguists, the psycholinguists, the neurophysiologists, and others, as well as of the communications engineers.

This survey, then, may seem unreasonably long, but there is at least one legitimate cause for this. Very little of the literature relevant to this speech venture has appeared in the electrical engineering literature (it has appeared in such journals as the *Journal of the Acoustical Society of America*, in *Language and Speech*, in *Language*, in *WORD*, in the *Journal of Experimental Psychology*, in the *Journal of Speech and Hearing Research*, and so on) so that those electrical engineers who have not had special cause (and it is certainly to them that this survey is directed) may be unaware either of its existence or of its abundance.

After this warning and apology, however, we should say that those who persevere, and who go to seek out the original literature, may discover that speech and language research is as exciting and intriguing as anything going.

The place of vocoders

Electrical engineers have heard, in recent years, a great deal about vocoders (voice coders).[1] It is useful,

This article is the first of a three-part series.

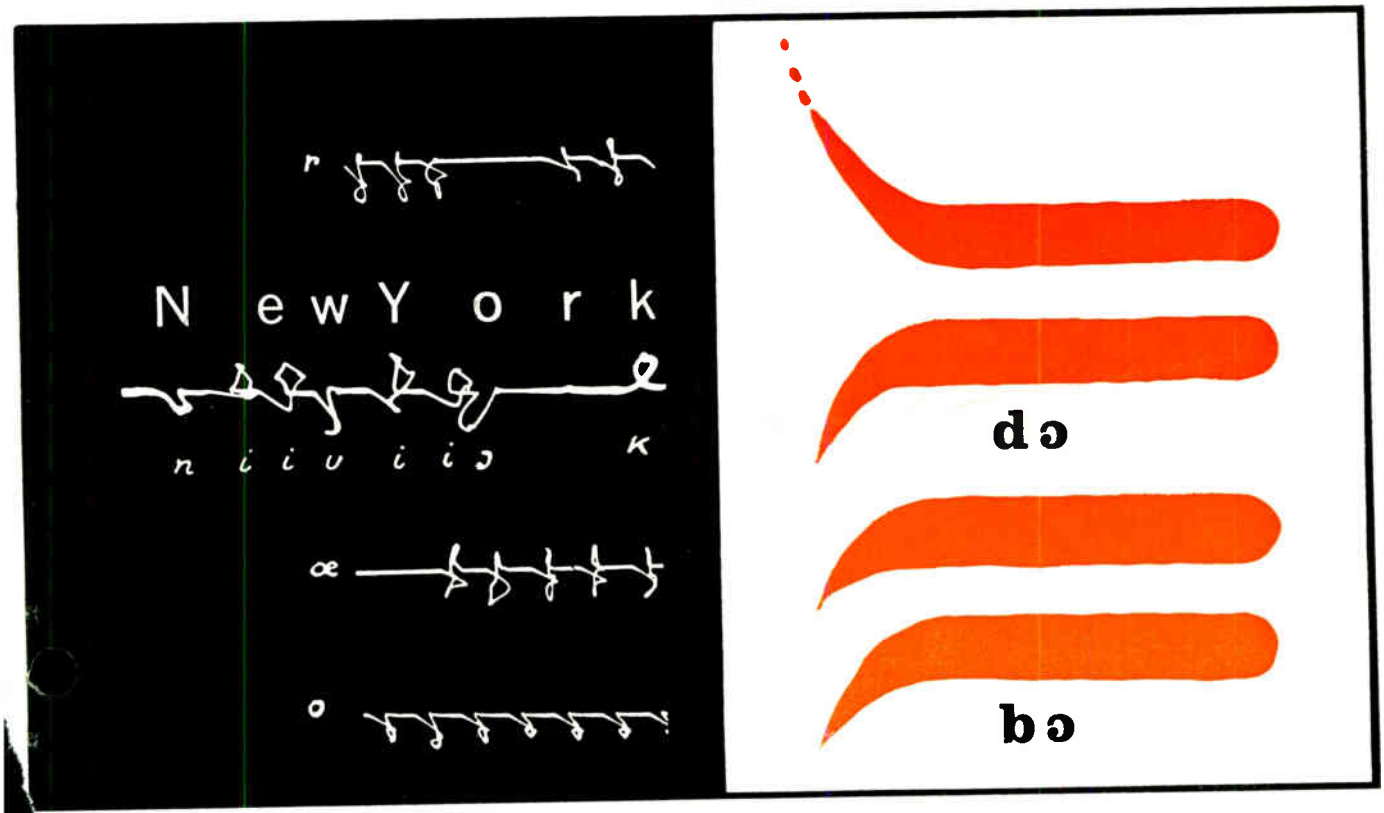


Fig. 1. Early Egyptian prealphabetic signs (left) were compared by Dreyfus-Graf to pictographs (center) produced by his speech recognizer. Today, strikingly simple hand-painted "cues" (right) can be used by machines to produce intelligible speech.

therefore, to draw immediately this distinction: these devices are basically intended as methods of economical voice communication[2]; they are not speech recognizers.

Actually, many speech-bandwidth compression systems have been developed—such as vocoders, amplitude or frequency limiters, and formant coders. These machines do not recognize speech; what they do is transmit sufficient verbal clues so that a human listener can piece together the linguistic content of the utterance.[3] A visual analogy of vocoder action can be found in abstract painting, as in, for instance, Picasso's early cubistic paintings, in which the viewer is brought to see (by an active process of composing on his part) the "natural" object embedded in the composition. In a similar manner, the output of speech-compression systems lacks naturalness, and, in fact, the cues for what constitutes naturalness in speech have yet to be singled out.[4]

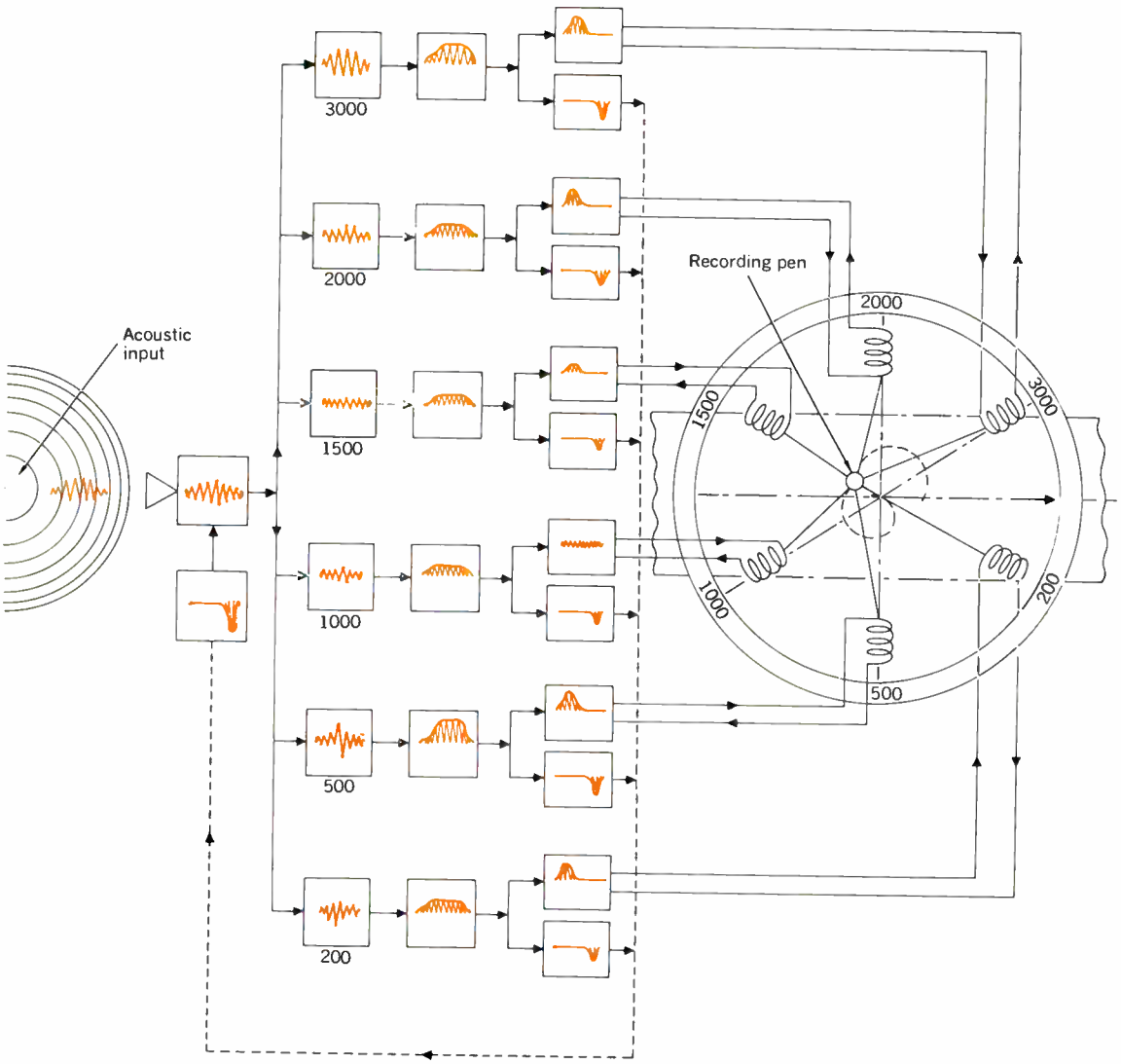
These remarks do not mean to imply, however, that speech-compression research has not contributed understanding to the automatic speech recognition problem; it certainly has. The point is only that these systems require the intervention of a *perceptive* human, which it is the aim of automatic speech recognizers to render unnecessary.

Categories and levels of speech studies

There are several broad categories of how human speech may be studied. Speech may be regarded as a sequence of articulatory events in the physiological structure. Speech may be studied as an acoustic disturbance freely propagating through air. And it may be studied as an auditory sensation.[5] Although investigations have gone on in all these categories, their objectives and their results have not necessarily been correlated and unified in a science of speech.

The human speech *recognition* processes, which the proposed machines are to mimic in greater or lesser degree, may be described at several hierarchical levels—at the acoustic, at the linguistic, and at the semantic. Such recognition may be described at other levels as well,[6] but for our purposes it is enough to know that most modern work on speech-recognizing automata has concentrated largely on the first, the acoustic level, that research has only recently begun in earnest on the second,

Fig. 2. Schematic of Dreyfus-Graf's phonetic "stenosonograph."



the linguistic level, and that questions concerning the semantic level are at this time virtually untouched.

These generalizations reflect the state of speech research today. Let us give them some substance.

Early attempts and first principles

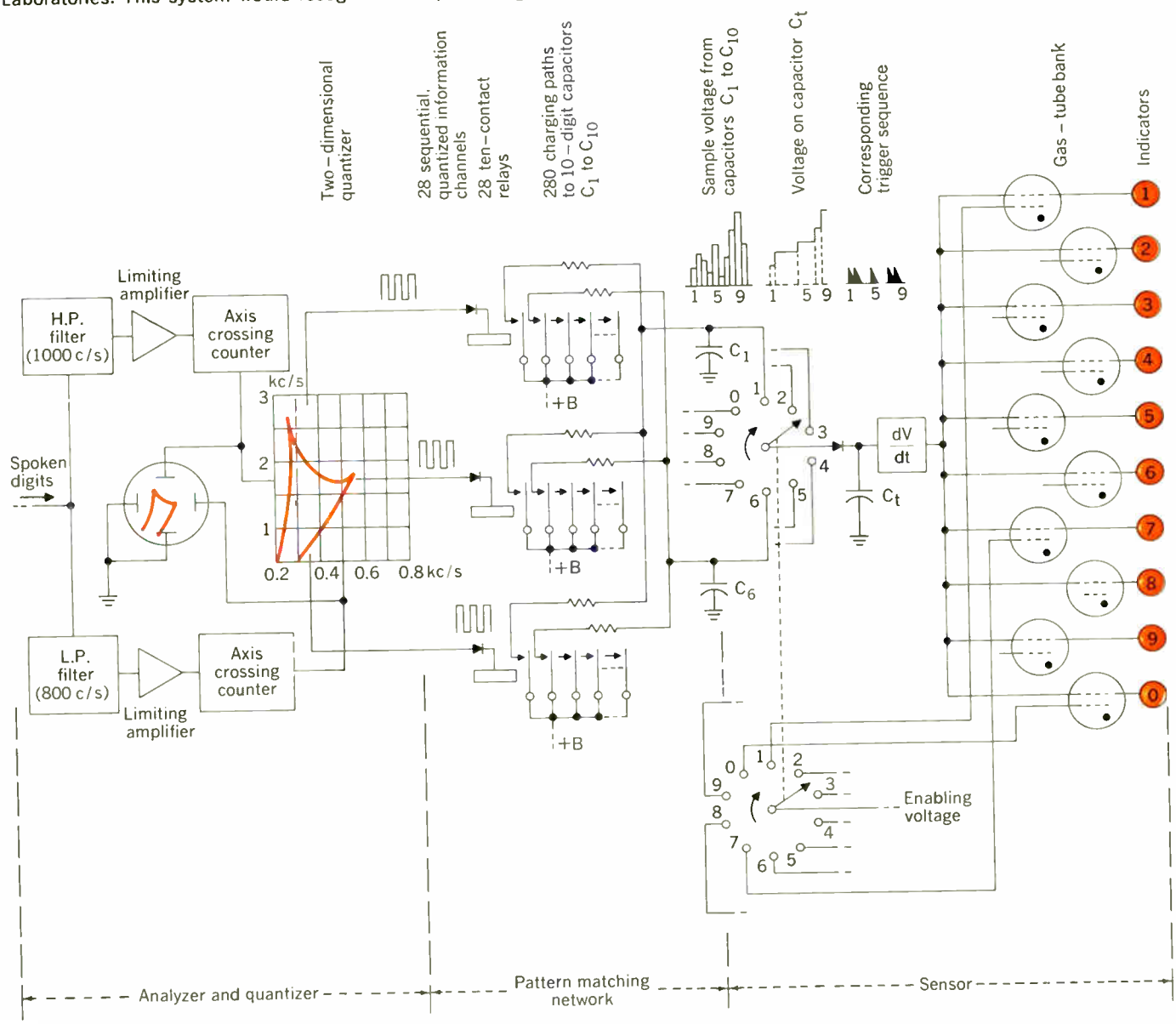
Five years ago, in an interesting survey on machine recognition of spoken words, Richard Fatehchand stated that only limited success had been achieved with speech recognition machines. No machine existed, he said, that would deal with continuous speech.[7] Today, there is still no such machine. Nor is it likely that there will be one before the end of this decade.

In the decade preceding, between 1950 and 1960, there had been developed a number of electronic machines that would recognize very limited vocabularies pronounced by particular speakers for whom the machines had been adjusted.

Probably the first automatic recognizer, or at least the first in the electronic era, was the sonograph, described in 1950 by its designer, Jean Dreyfus-Graf of Geneva, Switzerland, who had spent many years of research on the design.[8] The principal elements of his machine consisted of a microphone and an amplifier, followed by a bank of six filters for dividing down the acoustic spectrum (as he said, "to the six principal formants of the mouth orchestra"). The modified outputs of these filters controlled six deflecting coils, which in combination operated a pen recorder to provide diagrams for the input sounds. Dreyfus-Graf compared these output diagrams to the old prealphabetic Chinese or Egyptian pictographic signs, to which they were "similar in principle" (see Fig. 1). Figure 2 shows the basic configuration of the Dreyfus-Graf machine.

A more influential and more extensively tested recognizer, however, was developed slightly later at the Bell

Fig. 3. Schematic of the digit recognizer developed in 1952 at the Bell Telephone Laboratories. This system would recognize the spoken digits "oh" to "nine."



Telephone Laboratories. This system would recognize the spoken digits zero (*oh*) to *nine*. First described in 1952 by Davis, Biddulph, and Balashek,[9] it operated on a simple principle: it compared the spectrum of the acoustic input with the ten spectral patterns already stored. The spoken input digit was recognized on a best-match basis. In its implementation, this system was already considerably more sophisticated than the Dreyfus-Graf machine, as Fig. 3 indicates.

Another version of the Bell Labs system (called Audrey) was developed later in the decade (1958) by Dudley and Balashek.[10] It would recognize acoustic patterns corresponding to 16 different basic linguistic elements. In both these machines, the incoming acoustic signals were broken down into specific patterns, which were compared with patterns stored in the machine. Best-matches were determined by cross-correlation methods.

The 1952 machine, which dealt with each word as a single unit, would recognize *oh* to *nine*, spoken by an individual, with 97 to 99 per cent accuracy. Its accuracy fell, however, to 50–60 per cent, when its circuit was not adjusted for the particular speaker. The 1958 machine would recognize *oh* to *nine* with almost perfect accuracy when the circuit was optimized for a single speaker; other speakers of the same sex could, by modifying their voices, give the machine a 90 per cent chance of being right.

It should be pointed out that the recognizer built by Davis, Biddulph, and Balashek in 1952 was essentially a vowel "formant" tracker. In spoken vowels, there are concentrations of energy at certain frequencies, corresponding roughly to resonances in the tube of the vocal tract. When the lowest frequency of energy concentration is plotted against the next highest frequency for each spoken digit, the plot takes on a distinctive shape (see Fig. 4). These distinctive traces were utilized for the digit recognition. Because the regions of energy concentration are called formants, the general method of tracking the movements and characteristics of such regions is called "formant tracking." The principle of formant tracking, in differing physical implementations, has been employed in even the most recent attempts at automatic speech recognition.

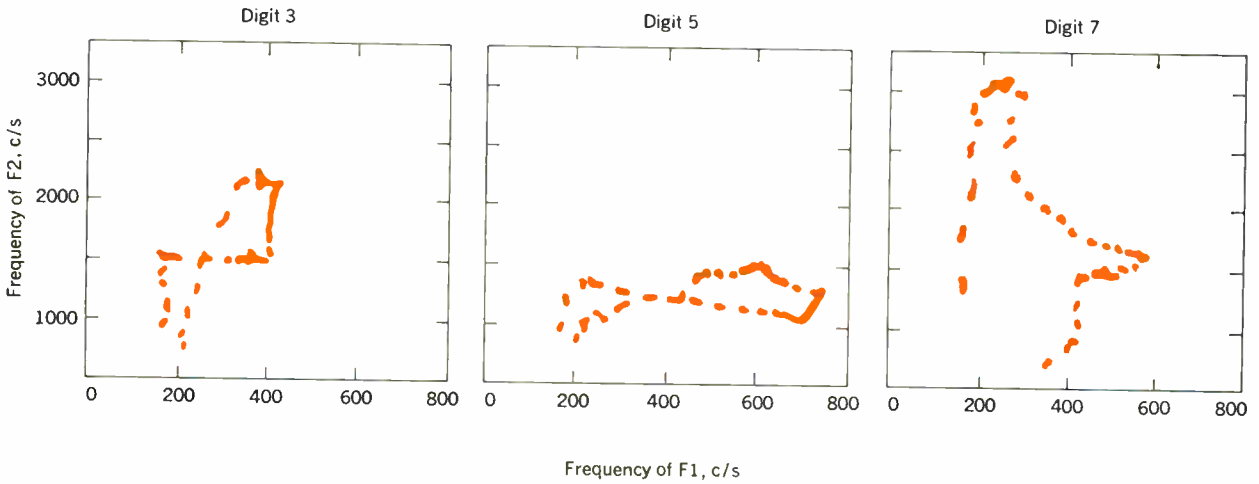
In 1956 an automatic speech recognizer based on an entirely different operating principle was designed at Northeastern University by J. Wiren and H. L. Stubbs.[11] This electronic machine was designed to sort out elementary sounds of speech (phonemes) by a process of successive binary decisions about the features or properties of the incoming signal. This system was based on the bold idea of *distinctive features* proposed originally by Roman Jakobson and elaborated by Jakobson, Fant, and Halle in 1952.[12, 13] An outgrowth of linguistic and acoustical studies, the distinctive-features approach postulated sets of features embedded in the highly redundant sounds of speech.

In the Wiren–Stubbs electronic implementation, the properties separated were the voiced sounds from the unvoiced, the turbulent (noiselike) from the nonturbulent; then the nonturbulent sounds were separated into the groups shown in the upper right of Fig. 5 and the unvoiced turbulent sounds were separated into the stops and fricatives as in the lower right. (At this point in the discussion, the reader should not worry about terminology. The important fact to carry forward is that the principle of binary classification has been applied to the selective sorting or screening out of distinctive linguistic features from an acoustic speech input.) Fairly good results were obtained from this system. For instance, for vowels in short words pronounced by 21 speakers, accuracy was above 94 per cent, which is probably comparable to what a human listener would do if he were presented with a succession of speech sounds.

The next significant attempt to be considered is a machine designed by Peter Denes and D. B. Fry in 1959.[14] Fry had suggested in 1956 that a human listener could not successfully identify speech sounds in isolated acoustic signals, and that the listener reduces ambiguities and confusions through use of linguistic information he already possesses. On this premise, Fry and Denes built a machine that incorporated certain linguistic information (this information was in the form of probabilities that one sound element would follow another—that is, how likely it is that a *t* will follow a *k*, and so on).

The principal aim of the designers was to see whether or

Fig. 4. Formant 2 versus formant 1 presentations of the digits reveal distinctive differences in shapes. Recognition depended upon these differences and upon their relative duration in the frequency space.



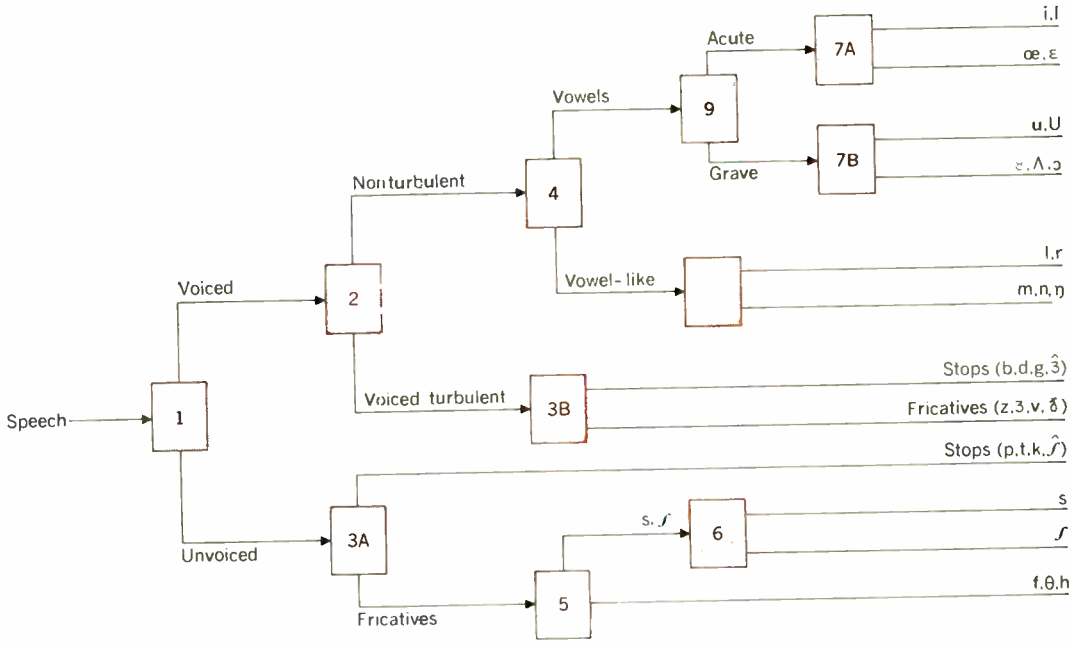


Fig. 5. Binary selection system for phoneme classification. Not all elements of this scheme were physically implemented.

not the use of linguistic information to modify the output of an acoustic recognizer would improve recognition results. They were *not* concerned with the refinement of the acoustic detector itself.

In its operation, the acoustic section (consisting of an acoustic spectrum analyzer and a spectral pattern matcher) examined the characteristics of the speech input sound wave, compared these with the repertory of characteristics stored in the machine, and made a preliminary decision. This information was then combined with statistical information from the linguistic store in a computational section (multiplying circuits, etc.), which then selected the most likely element in light of this combined information, and operated the appropriate key of a typewriter output.

The recognition repertory was limited to four vowels and nine consonants, and speech input material consisted of a list of isolated words. The linguistic data did not materially improve recognition of individual sounds, but word recognition accuracy was doubled. Whatever the interpretation, the important principle that their work projected was the use of linguistic "context."

In 1960, Peter Denes and M. V. Mathews made another kind of study involving linguistics.[15] In this case, they were, in part, trying to obviate the need for linguistic information by sharply restricting the vocabulary of the recognizer, thus, in effect, heightening the redundancy of the words to be recognized. The objective of this study was the recognition of whole words (the spoken digits zero through nine), relying only on their acoustic characteristics (by time–frequency pattern matching). The study was carried out by a digital computer simulation, with a further underlying intention of investigating the usefulness of computers in automatic speech recognition research. Their conclusion was positive: there was "little doubt that computers provide considerable advantages for solving many of the problems encountered in speech research."

I. Words used in recognition study

English words	Phonetic transcription
bit	bɪt
bet	bɛt
bot	bɑt
bat	bæt
but	bʌt
beat	bit
boot	but
bought	bɔt
Bert	bɝt
put	pʊt
book	bʊk

Other computer-based studies of speech recognition that should be mentioned at this point are those made by J. W. Forgie and C. D. Forgie at the Lincoln Laboratory. The Forgies have made a series of such studies.

One of these was a vowel recognition program (completed in 1959), which recognized ten English vowels in isolated words of the form /b/-vowel-/t/, words like *bit*, *bet*, *bot*, *bought* (see Table I), with an accuracy of 93 per cent.[16] The recognition procedure depended almost solely upon the locating of the first two vowel formants (F1 and F2), that is, upon a relatively simple use of two-dimensional patterns of amplitude and frequency. Figure 6 shows the general structure of the Forgie program.

An indication of the state of the relation between speech research and the speech recognition art at that time appears in this remark by the Forgies. "Much work has been done on the theory of vowel production, and statistics have been published on the characteristics of American English vowels, but one who attempts to design a vowel recognizer still finds that much of the information he needs must be found by trial-and-error procedures."[16]

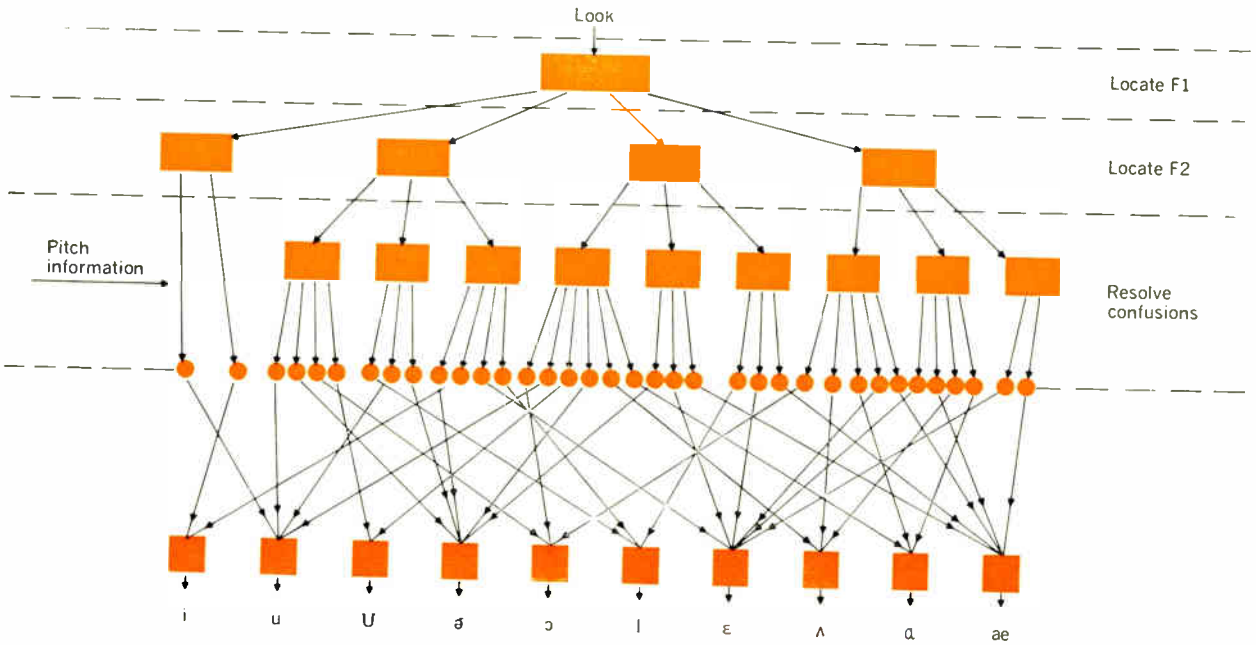


Fig. 6. General structure of the Forgies' vowel recognition program which classified each LOOK as belonging to one of the ten possible vowels.

A later computer study by the Forgies, in 1962, was more ambitious in another respect[17]—it aimed at recognizing the English fricative consonants /f/ and /θ/ in the initial and final positions in pairs of words like *fief* and *fie*, *thief* and *thigh*, *frill* and *thrill*, *Ruth* and *myth*, and so on. These fricative sounds present problems for both machines and humans because their spectra are so much alike and because their effects on adjacent vowels differ only slightly. Thus, it was necessary to rely on a number of different "cues," and in the final process of recognition to use a "voting operation" based on statistical probabilities. In this study, it was found that for final fricatives, human listeners and the computer did about equally well, but that for initial fricatives, the people did considerably better than the computer.

These latter studies, by Denes and by the Forgies, set into relief some points of interest. They both used computers, but the Forgies were using computers to pull out elementary speech sounds integrally embedded in whole words, whereas Denes was investigating, in part, the trade-off between the amount of linguistic information needed by a recognizer and the amount (the number) of words to be recognized.

To summarize, these representative early attempts, between 1950 and 1962, to devise speech recognizers all helped to crystallize certain guiding principles and concerns. They illustrated a unique relation between acoustic inputs and diagrammatic outputs, they made use of formant tracking, of pattern comparison (spectral matching), of binary-decision methods, they made use of computers, they attempted to tackle the so-called "segmentation problem," that is, to recognize *discrete* elementary speech sounds embedded in *continuous* short utterances (of word size), and they raised the question of the need for linguistic information (and its corollary in a practical situation, "*how much* context?"). Other points could be mentioned as well, depending on the direction of one's interests, but (following a principle of speech research that has been put forward more recently, namely, that the listener stores up and processes linguistic information in "chunks") this chunk should suffice for the moment.

All these early achievements, in relation to the complexities and richness of natural speech, were very limited—they dealt only with small vocabularies, isolated elementary sounds, limited numbers of speakers, utterances made in laboratory conditions, careful trial-and-error normalization of acoustic inputs—but each of these attempts gave some insight into the manifold problems of unlocking the secrets of speech coding, and of the ultimate magnitude of the problems inherent in the objective of automating speech recognition.

A shift in viewpoint

Most of the early efforts at building speech recognition machines were alike in that they dealt almost exclusively with the acoustic input signal. Throughout the period, from the mid-forties to the mid-fifties, it was the considered view of researchers that once they had found some method of analyzing acoustic signals into their basic component parts, the automation of speech recognition would quickly follow. Equipped with the basic principles of how linguistic elements were encoded into the acoustic outputs of speakers, these machines were to operate on grander vocabularies by a simple extension of their size without, hopefully, having to use giant computer facilities to carry out the necessary comparisons with the incoming acoustic signals.

But extensive research on speech at the acoustic level, in the effort to single out the acoustic cues to linguistic content, increasingly revealed the complexity of the speech process, and forced the realization that this viewpoint was far too simple. Single linguistic elements, spoken carefully by selected speakers, set apart by silent pauses, could be identified by machine, but when these elements were incorporated in continuous speech by many

different speakers their acoustic representations showed a dismaying variability—usually fatal so far as machine recognition was concerned. This raised a serious problem: how was a machine, faced with an ambiguous pattern, to know when one pattern interpretation was more appropriate than another?

Further research only made it more and more obvious that the gap between these limited word recognizers and a practical natural-speech recognizer was enormous. Natural, unconstrained speech seemed very "sloppy" to the engineering mind, and seemed almost beyond analysis. Not only were there variations between speakers in their acoustic outputs, but there were variations by the same speaker in different circumstances, in differing emotional states. All the rich, meaningful sounds uttered by humans in their linguistic intercourse—embodying multifoliate complexities and subtleties of expression—obscured the clear, quantitative picture of speech the engineer wanted.

When researchers like D. B. Fry in England began, during the mid-fifties, to stress the necessity for taking "linguistic constraints" into account in speech recognition machines, this view was received somewhat skeptically.[18] By the end of the decade, however, as research broadened its grasp on the physical nature of speech, this view began to gain acceptance, and now there is hardly a serious researcher in this field who does not begin or conclude his ideas about recognizers with the call for more attention to linguistic structure. Peter Denes says, for instance: "Automatic speech recognition is probably possible only by a process that makes use of information about the structure and statistics of the language being recognized as well as of the characteristics of the speech sound wave."[15]

Somehow, ways must be found of incorporating linguistic information in the decision-making functions of possible speech recognizers. Accordingly, the emphasis of speech research has been shifting its center. Alongside the still numerous studies of the purely acoustic factors of speech, there are growing numbers of studies of contextual factors and of the articulatory processes, as well as of human and lower-animal perceptual systems. (Meanwhile, the immediate aim of building automata that can recognize speech seems to be somewhat in abeyance.) At Bell Labs, a long-time leader in speech research, Dr. Flanagan, says, "We are not working on a speech recognizer at this time."[4] Peter Denes, now also at Bell Labs, and who admits to a definite long-range interest in speech recognizers, is now working with computer simulations of articulatory processes and their multiple parameters.

Thus, communications engineers, who approached the speech recognition problem from the information-theoretic point of view, and who had expected to reduce speech to sets of relatively simple physical measurements, have been gradually moving closer to the researchers of the other involved disciplines—not only to phoneticians, but to psycholinguists, to speech and hearing pathologists, and to pure linguists and philosophers of language as well. Such interdisciplinary spillover has, of course, become the characteristic style at the frontiers of modern research, as the researchers tear down the conceptual walls that have long persisted between the many scientific disciplines.

Nonetheless, at this very frontier many communications engineers who had been strongly committed to some aspect of automatic speech recognition seem to have

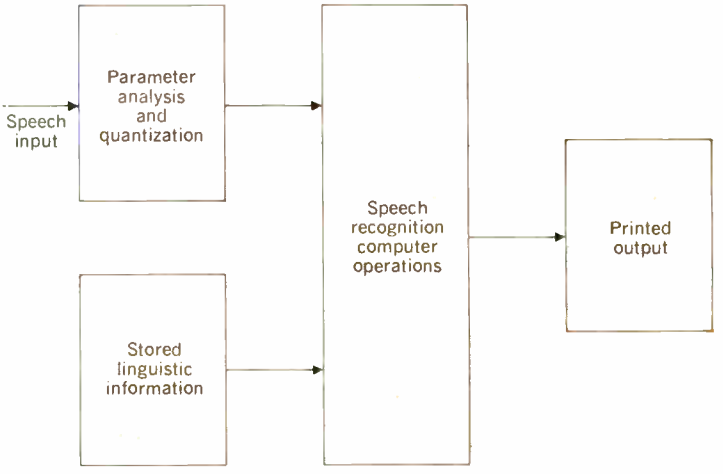


Fig. 7. Automatic speech recognition procedures must broadly take in the functions schematized here (after G. E. Peterson of the University of Michigan).

hesitated, disillusioned perhaps by the "dead-end" quality of many of their efforts to date. They remark that the problem of incorporating linguistic constraints in machines seems too formidable or intractable. Some engineers admit they have retreated into other types of research. What had seemed to them "around the corner" a few years ago has fled from their grasp. Automata that will understand natural speech seem farther away on the research horizon. The view is expressed, for instance, that the development of satisfactory statistical information on linguistics would involve "a tremendous amount of dog work"—and not dog work alone. The engineers in some cases confess that they simply do not know where to begin. "What is needed now," says one, "is a good idea." To grapple with the octopus of natural speech and natural language recognition seems to them almost too hopeless a task to undertake at this time.

Leon Harmon, also of Bell, who among other endeavors has been working for the past three or four years on automatic recognition of cursive script, takes another view. "We should consider," he says ruminatingly, "whether it is unfair of us to expect so much of the machine. Perhaps the interface between man and machine must be set at some other point to demand less of the machine."

However, not everyone is pessimistically inclined about the *eventual* prospects of automatic speech recognition, and in any event the pace of speech research has not abated. To a certain extent, our original question, "What is the present state of automatic speech recognition?" transforms itself into the question, "What is the present state of speech research?" In many respects, this is a much more interesting question to consider.

The general recognition problem

At this point in our discussion, the general problem, then, may be regarded as being composed of two major parts: a primary recognition based solely on the sound shapes of the acoustic signal; a secondary recognition of the linguistic (grammatical and syntactic) content based on the (presumably phonemic) output of the primary recognition level. These two major parts would undoubtedly

be implemented in a machine in many complex hierarchies of procedures and decision strategies.

In the final machine, it may be necessary to incorporate the faculties any ordinary listener possesses—knowledge of the meanings of utterances, rules of grammar, feelings for phonological probabilities, vast stores of general knowledge organized and codified in some form of associative system—in short, many of the interpretive faculties a listener can bring to bear on any utterance that comes into the purview of his ears. This latter portion of the recognition problem is, without question, much the bigger. Conceivably, the incorporation of such faculties in automata will depend on a deep-going investigation and quantification of the dynamic functions of the central nervous system (CNS). However, most neurophysiological investigations thus far have dealt only with peripheral events.

A general schematic representation of automatic speech recognition procedures that would take these two major halves into account is shown in Fig. 7.

In terms of this definition of the two halves of the general problem of automatic speech recognition, we shall, in this Part I, restrict our discussion to the achievements and methods of research on the acoustic level, the level of primary recognition, without which nothing else could follow. In Part II, we shall take up research on, and recent models of, the deeper perceptual processes, which includes physiological and psycholinguistic investigations, as well as studies of the structure and function of language above the level of sounds.

Terminology

Before going further into the various efforts to devise speech recognition machines, we must acquaint ourselves with some of the terminology of experimental phonetics, thus far avoided.

If a machine is to recognize speech, it must first of all, in some manner or procedure built into it, select from the "raw" continuous utterance those distinctive features, invariants, or "acoustic cues" that determine the linguistic content or the "message." Some authors speak of this selection process as a general problem of pattern recognition, in which one searches for a "recognition function" that appropriately pairs *signals* and *messages*.[19]

In handwriting, for instance, the signal is a more or less continuous two-dimensional line that forms curves, segments, and one-dimensional dots; the message is a sequence of discrete letters in a known alphabet. In recognizing the message embedded in the handwriting, the reader must rely on his knowledge of the alphabet, on certain invariant features of each letter, and he must rely on his knowledge of the language and possibly on other contextual information as well, for the signal may be noisy, i.e., sloppy or scrawly.

In speech, the signal consists of more or less continuous fluctuations of energy distribution in the acoustic domain, and the messages may be decoded as sequences, or strings, of discrete symbols called phonemes. These phonemes are viewed as the basic or elementary classes of sounds for a particular language. In English, roughly 40 such phonemic elements are distinguished (see Table II). One of the major long-term aims of research on automatic speech recognition has been to find a recognition function that relates the acoustic signals produced by the human vocal tract to these distinct phonemes. The end result of this process would be a machine that could listen to a human speaker and store, or reproduce in some symbolic form, an accurate phonemic transcription of what it heard (with the so-called phonetic typewriter, for instance).[20] Much of the experimental phonetics of the past decade and a half aimed at singling out, one by one, the significant linguistic features or cues of these phonemes embedded in the acoustic signal. Although a great deal has been learned, and many cues singled out, this work is still far from complete.

II. English phonemes

Phonetic Symbol	Key Word	Phonetic Symbol	Key Word
Simple vowels		Plosives	
ɪ	fit	b	bad
i	feet	d	dive
ɛ	let	g	give
æ	bat	p	pot
ʌ	but	t	toy
ɑ	not	k	cat
ɔ	law		
ʊ	book	Nasal consonants	
u	boot	m	may
ɜ	bird	n	now
ə̂	Bert	ŋ	sing
Complex vowels			
e	pain	Fricatives	
o	go	z	zero
aʊ	house	ʒ	vision
aɪ	ice	v	very
ɔɪ	boy	ð	that
ɪʊ	few	h	hat
		f	fat
Semivowels and liquids		θ	thing
j	you	ʃ	shed
w	we	s	sat
l	late		
r	rate	Affricatives	
		tʃ	church
		dʒ	judge

The articulation process

Almost all speech sounds are produced on the outbreath. The breath stream coming from the lungs passes through the vocal tract—the throat, mouth, and nasal cavities (see Fig. 8). This moving stream of air is acted upon by all the parts of the vocal system to create various acoustic disturbances from which a listener extracts linguistic information. The air stream first passes through an opening between the vocal cords, which are vibrating in voiced sounds, periodically modulating the stream in such a way as to produce a harmonically rich spectrum.

Complex patterns of shifting resonances are produced in this system by modifications of the size and shape of the vocal cavities through time-varying tongue and lip positions. The oral and throat cavities may or may not be coupled to the nasal cavities by the action of the valve at the rear of the mouth, called the velum. Turbulence (noiselike sound) is produced by the movement of the air across the edges of the teeth, and by partial closure of the vocal cords. In actual speech, these physical articulators are rarely stationary, but are enacting complex programs of gestures which have their analogs in the modifications of the acoustic output—output frequencies change

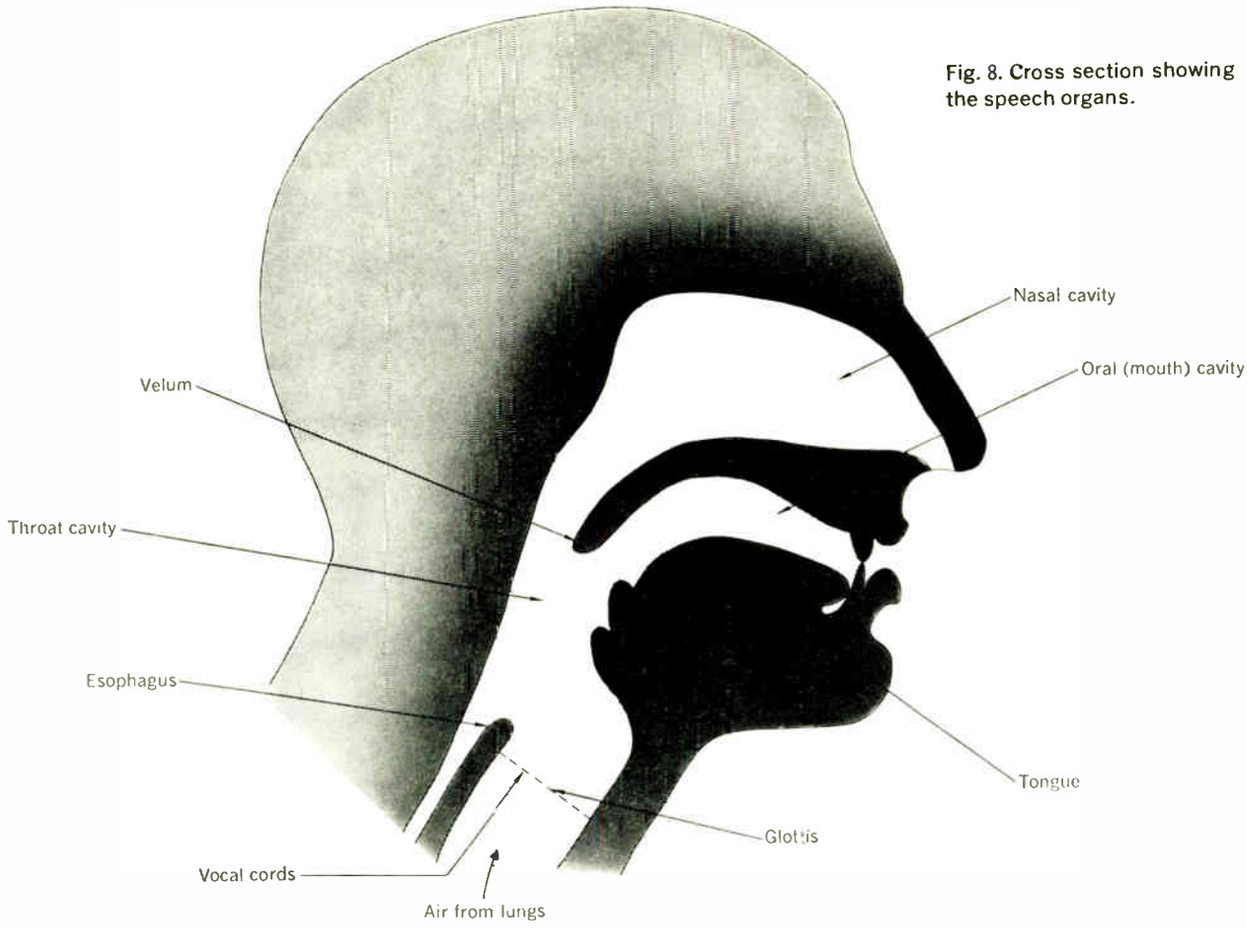


Fig. 8. Cross section showing the speech organs.

(perceived subjectively as changes in pitch), output intensities change (perceived subjectively as changes in loudness), duration of the signals vary (perceived subjectively as length), and so on. The varied coupling of the throat, oral, and nasal cavities produces changing patterns of resonant frequencies. Excitation harmonics in the neighborhood of a cavity resonance are strongly transmitted, forming fairly narrow frequency regions of energy concentration (the formants), the first three of which are the most important for speech. The general range of possible formant frequencies produced by the vocal tract also depends to some extent upon the relative size of the cavities. Thus, men with larger cavities tend to produce a lower range of such frequencies, and women a higher range. In addition, male voices, with their lower fundamental frequencies and closer harmonic spacing, often show more clearly defined formants than those to be found in female voices.

The linguistic outputs possible from this acoustic system are, as we all know, a lexicon of tens of thousands of distinctly different words. (The number of most frequently used words, that is, the normal working vocabulary, is roughly 30 000.) These words, in turn, are composed of syllables, of which there are said to be about 2000 distinct variations (in English). And these syllables, in turn, are built up of the roughly 40 distinct elementary sounds, the phonemes.

One can imagine, then, the potential economy to be achieved if a machine can be devised that will recognize sound patterns on the level of the phoneme.

We must be careful here to draw the distinction between orthographic and phonetic representations. Orthographic representation is the ordinary way of spelling words. A phonetic representation is also a spelling, but on the principle of one-sound, one-letter. For example, *to*, *too*, and *two* sound alike but are spelled differently in our normal orthography. Phonetically, there is just one spelling [tu]. Another example of how orthography has two spellings for one sound is in the words *keep* and *coop*. Phoneticians transcribe the first consonant of these two words with the letter [k]. English obviously has many different spellings for the same sounds, and the same spellings for different sounds, whereas phonetic transcriptions match one letter to one sound, or to one class of very similar sounds.

Sounds that are sufficiently different (in identical contexts) to cause differences in meaning are said to belong to different phoneme classes. Sounds that are more or less similar (whose differences are not sufficient to cause a change in meaning) belong to the same phoneme class. In terms of linguistic notation, when a symbol represents a speech sound in a particular context, it is put in brackets, as above [k]. When a phoneme is referred to, it is put in diagonals /k/.

Phonemes in different positions within words—initial or prevocalic, intervocalic, and postvocalic—often exhibit differing acoustic characteristics. These positional variants of the same phoneme are called allophones.

Each language has a different set of phonemes, which may range in number from a dozen to over five dozen.

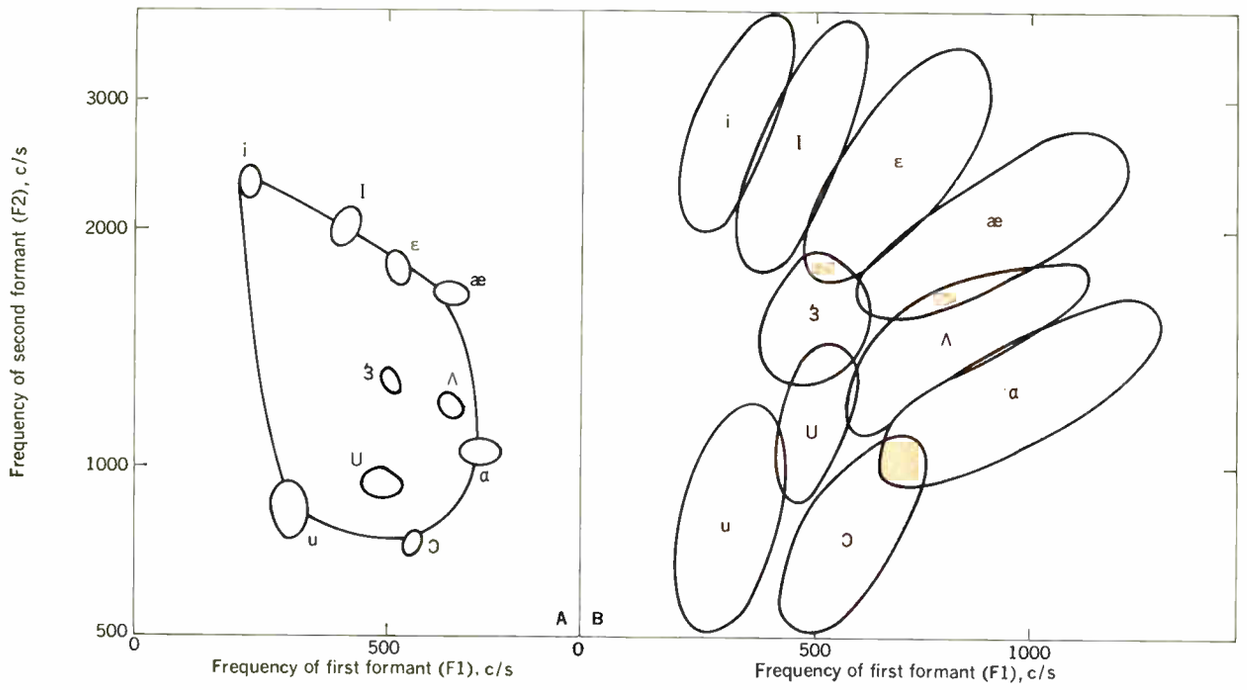


Fig. 9. A—Vowel sounds produced by an individual tend to form a fixed "vowelloop" in formant plots. B—These formant measurements made for a number of men, women, and children show a greater frequency spread.

Linguists generally posit 40 for English, give or take a few (depending upon the linguist).

Before going further, however, into a description of the elementary sounds produced by the vocal tract, we should dwell a little longer, and somewhat subjectively, on the important concept of the phoneme. Each phoneme represents, in effect, a distinct articulatory configuration required to produce it. If one goes down the list of phonemes (Table II), and reproduces aloud each sound, he soon notes that for each sound the different parts of his vocal tract assume a distinct initial posture, the "point" of articulation seems different, and the follow-through of producing the whole sound proceeds in a seemingly programmatic manner. Thus, these settings or configurations of the entire vocal mechanism may be thought of as gestures—verbal gestures—every one distinct from all the others when executed in this pure, isolated form.

The rapid stringing together, or successive performance, of these gestures gives rise to acoustic results, however, that make it difficult, subjectively, to credit this phoneme concept as being valid. As can be imagined, the performing of the gestures can become very sloppy. This, of course, is one of the characteristics of "natural" speech. Seen from the point of view of the articulatory region, the different structures (the lip position, the velum position, the tongue position, the opening or closing of the glottis) and the activity (the breathing, the vibration of the vocal cords, the places of turbulence in the vocal system) are, in natural, continuous speech, always heading toward "target" positions to execute the phonemes, but hardly ever getting to these targets because instructions are pouring in from the central nervous system to get moving on toward the next target sound, that is, to the next phoneme. In many situations, a speaker may fail to pronounce whole sounds, but the human listener understands nonetheless.

The acoustic output, then, of this effort to string phonemes together, to produce syllables and whole words, subjectively sounds quite different from the result of producing separately and carefully the phonemes which compose it. Frequency and power measurements of such whole words and of their single constituent phonemes also confirm such significant differences.

Elementary speech sounds

Traditionally, the acoustic outputs of our articulatory processes are classified into two broad groups of sounds: the vowels and the consonants. These vowels and consonants are dynamically combined in natural speech to form syllables and words. This much we have all been taught.

Further than this, we encounter many more subcategories and descriptive terms, which an engineering education usually does not provide. Speech sounds may be described in terms of articulator movements or position (i.e., how they are produced) or in terms of acoustic outputs. In articulatory terms, vowels may be described as having front-to-back and high-to-low tongue positions, whether nasalized or not, and as voiced or whispered. In psychological terms, vowels are said to have color or timbre. Certain vowel pairs are, of course, known as diphthongs.

Among the subclasses of consonants are the plosives, affricatives, fricatives, nasals, and vowel-like (or "resonants"). Vowel-like sounds may be subdivided into liquids and semivowels. The plosives, or stops, may be voiced or unvoiced, and they are also sometimes called oral stops. The continuant sounds (m, n, ŋ) are usually classed as nasal consonants (see Table II).

Prosodic features of speech are discussed primarily in

III. Range of formant frequencies

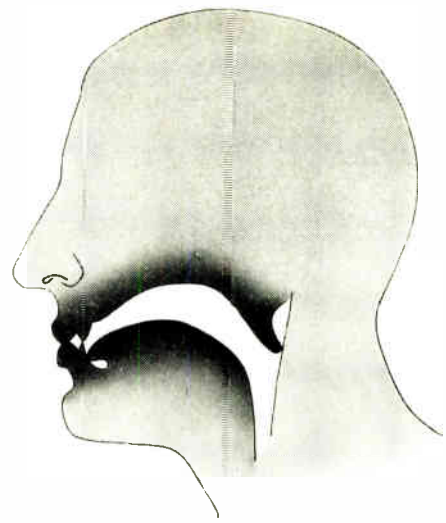
Vowel	Formant 1 Max	Formant 1 Min	Formant 2 Max	Formant 2 Min	Formant 3 Max	Formant 3 Min
u	480	210	1430	570	3300	1850
i	406	190	3100	2000	3900	2600
ɜ	652	360	2120	1130	2480	1400
ɑ	1040	592	1470	820	3180	2020
ɪ	534	206	2700	1710	3400	2340
ɛ	760	370	2570	1650	3300	2200
ʌ	910	550	1688	880	3250	1950

terms of the stress, pitch, and duration of individual speech sounds (or "segments") and combinations of these segments (as in syllables). For this reason, prosodic features are also sometimes called suprasegmental features. There may be several levels of prosodic organization in a language, starting at the syllable or word level, and going up to and beyond the sentence level, where prosody is often called intonation.

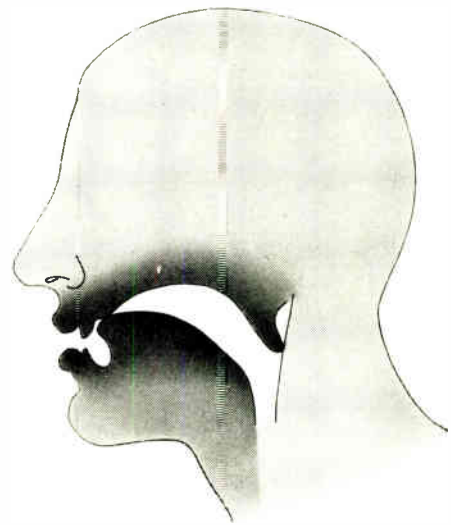
For the vowels, the vocal cords are usually vibrating (i.e., voiced), and the vocal tract is left relatively unimpeded. Different tongue hump positions, and rounding of the lips, produce the different vowels. The vowels usually have higher acoustic power than the consonants. Linguistic identification of vowels does not seem to depend entirely on the absolute frequencies of the formants, but on the frequencies relative to a speaker's total formant structure, which may vary slightly from person to person. For instance, it has been found that the "vowel-loop" [see Fig. 9(A)] for a single speaker tends to remain fixed in shape. Thus, it has been theorized, a listener who "tunes in" on the extremes of a particular speaker's loop frequencies hears the intermediate sounds in relation to this range of tone rather than to a fixed standard. When these formant measurements are made for a number of individuals [see Fig. 9(B)] the vowel regions become more diffuse and overlap; that is, the way one person pronounces /i/ may be similar to the way another person pronounces /I/. This is an example of one factor that a recognition machine must somehow take into account and "normalize." Table III provides another glimpse into how the formant frequencies range between persons (data taken from repetitive speakings of 33 men and 28 women).

Among the consonants, the four "glides" (/w/, /j/, /l/, and /r/) are transitory, being formed by rapid articulatory changes. The nasals (/m/, /n/, /ŋ/), however, can be sustained.

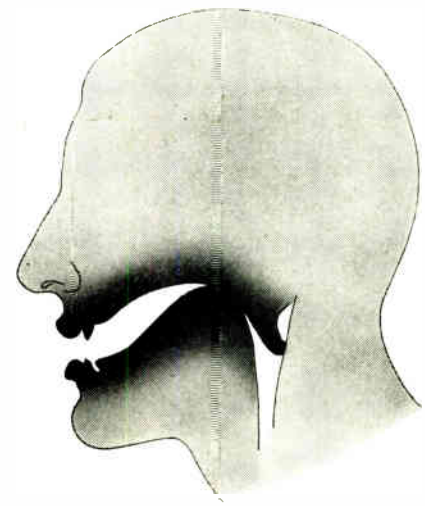
A predominantly turbulent air flow characterizes the fricatives, which can also be sustained. The air passes through a narrow opening at the front of the mouth and over the edge of the teeth. Vocal cords may or may not vibrate. For example, /s/ in *see* is an unvoiced fricative, while /z/ is *zoo* is voiced. Fricatives have low acoustic power. The fricatives are distinguished from affricatives and stops by the duration of the turbulent sound (noise) as well as by the rate at which the initial intensity of the noise rises. Indeed, Prof. Pierre Delattre of the University of California at Santa Barbara has tabulated no less than seven acoustic cues by which fricatives are distinguished



/p/ and /b/



/t/ and /d/



/k/ and /g/

Fig. 10. The three locations of closure for producing stop consonants.

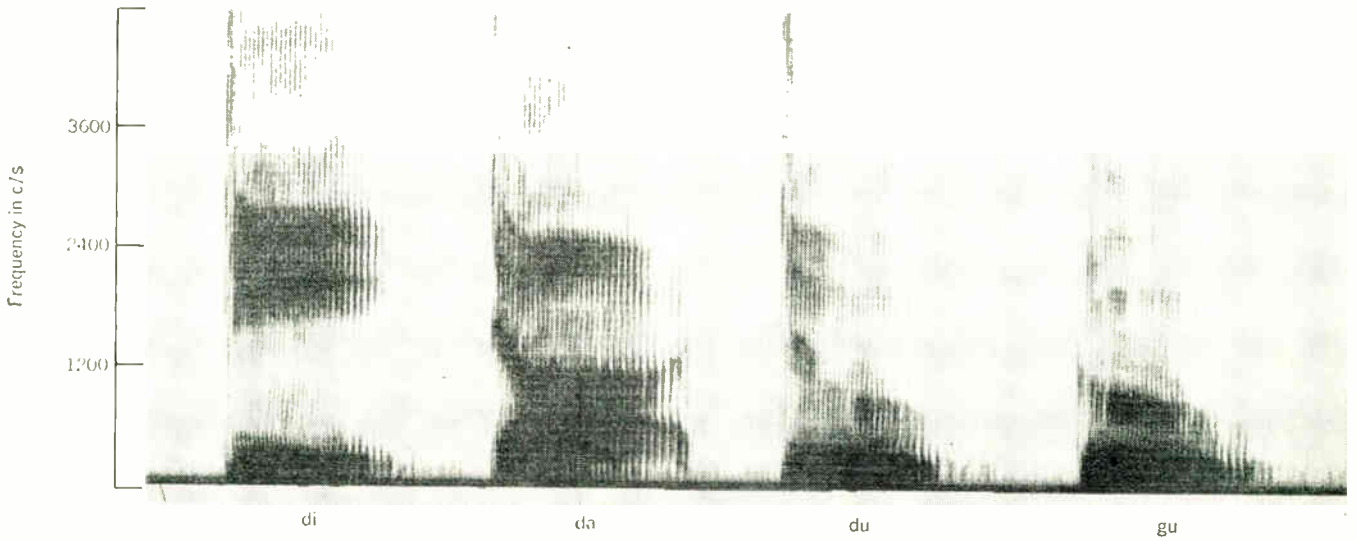


Fig. 11. Spectrograms of the sounds di, da, du, and gu, in which certain of the consonant transitions can be seen. These particular spectrograms, made under less than perfect recording conditions, give some indication of the problems a recognition machine might have in making identifications.

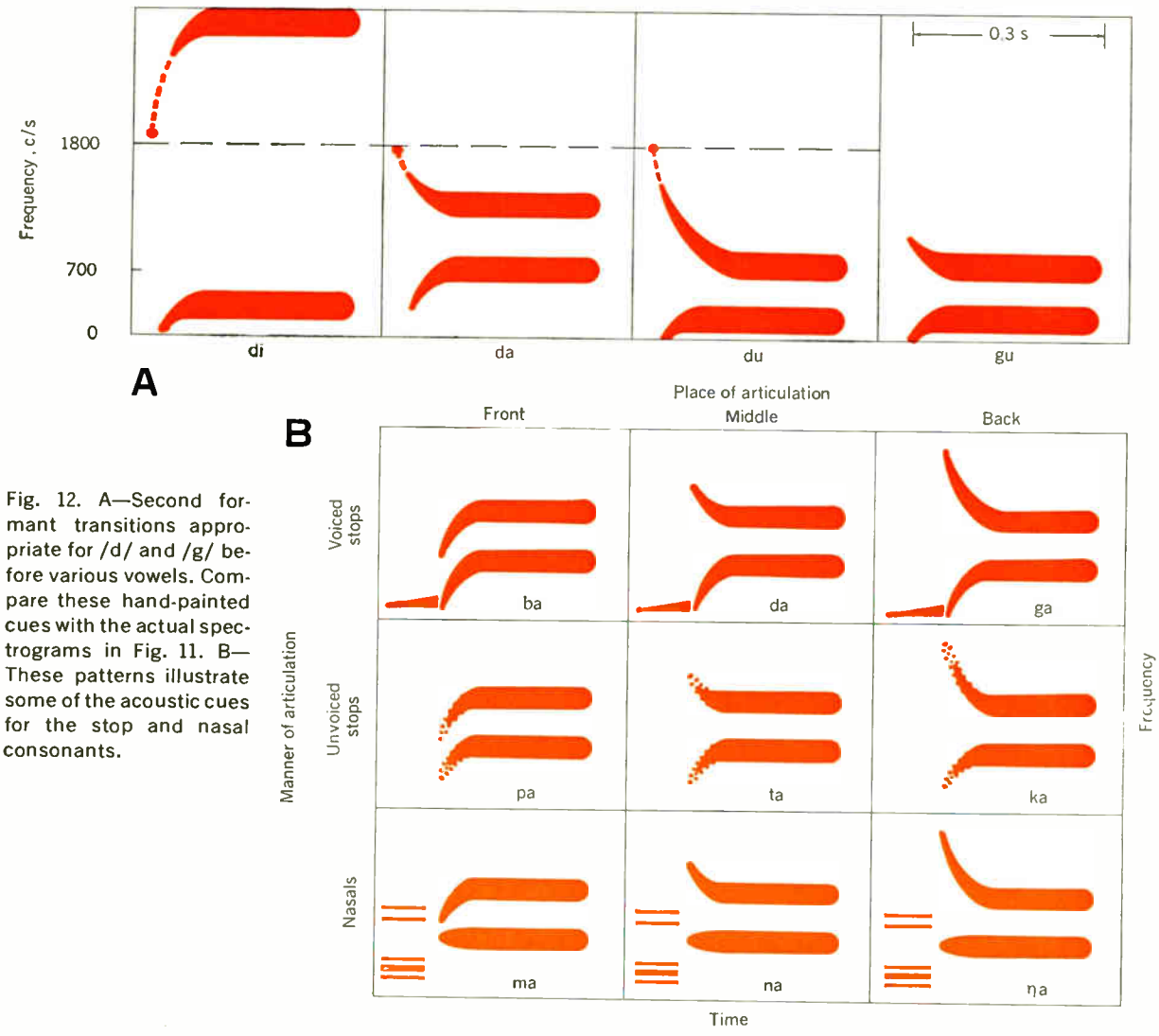


Fig. 12. A—Second formant transitions appropriate for /d/ and /g/ before various vowels. Compare these hand-painted cues with the actual spectrograms in Fig. 11. B—These patterns illustrate some of the acoustic cues for the stop and nasal consonants.

among themselves and from other sounds of the language.

The plosives (explosives) or stops are transient. A silent interval is formed as an air blockage is formed by the lips or tongue, which, when removed, is followed by a very short period of intense turbulence (the burst). A plosive may be voiced or unvoiced, and is of low acoustic power. This class of sounds has probably been the most intensively studied. Figure 10 shows the three locations of closure in the production of English stop consonants.

It should be noted that all of the above, admittedly brief, descriptions will make more sense if the reader pronounces aloud the phonemes in Table II, and correlates his subjective impressions with each description.

A thoroughgoing "engineering-type" description of the generation and characteristics of speech sounds has been prepared by D. B. Fry and Peter Denes.[21] Those who are interested in the development of the engineer's outlook on the speech-making processes should not neglect Homer Dudley's paper, "The Carrier Nature of Speech," published in *The Bell System Technical Journal* in 1940, and now regarded as a classic.

Instrumental methods

Like so many other fields of research, advances in phonetics have depended importantly on the development of new instruments. The analytical methods employed before the War certainly contributed to the store of information on the acoustic cues of speech, but these analyses in many cases led to erroneous conclusions,[22] and they had nothing like the liberating impact on speech research as did the development at the Bell Telephone Laboratories in 1945 of the sound spectrograph,[23] and the subsequent development at several laboratories of speech synthesizers that used various means to transform spectrographic patterns to produce intelligible speech.

The importance of the sound spectrograph lay in the fact that it provided a visual image of the spectra of speech sound. It was in effect the automation of Fourier analysis of speech spectra. It immediately made evident acoustic factors of speech that had not been suspected, and helped to consolidate or eliminate various aspects of the theories that analytical methods had only gradually been yielding.[22] Sound spectrograms (which have become best-known as "visible speech")[23] are composed on a raster of lines, ranging in frequency from bottom to top, in duration from left to right; they appear darker wherever a particular frequency rises in intensity above a certain level. Figure 11 shows spectrograms of the sounds /di/, /da/, /du/, and /gu/. The darker bands, as has been described earlier, are the formants, the lowest being the first formant (F1), the next highest being the second formant (F2), and so on. As can be seen in these sounds, the formants in places change their frequency region quite rapidly. These *formant transitions*, which spectrography made clear, are the acoustic counterparts of articulatory movements, and their elucidation and their role in the perception of consonants is considered to be one of the greatest contributions of phonetics research during the 1950s.[22]

Sound spectrographs have been designed in a number of variants, giving them more flexibility for greater sophistication in experiments.[24] These new instruments do not, however, change the basic method of presentation of spectral information; their chief improvements have been in their mechanical design and circuits. In addition to these, there have been built a number of special-purpose instruments whose main objective is to obtain real-time spectrograms of long samples of speech, reflecting to some extent a shift of emphasis in speech research. The earlier spectrographs presented short samples of speech, most suitable for speech elements on the phonemic and syllable level. To study prosodic features effectively, it is necessary to develop full analyses on the sentence level.

New types of display have also been designed recently. Prestigiacomo at Bell Labs has produced contour spectrograms that show relative intensities that do not show up on the conventional spectrograms. These relative intensity patterns are claimed to be the clue to individual speaker identification.[25] Franklin S. Cooper of the Haskins Laboratories, in an excellent survey,[24] describes some of these new instruments developed at Haskins, at Bell Laboratories, at the Speech Transmission Laboratory of the Royal Institute of Technology in Stockholm, at Columbia University, at the Communication Sciences Laboratory of the University of Michigan, and at the Air Force Cambridge Research Laboratories. (It should be noted, incidentally, that AFCRL's Data Sciences Laboratory, under the direction of Weiant Wathen-Dunn, has sponsored a great share of recent speech research.)

In whatever form, sound spectrographs play a central role in speech research laboratories, and in conjunction with the speech synthesizers that use spectrograms, they have set the major trends of speech research for more than a decade.

However, sound spectrograms also presented their dangers; they presented almost too much information. Provided with over 8000 c/s of acoustic detail, the investigator (as Fant warned, and as Cooper found worthy of quoting)[24] "too easily drowns in a sea of details of unknown significance if he attempts to make use of all observable data."

What the investigators needed, in fact, was some technique that could circumvent two problems inherent in the humanly produced spectrograms: one was the unreliability of the human speaker, that is, his variability in output even when he tried to repeat sounds exactly; the other was an even deeper human constraint—the speaker's inability to change his spectral pattern at will.

The development of the Haskins synthesizer in the early '50s, then, opened the way for a programmatic method of exploration. Elements of synthetic spectrograms were successively suppressed, and the patterns thus amputated were run through the synthesizer. By listening to the result, the experimentalist was able to determine, step by step, which acoustic elements formed the acoustic cues for recognition. This work soon revealed the importance of the first three formants in vowel perception, and the results of this work led the Haskins researchers to make increasingly simplified spectrograms, which still produced intelligible speech.

It should be realized, of course, that the work of reduction has proceeded slowly and methodically, so that it sometimes has taken years of work between the time a single linguistic cue was isolated until it had been definitively analyzed.

As the major acoustic cues for the phonemes were progressively disentangled, the emphasis on this type of research has shifted. Now, there is a need for synthesizers that are closer to normal speech for making studies of

stress and intonation; also, the studies of the relative importance of individual cues for sounds when multiple cues exist demand controlled changes in the total patterns derived from natural speech. For this kind of research, a new synthesizer called a Digital Spectrum Manipulator has been developed at Haskins Laboratories, with which it will be possible to make "microsurgical" modifications to speech spectrograms.[24]

It should perhaps be emphasized at this point that these instruments have been used only on the *acoustic* level of study, and even the most recent refinements of these methods have moved in a well-established direction. It might not be too much of a distortion to say that these studies, aside from their positive values, have provided weighty evidence that it is not feasible to build machines to recognize speech based on the acoustic level alone, and they have shown that new methods, new instruments, new experiments, and new directions would be required if the dream of automation of speech recognition were to come nearer achievement.

More recent studies then, in the past few years, show an unmistakable shift in direction and emphasis. One such study also makes use of a synthesizer, but one of a very different sort; it is the very promising and important development at M.I.T. by K. N. Stevens and his colleagues of an articulatory analog of the human vocal tract.[19] This development, however, brings up other than instrumental issues, and embodies an integral stream or program of research, founded on a rather different philosophical and experimental outlook, thus requiring a separate disquisition. Incorporating as it does the acoustic information we have been discussing, and assembling, as it is, the "generative" features of language, it can be thought of as forming a bridge between the level of acoustic research and the level of linguistic research (and for these reasons, it is discussed in Part II).

However, there is another research tool which promises to open, as it has already in so many other fields, whole new objectives of speech research. That tool, of course, is the computer.

By all accounts, the entrance of computers promises to open many new directions of speech research. Their powers as tools of analysis, synthesis, or simulation, as digesters and sorters of massive quantities of atomic data, are well known. Profound changes in experimental phonetics and in statistical analyses of language are expected.

The computer began to come into use in speech research towards the end of the '50s, as we have already seen, and by now it has become so important a tool that Dr. Stevens of M.I.T. could remark that he thought there already was almost too much emphasis on its use (private communication).

Let us cite just a few of its applications: Caldwell P. Smith at the Air Force Cambridge Research Laboratories has used a digital printout of time- and frequency-quantized spectrograms so as to provide both the pattern and numerical aspects of such spectrograms.[26] Bernard Gold at M.I.T.'s Lincoln Lab has set up a computer program for extracting pitch information from the waveform of voiced sounds.[27] As early as the spring of 1958, James Forgie at the Lincoln Laboratory was devising computer recognition programs. He is at present working on an extensive program for recognizing all the fricatives in various vocalic positions (work is unpublished), and is planning to devise computer programs that will recognize a vocabulary of 1000 words, a program which is intended for use with the Lincoln Sketchpad program. One of his colleagues, Constance McElwain, has set up a program for "degarbling" samples of English text, which had been garbled by a machine reading hand-sent Morse code.[28] She has also worked on the detection of unstressed syllables.

Mathews, Miller, and David,[29] Pinson,[30] Flanagan,[31] Denes,[15] and others, all at Bell Telephone Laboratories, have made extensive analyses using computers. The 1960 work of Denes has already been discussed. In 1963, he reported on a program on the statistics of spoken English,[32] and most recently he has started on a new program of articulatory studies, which will allow the investigator to become a dynamic part of the experiments.[18]

These are just a few examples. Computers have their disadvantages, too—real-time speech production that is generated from stored rules is difficult, and they are expensive. Nevertheless, Franklin S. Cooper of the Haskins Laboratories (from whom many of these observations on instrumentation are derived) states that "an awareness of computer capabilities is becoming a minimal requirement for following research in experimental phonetics."[24]

This year, there will be held the first International Conference on Computational Linguistics, which proposes to include all uses of computers to manipulate natural or artificial languages.

The search for the acoustic cues

The study of the information-bearing elements in speech has progressed steadily, and in a definite direction, although perhaps not entirely systematically over the past decade and more. A sampling of the published papers over this period should give some feeling for the progress.

The methods of these studies differed. For instance, Gordon E. Peterson, originally at Bell Telephone Laboratories, conducted analytical studies of vowels. In 1953, he presented data on two front vowels spoken by different types of speakers, and gave evidence that a listener identifies vowels by frequency positions of the first and second formants.[33] He used similar analytical methods and instrumentation in his later work, reported on in 1961, which summed up much of his earlier work at Bell.[34] In this work, also on vowels, he suggests that studies of humanly produced vowels are handicapped, and are more satisfactorily carried forward through speech synthesis methods. More recent work done under his direction at the Communications Sciences Laboratory at the University of Michigan includes a massive study of the allophones (variants of phonemes) of the phonemes /r l w y h/. Four positional variants of these sounds were included in the study.[35] (An interesting automatic speech recognition program has also come most recently from Peterson's laboratory. It is discussed later in this article.)

Another type of study, also reported in 1953, was by C. M. Harris, who made rearrangement experiments with sounds and showed that the interaction between contiguous speech sounds was perceptually significant.[36]

About that same time, researchers at the Haskins Laboratories in New York were well embarked on their extensive program of investigation of the acoustic cues

Fig. 13. Hand-painted spectrographic pattern for the word "typical."

of speech through their synthetic speech methods. Their earlier work, appearing in 1952, began with the study of various acoustic cues in isolation;[37] later in the decade, they went on to study combinations of cues provided simultaneously. Certain of their experiments, reported in 1955, showed that both the second and third formant transitions play a role in the perception of the voiced stops /b/, /d/, and /g/.[38] A follow-up of this work, in 1958, was carried out by H. S. Hoffman, who tested listeners with synthetic speech containing all possible combinations of single, double, and triple simultaneous cues. He showed that burst frequency was also a cue in the perception of voiced stops.[39] Still other experiments, in 1956, showed that the tempo of the transitions was sufficient to distinguish members of a class of voiced stop consonants from corresponding members of the class semivowels and vowels of changing color.[40] A fine interpretive article (appearing in 1957) on the Haskins work done during this period is that of Alvin Liberman.[41]

In 1957, the Haskins researchers specified the major acoustic differences between the set of consonants /w r l y/ in the intervocalic position;[42] in the same year, they studied how listeners lumped acoustically varied sounds into phoneme categories;[43] in 1958, they described the cues for unvoiced fricatives and their voiced counterparts;[44] also in 1958, they described the effects of third-formant transitions;[45] they also studied the distinctions between voiced and voiceless stops in initial position;[46] and so on.

There are two superb summations of the work of this acoustic research. One, by A. M. Liberman and his colleagues, catalogues "rules" for the acoustic cues required to synthesize speech.[47] In this paper, there are summarized the results of ten years of intensive investigation into the respective roles played by acoustic and articulatory phenomena in speech perception. With the rules devised in the Haskins work, it is possible to hand-paint the proper elements to create understandable speech through the use of a special Pattern Playback (or its vocoded twin, "Voback") machine. It adds much to one's understanding of the relation of linguistic-to-acoustic elements to see what these hand-painted cues look like. For instance, Fig. 12(A) shows some of the second-formant transitions appropriate for recognizing /d/ and /g/ before various vowels. Figure 12(B) shows patterns of some of the acoustic cues for the stop and nasal consonants. Figure 13 shows the cues for the word *typical*.

Figure 14(A) shows how the various categories of rules are combined to specify a word pattern, in this case, for the word *labs*. Compare this artificial pattern with actual spectrograms in Fig. 14(B) of two different persons saying the same word. These two figures indicate qualitatively how much redundancy (linguistically speaking) and possibly noise exists in the human acoustic output.

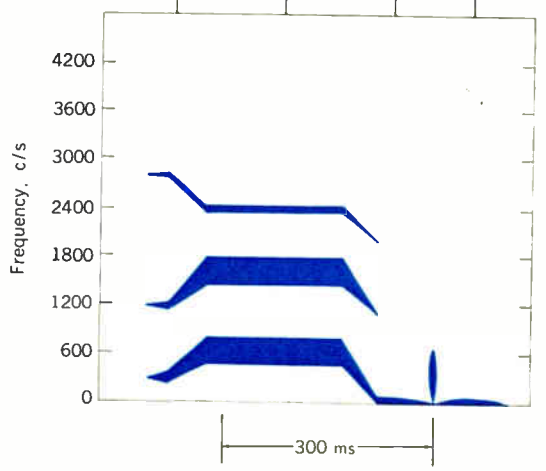
The other summation of acoustic research, for the ten years until 1957, is that of Pierre Delattre.[22] His paper provides an excellent view of the historical development of the work that led to the isolation of the many acoustic cues, which he breaks down for all the classes of sounds (fricatives, nasal stops, oral vowels, etc.) and his summation also dates the beginnings of research on the prosodic elements of speech (stress, rhythm, intonation). In addition, he supplies a bibliography of the major papers of that era of speech research, consisting of more than 50 references, the significance of which he marks in the appropriate places.

More recent papers in the important Haskins "opus" include a study of the effect of learning on speech perception (in 1961), which showed that there is an increased discrimination across phoneme boundaries,[48] an elaboration of their method of speech synthesis by rules[49] (in 1962), and a description of their provocative and much-

Fig. 14. A—The categories of rules devised at Haskins Laboratories combined to specify a word pattern. The word synthesized is "labs." B—Actual spectrograms of the words "labs" uttered by two different speakers.

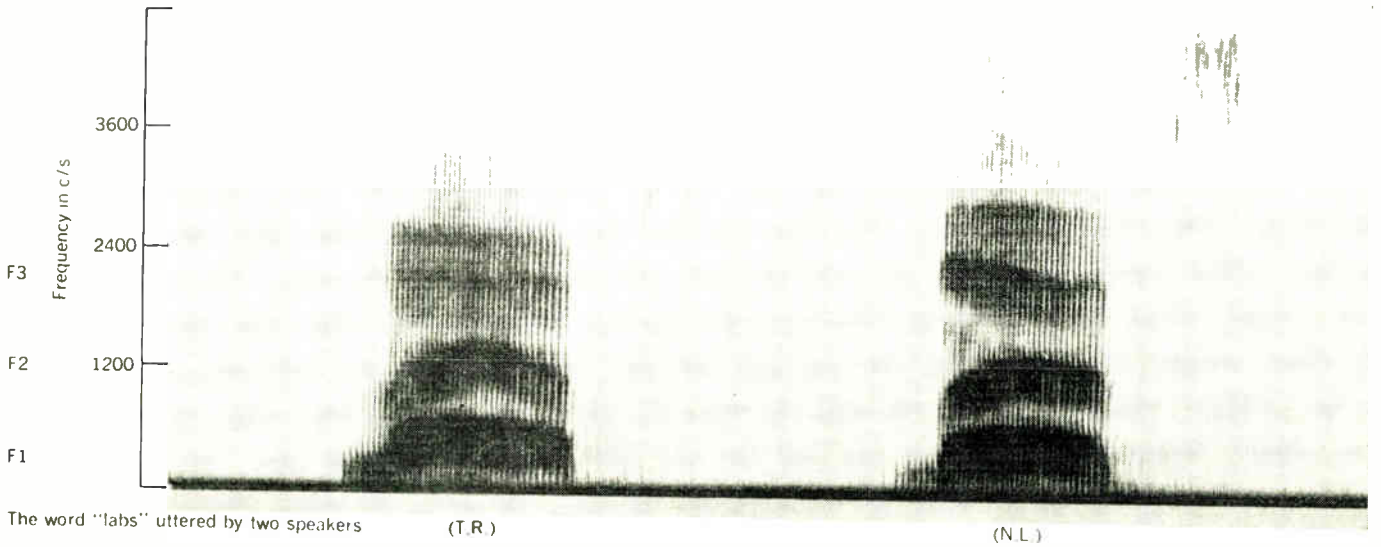
SYNTHESIS BY RULES: /læbz/

Manner	*Resonants* /wrly/: Periodic sound (buzz); formant intensities and durations are specified. F1 locus is high. Formants have explicit loci.	*Long vowels* /ieɛæaɔo/: Periodic sound (buzz); formant intensities and durations are specified.	*Stops* /pbtdkg/: No sound at formant frequencies; i.e., "silence." Burst of specified frequency and band width follows "silence." F1 locus is low. F2 and F3 have virtual loci.	*Fricatives* /fvθðszʃʒ/: Aperiodic sound (hiss); intensity and band width are specified. F1 locus is intermediate. F2 and F3 have virtual loci.
Place	/l/: F2 and F3 loci are specified.	/æ/: Formants frequencies specified.	*Labials* /pbfvm/: F2 and F3 loci are specified. Frequencies of buzz and hiss are specified.	*Alveolars* /tdsz/: F2 and F3 loci are specified. Frequencies of buzz and hiss are specified.
Voicing	(The voicing rules are only applied to those phonemes for which the condition of voicing has differential value. For the resonants and vowels, which are invariably voiced, the acoustic features correlated with voicing are specified under Manner.)		*Voiced* /bdg/: Voice bar. Duration of "silence" is specified. F1 onset is not delayed.	*Voiced* /vðzʒ/: Voice bar. Duration of "silence" is specified. F1 onset is not delayed.
Position		Vowels in final syllable: Duration is double that specified under Manner.		



A

B



The word "labs" uttered by two speakers

IV. Acoustical parameters of speech

F1 —frequency of vowel or consonant first formant
F2 —frequency of vowel or consonant second formant
F3 —frequency of vowel or consonant third formant
F_{Z1}—frequency of consonant first antiresonance
F_{Z2}—frequency of consonant second antiresonance
F_0 —fundamental voice frequency
d —duration of successive vowels and consonants
α —instantaneous speech power
$\bar{\alpha}$ —average speech power

debated motor theory of speech perception, put forward in September 1962 at the Speech Communication Seminar at the Royal Institute of Technology in Stockholm.[50] And most recently, they have reported on their electromyographic studies of the tongue during speech production.[51] But these later papers reveal a new direction of research, beyond the acoustic level, and so are better treated in Part II.

All in all, the Haskins opus, starting in 1950 and continuing until the present time, provides us with a trunk line into the heart of the acoustic research of this period. Even though Haskins has never yet attempted to design speech recognition machines, their research, as perhaps even this superficial account may convey, forms an important component of the work towards the objective of developing nontrivial machines.

A summary of speech parameters

Conceptually, there are many ways that the acoustic variables or acoustic features of speech could be specified and quantified. That is, there are many sets of relevant pattern features that might be used in an automatic speech recognition system, but thus far authors have not specified or selected the most important or informative features.[52] This failure may be due in part to the fact that not all the most relevant features, and their interrelationships, have been made clear in acoustic studies.

However, in lieu of this complete and final picture of the most relevant features or patterns in speech, let us look at some of the lists of information-bearing acoustical parameters that have been used in the limited recognition machines, and that have been proposed as possible candidates for machines of the future.

Gordon E. Peterson, Director of the Communications Sciences Laboratory at the University of Michigan, in a general and philosophical discussion of procedures for automatic speech recognition, assembled a set of information-bearing acoustical speech parameters.[53] These measurable parameters are given in Table IV.

An "acoustical speech parameter" is defined by Peterson as a unidimensional time function derivable from a physical analysis of an acoustical speech sound class. Speech waves may be characterized by four such classes of sounds.

1. Quasi-periodic sounds: These involve recurrent excitation by one or more vibrating mechanisms (vocal cords, velum, tongue tip, lips) plus resonance (and sometimes antiresonance) due to the source and transfer functions of the vocal cavities. Spectrum and overall amplitude may vary with time. Parameters: fundamental frequency and resonance characteristics (amplitudes, bandwidths, and frequencies of resonances and antiresonances).

2. Quasi-random sounds: Essentially continuous spectrum (frictionally produced); both spectrum and overall amplitude may vary with time. Parameters are the resonance characteristics.

3. Gaps: Periods of silence in speech. Parameter is overall instantaneous speech power.

4. Impulses: These explosive or implosive sounds follow gaps. Parameter: (impulsive rise time and peak level) overall instantaneous speech power.

Various combinations of these basic sound classes may occur.

For linguistic (phonemic) interpretations: The vowels and continuant consonants are identified primarily by the character of the resonances. Fundamental voice frequency may also be important for identifying vowels. Gaps and impulses are important for identifying plosives.

The three essential prosodic parameters of speech are defined by Peterson as vowel and consonant duration, fundamental laryngeal frequency, and speech production power. All these parameters, said Peterson, merit much further research.

Philip Lieberman of the Air Force Cambridge Research Laboratories has recently reported on studies involving these last two factors, studies that have led him to postulate a perceptual model in which "intonation" is given a central role in providing acoustic cues that allow a listener to segment speech into blocks or chunks for syntactic analysis.[54] This interesting research, however, lifts our viewpoint from the level of the phoneme, and its acoustic correlates, to the level of syntax; so a discussion of Lieberman's work is postponed till Part II.

New automatic recognition techniques

In an earlier section of this survey, we considered some of the first attempts at building automatic speech recognition machines. To conclude, let us now look at some of the most recent attempts. There are, in fact, two systems that are worth taking in conjunction, and which may help to set the final lines on the perspective we have been attempting to draw.

Both systems are new, both have been able to rely on the strength and the discoveries of the acoustical research of the past decade, both are sophisticated in their approaches, and both are treating recognition almost solely on the acoustical level with the full understanding that this is the primary or lowest level for what must eventually be a multilevel hierarchy of processes. Thus, these two approaches are a logical outcome of the present spirit of speech research, and are representative of the present state of the art.

Both systems have been consciously limited to what is possible. Neither has attempted connected speech. They differ in that one approach is based on the use of a computer for tracking the distinctive features of vowels; the other approach uses neural logic, and recognizes the more difficult consonant sounds by the frequency–energy relationships that vary with time. Both approaches look promising.

Recognizing distinctive features by computer. The distinctive-feature description of speech of Jakobson, Fant, and Halle has been mentioned earlier. Although this scheme holds a strong position in the thinking of speech researchers, its possible value as applied to automatic speech recognition has only partially been explored. Its earliest implementation was in the electronic

successive-binary-selection system of Wiren and Stubbs, discussed earlier.[11]

Now, J. F. Hemdal of the University of Michigan and G. W. Hughes of Purdue University have devised a computer recognition program to extract the physical correlates of the distinctive features.[55] Their program is designed to recognize ten cardinal vowels, nine diphthongs, and takes into account the effects of the consonant environments in which these vowels and diphthongs occur.

In the implementation of this program, 227 CVC (consonant-vowel-consonant) nonsense syllables, plus 50 short monosyllabic common words and samples of continuous speech, were recorded on magnetic tape under normal conversational conditions. These nonsense syllables and words were constructed in such a way that all CV and VC combinations would occur. These speech data were put into an IBM 7090 computer in spectral form, obtained by sampling the rectified and smoothed outputs of 35 bandpass filters. Each sample from each filter was quantized into one of 1024 possible levels by an analog-to-digital converter and punched on data-processing cards. This information formed the basis for the recognition program.

The following four distinctive feature pairs were sufficient to provide vowel recognition: (1) acute/grave, (2) compact/diffuse, (3) flat/plain, and (4) tense/lax. The physical (acoustical) correlates of these feature pairs, which were tracked in the program, were determined somewhat as follows:

1. Acute/grave (High second formant/low second formant)
2. Compact/diffuse (High first formant/low first formant)
3. Flat/plain (F1 + F2 threshold/F1 + F2 threshold)
4. Tense/lax (Longer duration and greater departure from a neutral position/shorter duration and less departure from a neutral position)

A slight amplification of these terms is undoubtedly in order. (It will help to look at Fig. 15, which is an idealized F1–F2 plane with vowel regions shown with the first three feature boundaries.) For (1): grave phonemes show more intensity in the lower portion of the frequency spectrum as opposed to acute phonemes. For vowel phonemes: when the second formant is closer in frequency to the third formant than to the first formant, the vowel is probably acute. For (2): the first formant frequency is

Fig. 15. Idealized F1–F2 plane with vowel regions marked off by the first three distinctive-feature boundaries employed in the Hemdal–Hughes recognition program.

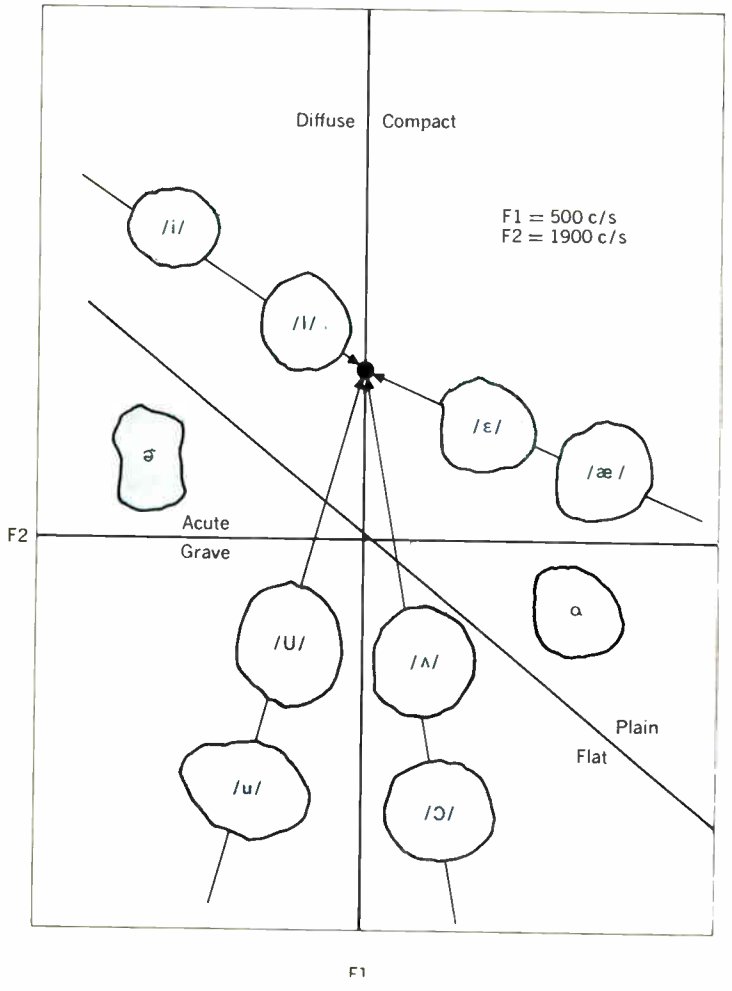
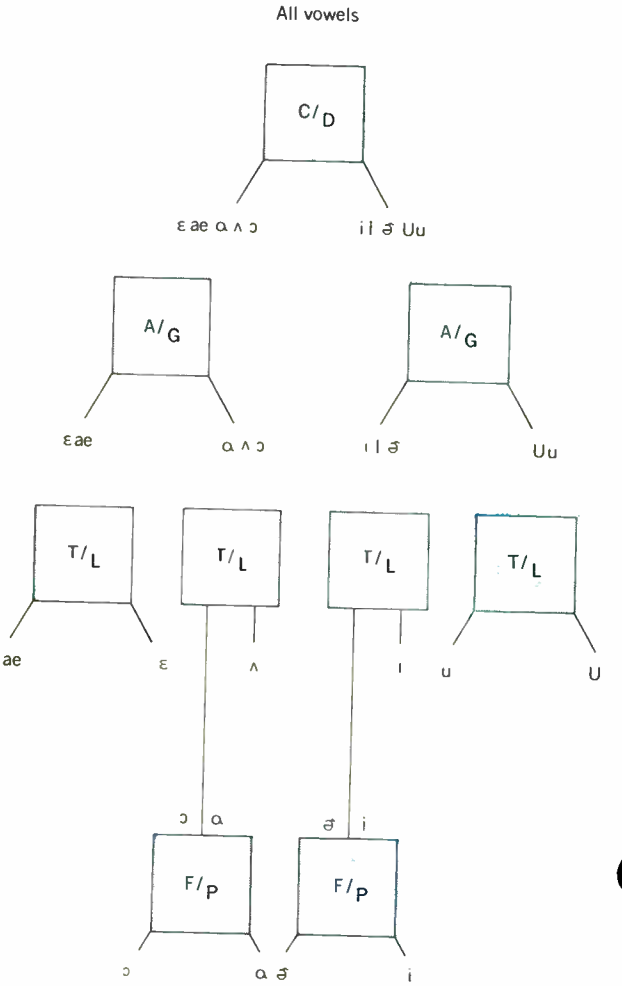


Fig. 16. Decision tree for the ten vowels of the Hemdal–Hughes program. Four binary feature-pairs were found sufficient to distinguish all these vowel sounds.



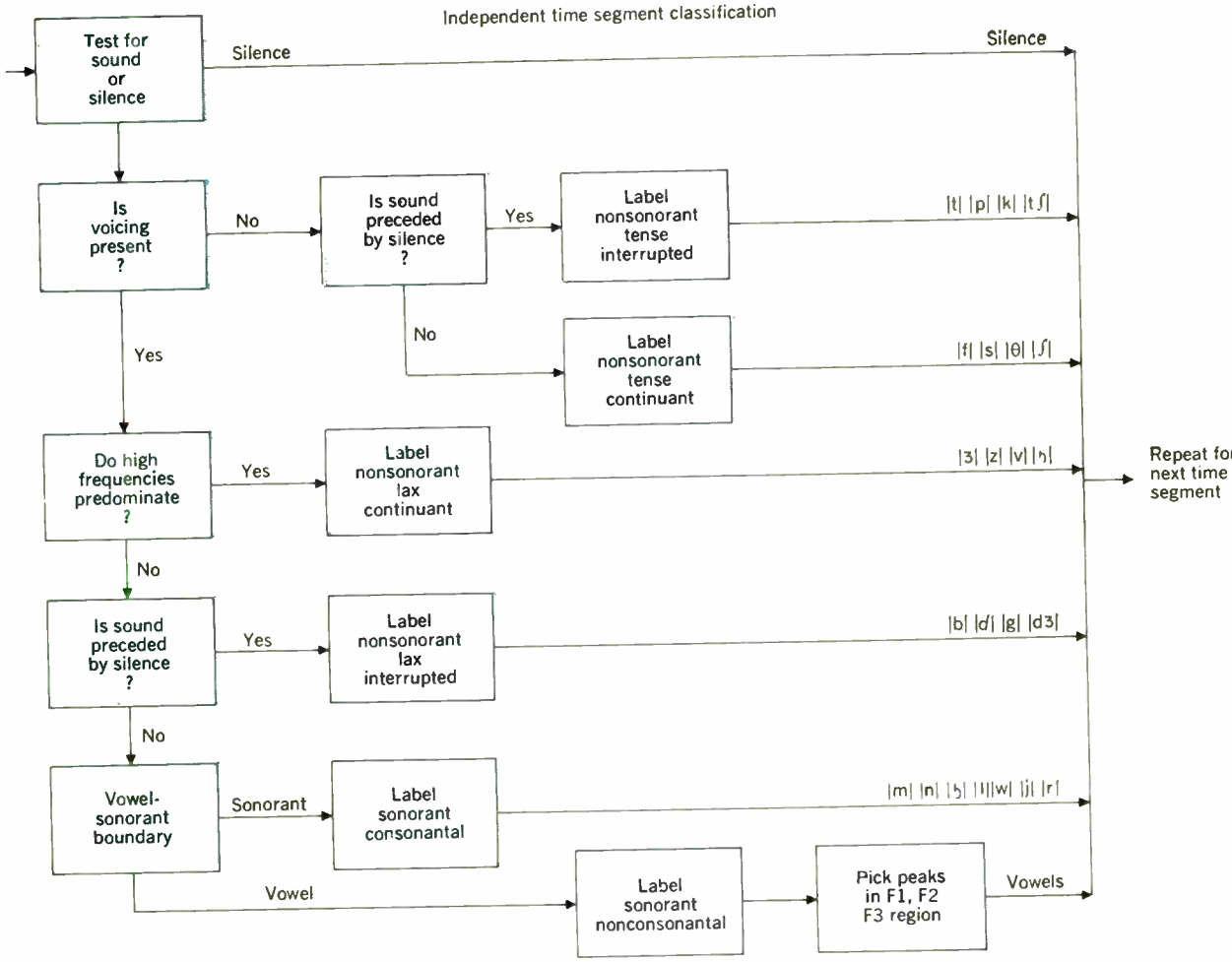
sufficient for identifying the compact/diffuse feature—F1 threshold was set at 500 c/s. For (3): a downward/-upward shift of a set of formants or all formants in the spectrum characterizes the flat/plain feature; thus, the physical correlate was determined by the sum of F1 and F2. For (4), the tense/lax feature pair ("perhaps the least well known of the vowel features"), there is a lengthening of a tense vowel and a shift of the formant frequencies away from a neutral position. Thresholds varied for each speaker (requiring normalization), but the form of making the decision was maintained. Hemdal and Hughes say if a recognition scheme such as this based on distinctive features were completely implemented, some kind of device would be needed to normalize the signal input of each speaker before the formants could be tracked and a decision made.

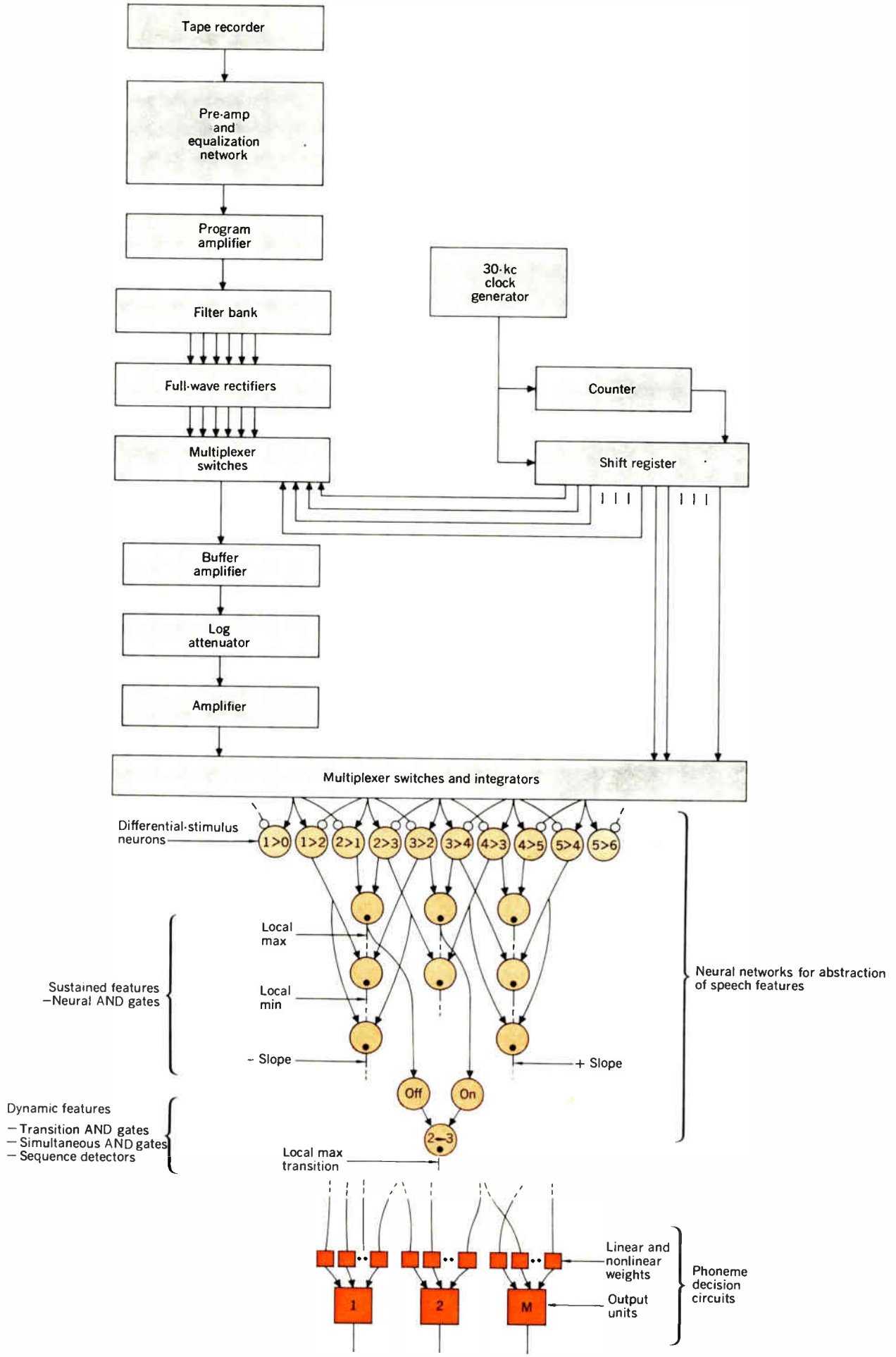
The Hemdal–Hughes decision tree for the ten vowels, in Fig. 16, shows how the four pairs sort out the vowels. An indication of how the computer program was set up appears in Fig. 17. Acoustic data are examined in either single time segments or in combinations of segments until the various consonants, liquids, etc., are sorted out as shown. Once it is known that a particular segment is a vowel, the computer determines approximate formant frequencies, and each vowel time segment is classified in accordance with the distinctive features tracked (as in Fig. 17).

The computer recognition results were evaluated by comparing them with the responses of 25 listeners who heard the same speech sounds. There was a close correlation of the computer and human responses—the computer accuracy was 92 per cent and the human accuracy was between 96 and 88 per cent—for words spoken in isolation. For connected speech, accuracy was poor. The positive results of this research, thus far, seem to strengthen the view that the Jakobson, Fant, and Halle distinctive feature approach is useful for automatic speech recognition.

Speech recognition using neural-like logic. Probably one of the most interesting physical implementations of phoneme recognition systems is that using neural-like logic elements.[3,56] Biological neurons, which are regarded as the basic information-processing elements of the animal nervous system, have been under intensive investigation for the last few years, and various electronic models of neurons have been designed which are more or less faithful representatives of the biological originals.

Fig. 17. In the Hemdal–Hughes computer program, time segments are first classified by an energy measurement as being either speech or silence. Then the speech segments are sorted into nonsonorant or sonorant phonemes through other property measurements—absence of energy below 350 c/s shows lack of voicing; frequency components above 4000 c/s indicate turbulence, etc. Vowels are separated (from nonvowel sonorants) on the basis that they are stronger sounds. Finally, the vowel formants are located by a spectral peak-picking routine.





Such "artificial" neurons appear to have some ideal characteristics for both aural and visual pattern recognition tasks—they lend themselves to parallel information processing, and can maintain a quantitative measure of probability throughout all logical operations, thus providing an assurance level that a particular pattern feature is or is not present. Not only can they be used to indicate the presence or absence of a feature, but they can also measure the amount by which a feature is present. In the RCA phoneme-recognition system, this capability of making analog measures of quantity has been found to be essential for separating nearly identical phonemes with overlapping characteristics.[3]

Although the ultimate objective of this program is to develop a speech recognition system that will recognize continuous speech, most work to date has been directed toward developing the logical networks required for recognizing the more difficult consonant sounds (plosives, fricatives, and vowel-like). The recognition equipment built thus far[3] uses 500 neural-like elements (called analog threshold logic or ATL). A block diagram of the system appears in Fig. 18. The system can abstract both relatively sustained and complex dynamic spectral variations (rapid speech transients) over a 60-dB dynamic range, and it operates in real time.

The system takes into account the fact that the features of each consonant phoneme are modified by the features of the phoneme preceding and following it (i.e., its "local" context). The speech samples consisted of isolated CVC sounds uttered by six different speakers (the consonant to be recognized in the initial position, in combination with ten different following vowels, and the same final consonant /d/ for all sounds)—e.g., *cud, could, dead, sad, ved, heard, yawd, lewd, wooed,* etc. Despite the fact that the consonants exhibited considerable overlapping in their features, recognition scores were quite high.

Future efforts in this program will aim at including vowel recognition, and studies will be made of variations in phoneme features for intervocalic and final positions. However, the researchers state that all of the principles utilized to construct recognition networks for isolated sounds are directly extendable to continuous speech.

The most important accomplishments of this program are best put in their authors' words:

"A significant deviation . . . between the present work and past investigations has been the type of features utilized for the recognition of the individual phonemes. In past investigations, the location in the spectrum of the formants and their movements with time have been considered to be the significant features of speech. The results of the present study, however, indicate that for machine recognition of speech, the features that are more invariant and more easily abstracted by machine are the spectral regions of increasing and decreasing energy (positive and negative slopes). This is not to say that either formant or pole-zero analysis of speech is not significant from the standpoint of human recognition or speech synthesis; rather, it is a statement to the effect that for machine recognition of speech it is far easier to abstract the regions of increasing and decreasing spectral energy. A striking example of the invariance of the slope features is the fact that a single onset transition of slope features was sufficient for the recognition of a semivowel in combination with ten following vowels for all six male speakers used in the investigation. The formants, on the other hand, undergo wide ranges of movement within the spectrum for the ten following vowels. The invariance of slope features and the ease with which they could be implemented for machine recognition are two of the most significant findings of the present study. It should be mentioned that the spectral locations of the formants and antiformants are available in the present equipment and were compared directly with the slope features for all of the phonemes investigated. However, the actual recognition networks . . . *do not utilize a single formant or antiformant.*"[3]

Fig. 18. This block diagram of a neural-type speech processor gives just a slight indication of the complexity involved. Basically, speech spectra are divided into 19 segments by an overlapping bank of bandpass filters whose outputs are operated on in various ways to produce a degree of amplitude-independent feature abstraction. Envelope shape of the spectrum and its time variations are obtained from 36 difference-taking circuits. Detailed descriptions of the neural network operations have appeared in many reports.

It was in response to this approach that Dr. C. Gunnar M. Fant of Sweden most recently remarked that "There has been an overemphasis in tracking formants," and he expressed an interest in and sympathy for approaches that did not rely on formant tracking. He went on to relate an anecdote about one of his recent visits to a speech symposium in Moscow. While he was there, a Russian colleague had queried him: "Oh, are you still tracking formants? That is old-fashioned. We don't do that anymore."

Conclusions

These two systems, then, bring us up to the present time. In a sense, they mark the extent of one aspect of the automatic speech-recognition art, and they raise provocative questions. Although they both consciously work primarily on acoustic recognition, and they both stress that linguistic information will be required in an ultimate machine, their immediate strategies (apart from the physical implementation) appear to be rather different. For instance, John Hemdal of the University of Michigan, in response to the question of how his system would "tune in" on different speakers, says: "We expect the ultimate recognition machine to be adaptive in some sense—that is, adapting to new speakers." Whereas T. B. Martin of RCA, in response to a similar question, says: "Rather than monitoring speakers, we wish to get the real invariants (of the speech sounds)."

More specific questions were directed at the designers of the neural logic system by Professor A. S. House (of the University of Purdue), a propounder of good questions:

"What kind of difficulties do you foresee when you add final consonants, when you add more speakers, when you add noise? How will your system compare with the many other types of systems, both simple and complex, that do these types of recognition? What justifies your greater complexity?"

Of both systems, he asks the question: "What happens when the system is extended logically to include the whole inventory of speech sounds, that is, of natural speech?"

At this point in time, such questions remain unanswered, and it is at this point that the surveyist must necessarily leave off.

The author is indebted to many persons who kindly gave assistance and guidance. He especially thanks: Dr. K. N. Stevens and Prof. Morris Halle, of M.I.T.; Dr. Peter Denes, Leon Harmon, Dr. J. Flanagan, and Dr. E. E. David, all of Bell Telephone Laboratories; Weiant Wathen-Dunn, of the Air Force Cambridge Research Laboratories; James Forgie, of the Lincoln Laboratory; Dr. H. Rubenstein, of the Harvard Center for Cognitive Studies; and Thomas P. Rootes, Jr., of Haskins Laboratories, who generously read the manuscript and offered many suggestions.

REFERENCES

1. Edwards, P. G., and Clapper, Jr., J., "Better Vocoders Are Coming," *IEEE Spectrum*, vol. 1, no. 9, Sept. 1964, pp. 119–129.

2. Olson, H. F., "Speech Processing Systems," *Ibid.*, no. 2, Feb. 1964, pp. 90–102.

3. Martin, T. B., *et al.*, "Speech Recognition by Feature-Abstraction Techniques," Tech. Doc. Report No. AL TDR 64-176. AF Avionics Lab., Wright-Patterson AF Base, Ohio, Aug. 1964.

4. Flanagan, J., Private communication, Bell Telephone Laboratories, Murray Hill, N.J.

5. Lawrence. W., "Role of Synthetic Speech in Speech Research," *J. Acoust. Soc. Am.*, vol. 36, no. 5, May 1964, p. 1022.

6. Miller, G. A., "The Psycholinguists, On the New Scientists of Language," *Encounter*, vol. 23, no. 1, 1964.

7. Fatehchand, R., "Machine Recognition of Spoken Words," in *Advances in Computers*, vol. 1, F. L. Alt, ed. New York: Academic Press, Inc., 1960, pp. 193–229.

8. Dreyfus-Graf, J., "Sonograph and Sound Mechanics," *J. Acoust. Soc. Am.*, vol. 22, Nov. 1950, pp. 731–739.

9. Davis, K. H., *et al.*, "Automatic Recognition of Spoken Digits," *Ibid.*, vol. 24, no. 6, Nov. 1952, p. 637.

10. Dudley, H., and Balashek, S., "Automatic Recognition of Phonetic Patterns in Speech," *Ibid.*, vol. 30, 1958, pp. 721–732.

11. Wiren, J., and Stubbs, H. L., "Electronic Binary Selection System for Phoneme Classification," *Ibid.*, vol. 28, 1956, pp. 1082–1091.

12. Jakobson, R., Fant, C. G. M., and Halle, M., "Preliminaries to Speech Analysis," Tech. Report No. 13, Acoust. Lab., M.I.T., Cambridge, Mass., 1952.

13. Jakobson, R., and Halle, M., *Fundamentals of Language*. 's Gravenhage, Netherlands: Mouton & Co., 1956.

14. Denes, P., "The Design and Operation of the Mechanical Speech Recognizer at University College, London," *J. Brit. Inst. Radio Engrs.*, vol. 19, 1959, pp. 219–229.

15. Denes, P., and Mathews, M. V., "Spoken Digit Recognition Using Time-Frequency Pattern Matching," *J. Acoust. Soc. Am.*, vol. 32, Nov. 1960, pp. 1450–1455.

16. Forgie, J. W., and Forgie, C. D., "Results Obtained from a Vowel Recognition Computer Program," *Ibid.*, vol. 31, Nov. 1959, pp. 1480–1489.

17. Forgie, J. W., and Forgie, C. D., "A Computer Program for Recognizing the English Fricative Consonants /f/ and /θ/," presented at Fourth International Congress on Acoustics, Aug. 1962.

18. Denes, Peter, Private communication, Bell Telephone Laboratories, Murray Hill, N.J.

19. Halle, M., and Stevens, K., "Speech Recognition: A Model and a Program for Research," *IRE Trans. on Information Theory*, vol. IT-8, no. 2, Feb. 1962, pp. 155–159.

20. Sakai, T., and Doshita, S., "The Phonetic Typewriter," (Kyoto Univ., Japan), in *Information Processing 1962*, Proc. of IFIP Congress 62, C. M. Popplewell, ed. Amsterdam: North-Holland Publishing Co., 1963, pp. 445–449.

21. Fry, D. B., and Denes, P., "The Role of Acoustics in Phonetic Studies," in *Technical Aspects of Sound*, vol. 3, E. G. Richardson and E. Meyer, eds. Amsterdam: Elsevier Publishing Co., 1962, pp. 1–69.

22. Delattre, P., "Acoustic Cues in Speech: First Report," available from Haskins Laboratories, N.Y.; first appeared in French in *Phonetica*, vol. 2, 1958.

23. Potter, R. K., Kopp, G. A., and Green, H. C., *Visible Speech*. Princeton, N. J.: D. Van Nostrand Co. Inc., 1947.

24. Cooper, F. S., "Instrumental Methods for Research in Phonetics," *Proc. of Fifth International Congress of Phonetic Sciences*, Münster, Germany, Aug. 1964.

25. Kersta, L. G., "Voiceprint Identification," *J. Acoust. Soc. Am.*, vol. 34, 1962, p. 725.

26. Smith, C. P., "Voice-Communication Method, Using Pattern Matching for Data Compression," *Ibid.*, vol. 35, 1963, p. 805.

27. Gold, B., "Computer Program for Pitch Extraction," *Ibid.*, vol. 34, 1962, pp. 916–921.

28. McElwain, C. K., and Evens, M. B., "The Degarbler—A Program for Correcting Machine-Read Morse Code," *Information and Control*, vol. 5, no. 4, Dec. 1962, pp. 368–384.

29. Mathews, M. V., *et al.*, "Pitch-Synchronous Analysis of Voiced Sounds," *J. Acoust. Soc. Am.*, vol. 33, 1961, pp. 179–186.

30. Pinson, E. N., "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths," *Ibid.*, vol. 35, 1963, pp. 1264–1273.

31. Flanagan, J. L., "Computer Simulation of Basilar Membrane Displacement," *Proc. IVth Int'l. Cong. Acoustics*, Copenhagen, Denmark, Aug. 1962.

32. Denes, P. B., "On the Statistics of Spoken English," *J. Acoust. Soc. Am.*, June 1963, p. 892.

33. Peterson, G. E., "The Information-Bearing Elements of Speech," in *Communication Theory*, W. Jackson, ed. New York: Academic Press, Inc., 1953.

34. Peterson, G. E., "Parameters of Vowel Quality," *J. Speech and Hearing Res.*, vol. 4, no. 1, March 1961.

35. Lehiste, I., "Acoustical Characteristics of Selected English Consonants," Report no. 9, Communications Sciences Lab., U. of Mich., Ann. Arbor, Mich., July 1962.

36. Harris, C. M., "A Study of the Building Blocks in Speech," *J. Acoust. Soc. Am.*, vol. 25, 1953, p. 962.

37. Cooper, F. S., *et al.*, "Some Experiments on the Perception of Synthetic Speech Sounds," *Ibid.*, vol. 24, Nov. 1952, p. 597.

38. Delatre, P. C., *et al.*, "Acoustic Loci and Transitional Cues for Consonants," *Ibid.*, vol. 27, July 1955, p. 769.

39. Hoffman, H. S., "Study of Some Cues in the Perception of the Voiced Stop Consonants," *Ibid.*, vol. 30, Nov. 1958, p. 1035.

40. Liberman, A. M., *et al.*, "Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds," *J. Exp. Psy.*, vol. 52, no. 2, Aug. 1956, p. 127.

41. Liberman, A. M., "Some Results of Research on Speech Perception," *J. Acoust. Soc. Am.*, vol. 29, Jan. 1957, p. 117.

42. Lisker, L., "Minimal Cues for Separating/w, r, l, y/ in Intervocalic Position," *WORD*, vol. 13, no. 2, Aug. 1957.

43. Liberman, A. M., "The Discrimination of Speech Sounds Within and Across Phoneme Boundaries," *J. Exp. Psy.*, vol. 54, no. 5, Nov. 1957, p. 358.

44. Harris, K. S., "Cues for the Discrimination of American English Fricatives in Spoken Syllables," *Lang. and Speech*, vol. 1, pt. 1, Jan.–Mar. 1958, p. 1.

45. Harris, K. S., *et al.*, "Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants," *J. Acoust. Soc. Am.*, vol. 30, no. 2, Feb. 1958, p. 122.

46. Liberman, A. M., "Some Cues for the Distinction Between Voiced and Voiceless Stops in Initial Position," *Lang. and Speech*, vol. 1, pt. 3, July–Sept. 1958, p. 153.

47. Liberman, A. M., *et al.*, "Minimal Rules for Synthesizing Speech," *J. Acoust. Soc. Am.*, vol. 31, no. 11, Nov. 1959, p. 1490.

48. Liberman, A. M., "An Effect of Learning on Speech Perception: The Discrimination of Durations of Silence With and Without Phonemic Significance," *Lang. and Speech*, vol. 4, pt. 4, Oct.–Dec. 1961, p. 175.

49. Cooper, F. S., "Speech Synthesis by Rules," *Proc. of Speech Communication Seminar*, Stockholm, Sweden, 1962.

50. Liberman, A. M., "A Motor Theory of Speech Perception," *Ibid.*

51. MacNeilage, P. F., and Sholes, G. N., "An Electromyographic Study of the Tongue During Vowel Production," *J. Speech & Hearing Res.*, vol. 7, no. 3, Sept. 1964.

52. Lai, D. C., "A Criterion for the Selection of Speech Features in Speech Recognition Based on Comparison of Experiments," *Proc. of the Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Mass., Nov. 1964, to be published.

53. Peterson, G. E., "Automatic Speech Recognition Procedures," *Lang. and Speech*, vol. 4, pt. 4, Oct.–Dec. 1961, pp. 200–219.

54. Lieberman, P., "Intonation and the Syntactic Processing of Speech," *Proc. of the Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Nov. 1964, to be published.

55. Heindal, J. F., and Hughes, G. W., "A Feature Based Computer Recognition Program for the Modeling of Vowel Perception," *Ibid.*

56. Zadell, H. J., *et al.*, "Acoustic Recognition by Analog Feature-Abstraction Techniques, *Ibid.*

Liquid-metal magnetohydrodynamics

The merits of various MHD cycles are investigated to establish the potential of the liquid-metal concept for commercial power systems. A comparison is made of the plasma and liquid-metal MHD cycles

Michael Petrick *Argonne National Laboratory*

For all practical purposes, effective further development of the steam cycle is rapidly nearing an end. The relatively small gains in efficiency that have been achieved by increasing the pressure and temperature to very high levels in the steam cycle have not been reported to be overly successful from an economic viewpoint. As a result, the search is broadening for new or unconventional power systems that can meet the ever-increasing demand for additional power more efficiently and economically.

From a thermodynamic viewpoint, steam is a poor working fluid because it absorbs too small a fraction of the heat input at the maximum cycle temperature. However, when all pertinent factors are considered, no other single fluid has been found to be superior to steam. A logical extension of the technology, therefore, is the combination of several working fluids into a binary cycle. The advantages of such a cycle have long been recognized and a substantial effort has already been made in this direction with the development of the mercury–steam turboelectric cycle. The acceptance of this concept has been delayed because of the serious problems encountered during its introduction. Many of these problems have been resolved but the maximum temperature of the system has been kept below 1000°F.

Based upon developments in reactor and materials technology and in magnetohydrodynamics (MHD), a power system is evolving which appears to have a strong potential for commercial development: coupling a liquid-metal topping cycle with a conventional steam-bottoming cycle. The topping cycle would consist of a liquid-metal generator tied directly to a liquid-metal-cooled reactor and operating in the temperature range of 1000°F–1600°F, and possibly reaching temperatures as high as 2000°F.

Recent advances in the field of magnetohydrodynamics and, in particular, liquid-metal magnetohydrodynamics indicate that the conversion of heat into electricity through an MHD device appears quite feasible. A number of cycles have been proposed for a liquid-metal MHD power system. The cycles are similar in that they are based on the conversion of thermal energy into kinetic energy or stagnation head, which is then converted into electric energy by an MHD generator. The cycles differ primarily in the manner in which the conversion of the thermal energy is achieved.

The introduction of a topping cycle above 1000°F will create additional problems that will have to be resolved—special design problems arising from the ultra-high temperatures and material limitations as they affect component fabrication, cost, reliability, and, most important, longevity. The lifetime of power equipment vs. the maximum operating temperature is shown in Fig. 1. The illustration was taken from a recent evaluation of the SNAP program[1] and depicts typical experience gained from commercial rotating equipment, ducted

gases, and reactors. Also shown in Fig. 1 are the goals of the more ambitious AEC programs. It appears that the turboelectric topping cycle may very well be limited to a temperature range of 900–1400°F due to the very serious problems that are encountered beyond this range in turbine design and in other system components such as valving, pumps, etc. The liquid-metal MHD topping cycle, on the other hand, shows great promise for being developed as an efficient power system, capable of operating at temperatures up to 2000°F. Unlike the conventional turboelectric generator, the MHD generator contains no moving solid parts that are subject to extreme temperature and dynamic stress or require close machine tolerances. As a result, the MHD generator can operate under conditions of high temperatures, and in highly corrosive and erosive atmospheres and temperatures that conventional energy-conversion devices could not tolerate for prolonged periods of time.

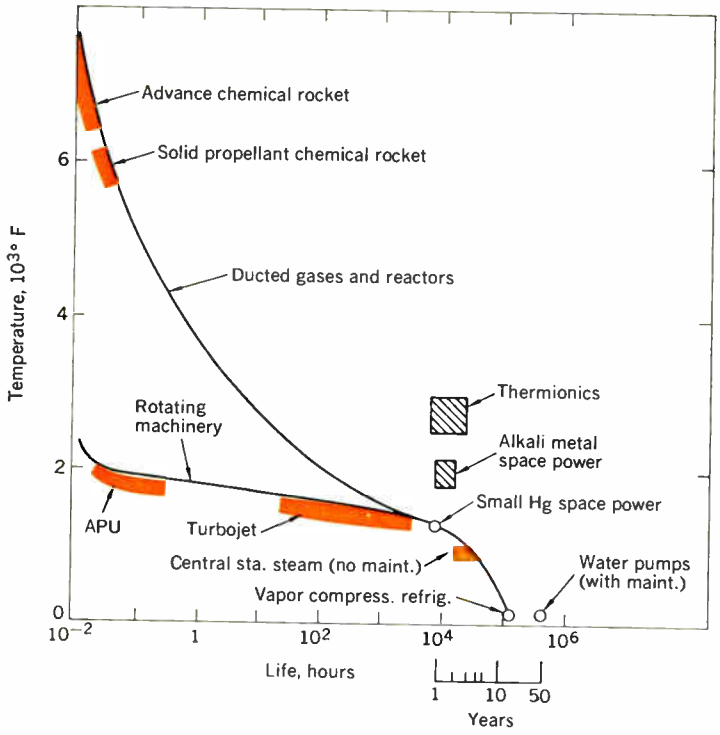
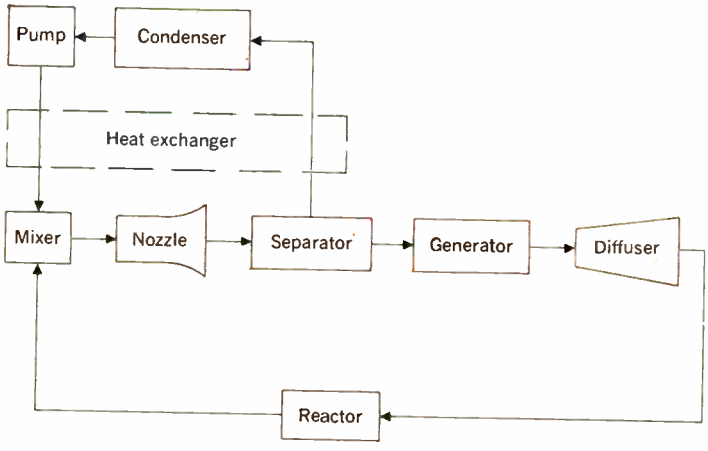


Fig. 1. Curve illustrating maximum operating temperature vs. life for power equipment.

Fig. 2. Schematic block diagram showing two-phase two-component cycle proposed by Elliott.



Cycle studies

Liquid-metal MHD cycles. The basic energy-conversion steps in the liquid-metal MHD cycles that have been proposed are: (1) the transfer of heat from the heat source to the liquid; (2) conversion of part of this heat to vapor enthalpy (subsequently rejected by the condenser); (3) conversion of the remaining heat to kinetic energy of the liquid in the nozzle; and (4) conversion of most of this kinetic energy to electric power in the generator. These steps are essentially similar in all the proposed cycles as is evident from the following descriptions.

The two-component two-phase MHD cycle initially proposed and analyzed by Elliott[2] as a space power cycle is schematically illustrated in Fig. 2. One fluid circulates in the vapor loop and the other in the liquid loop. The fluid circulating in the vapor loop leaves the condenser as condensate and is pumped (by an electromagnetic pump) to the mixer, where it vaporizes on contact with the liquid. The vapor expands with the liquid through a two-phase nozzle, separates from the liquid in the separator, and recondenses in the condenser. To raise the cycle efficiency, a heat exchanger cools the vapor while preheating the condensate. In the liquid loop, the liquid is heated in the reactor and cooled as it vaporizes the condensate in the mixer. The liquid is then accelerated by the vapor in the nozzle, separated from the vapor in the separator, decelerated by the production of electric power in the MHD generator, and returned through the diffuser to the reactor.

The cycle proposed by Prem[3] is also bicyclic. It can be operated with either one or two components. The single-component version of the cycle, which is especially suited for large power stations, is illustrated in Fig. 3. Liquid metal in the first loop is partially vaporized by the heat source. The resulting two-phase fluid expands upon passing through a supersonic nozzle and transforms its thermal energy to kinetic energy. Downstream of the nozzle and at considerably lower temperatures, atomized liquid from the second loop is injected into the two-phase high-velocity stream. Due to momentum exchange, the injected stream is accelerated; at the same time, due to mass and heat transfer between the two streams, the vapor component of the two-phase fluid is condensed. The resulting fluid stream enters the generator predominantly as a liquid phase having a high velocity. A major fraction of the fluid kinetic energy is then transformed into electric energy within the generator. The second loop carries a portion of the metal through a heat exchanger where the waste heat of the liquid-metal cycle is rejected.

The two-phase one-component cycle proposed by Petrick and Lee[4] is schematically illustrated in Fig. 4. The cycle is a simplified version of the two-component two-phase cycle. It consists of five basic components: the two-phase nozzle, MHD generator, condenser, diffuser, and reactor. A two-phase mixture or a saturated vapor is removed from the reactor and is passed through the nozzle where its kinetic energy is increased. From the nozzle, the two-phase fluid passes directly through

the MHD generator, where the electric energy is extracted, before passing to the condenser. The fluid is returned from the condenser via a diffuser to the reactor. Upon comparing this cycle with the two-component two-phase cycle of Fig. 2, it is evident that the separator and the vapor loop are eliminated, thus simplifying the cycle and increasing its efficiency. In addition, solubility and entrainment problems are also eliminated.

The condensing injector liquid-metal MHD power cycle proposed originally by Jackson and Brown[5] is shown schematically in Fig. 5. The cycle consists of a vapor loop (reactor loop) and a liquid loop (heat rejection loop). The vapor is generated in the reactor heat source and passes into the condensing injector where it is mixed with the liquid stream emerging from the waste heat exchanger. In the condensing injector, the vapor is condensed and a high stagnation head liquid is generated. The fluid passes through the MHD generator, where electric energy is extracted at the expense of the stagnation pressure head, and then is separated into two streams; one passes into the vapor loop and hence to the reactor where it is vaporized, and the other is sent to the liquid loop where heat is rejected in a heat exchanger.

Results of cycle analyses. Extensive studies were made of the two-component two-phase cycle, the condensing injector cycle, and the one-component two-phase cycle. Pertinent typical results from these studies follow. In addition, recent data on the potential of the Prem cycle[6] are also presented. The data refer to a liquid-metal MHD system operating between two temperature limits. Although the results of these studies would be applicable to a variety of power systems that could be used in space, under the ocean, in limited access terrestial locations, etc., the primary emphasis in this article will be on the potential application of the MHD cycle as a topping unit on a central station power system. The working fluids studied were mercury, mercury potassium alloy (40 mol %K), potassium, cesium, and sodium. The detailed results of these extensive cycle studies will be available in a forthcoming report.[7]

The two-phase two-component cycle. The variation of the cycle efficiency as a function of the liquid and gas flow rate, exit nozzle pressure, and generator efficiency is shown in Fig. 6. The temperature before the nozzle T_1 was set at 2470°R (2010°F) and the pressure P_1 at 150 psia; saturation pressure of potassium vapor at T_1 is 150 psia. The lithium vapor pressure at this condition is small (2 psia) but must be considered in the cycle analysis. If $P_1 > 150$ psia, the potassium would not vaporize prior to entering the nozzle.

It is apparent that the overall efficiencies for this cycle are very low; the theoretical maximum efficiency is < 16 per cent. The latter condition is based on the assumptions that $\epsilon_g = 100$ per cent and $P_2 = 0.25$ psia. These parameters are actually limits and do not represent realistic attainable conditions. It is obvious, however, that to attain the maximum cycle efficiency, P_2, the nozzle exit pressure, should be kept as low as possible and ϵ_g as high as possible. The lower the value of the nozzle exit pressure the greater will be the conversion of thermal to kinetic energy. As the efficiency of the generator ϵ_g is increased, greater amounts of power can be extracted from the cycle.

It should be noted that the parameter ranges shown in Fig. 6 are applicable to a space power system. These

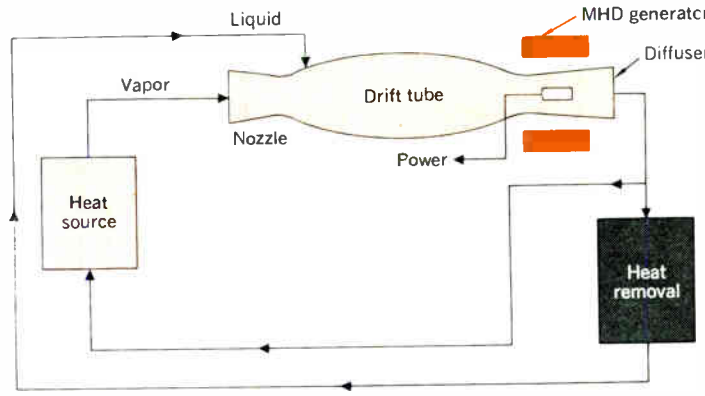
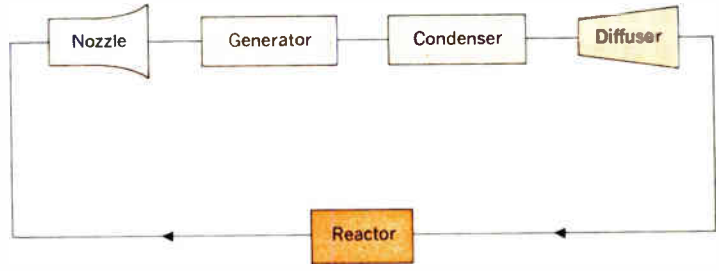


Fig. 3. Schematic diagram of two-phase cycle proposed by Prem.

Fig. 4. Schematic diagram of simplified version of two-phase two-component cycle proposed by Petrick and Lee.



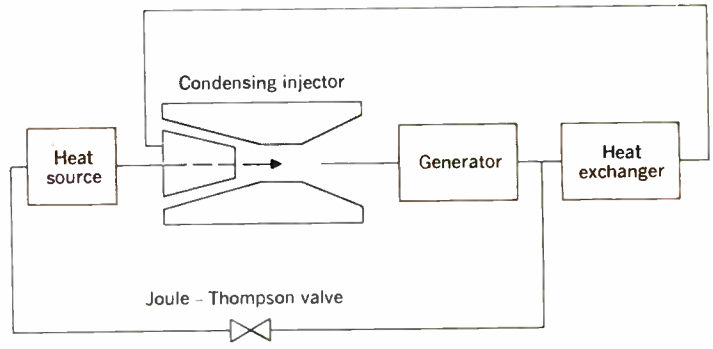
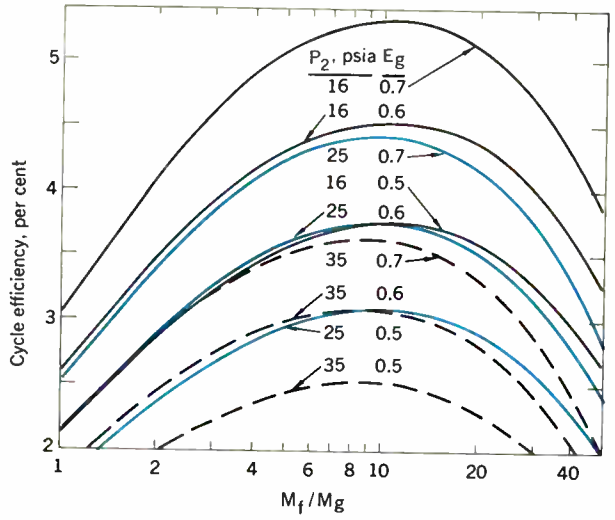
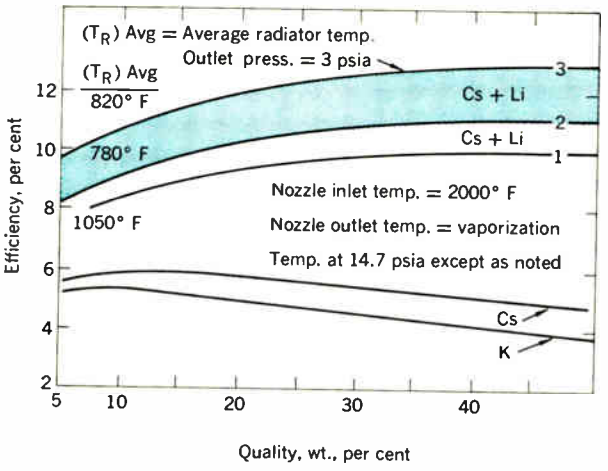


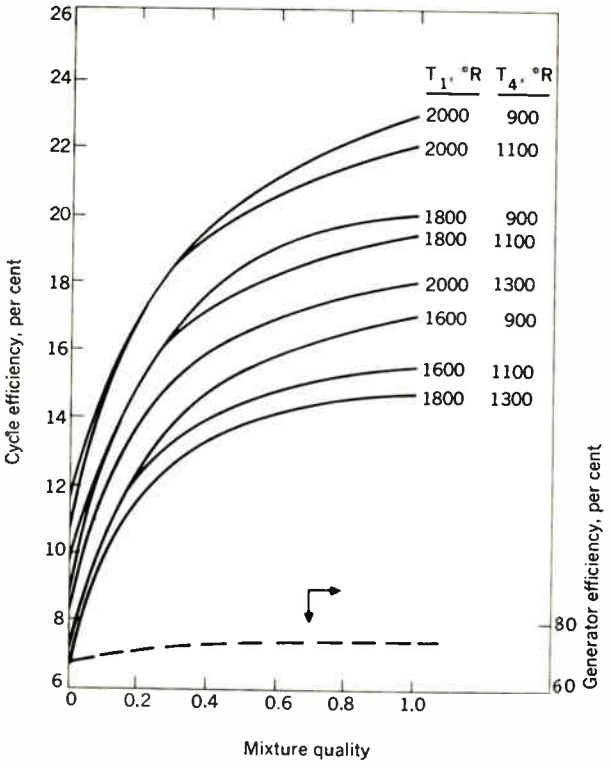
Fig. 5. Schematic diagram of condensing injector liquid-metal MHD power cycle proposed by Jackson and Brown.

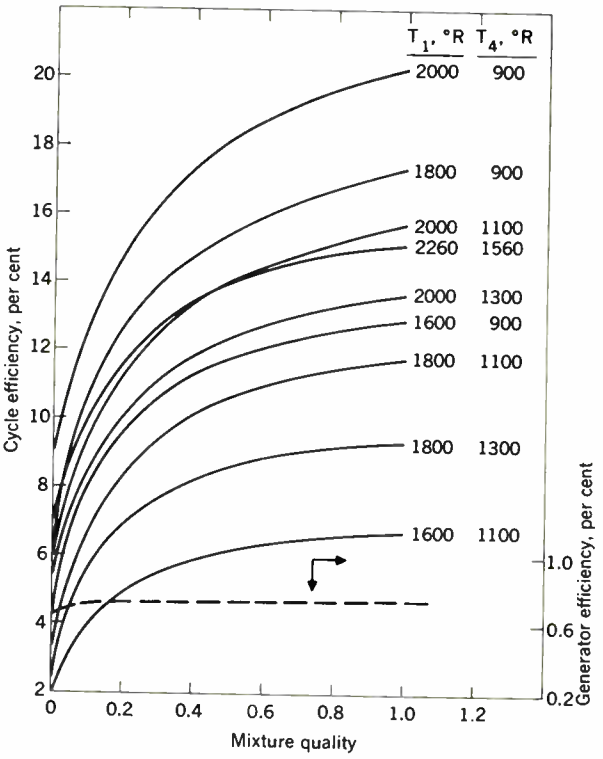
results, however, serve to demonstrate a serious limitation of the cycle—the excessive kinetic energy loss that occurs in the separator. The total kinetic energy loss in the separator was calculated to be ~46 per cent; this corresponds to a velocity loss between nozzle exit and MHD generator entrance of ~26 per cent. According to Elliott's recent separator experiments,[8] the velocity losses in a conical separator were found to range between 17 and 30 per cent. The calculated values therefore appear to be quite accurate. The kinetic energy loss represents a direct power loss since the power extracted from the MHD generator is proportional to the entering kinetic energy of the working fluid.

Two-phase one- or two-component cycle. Theoretical efficiency analyses were obtained from Prem[6] for both the one- and two-component versions of this power









conversion cycle. Figure 7 shows typical predicted efficiencies. No attempt was made to optimize the cycle in these analyses. For the single-component cycle, results are presented for potassium and cesium as working fluids. As can be seen an optimum cycle efficiency occurs for an inlet quality entering the supersonic nozzle of about 10 per cent. The operating conditions for these results were an inlet temperature of 2000°F and an outlet temperature set equal to the vaporization temperature of the fluid at 14.7 psi. Curves 1, 2, and 3 in Fig. 7 represent predicted efficiencies for the two-component cycle operating on an immiscible cesium–lithium mixture. Curve 1 shows the efficiency at an average radiator temperature of 1050°F; curve 2 indicates the improvement in the cycle efficiency if the sink temperature is decreased to 780°F., and curve 3 shows the improvement in the efficiency that can be obtained if the working fluid is expanded to 3 instead of 14.7 psi.

The predicted conversion efficiencies shown were based on the assumption of 0.70 for both the mechanical and electrical conversion efficiencies. The mechanical efficiency allows for losses in the nozzle and frictional

Fig. 6. (upper left) Cycle efficiency as a function of system parameters for the Elliott power cycle.

Fig. 7 (left). Cycle efficiency as a function of system parameters for the Prem power cycle.

Fig. 8. (lower left). Efficiency of the two-phase one-component cycle for mercury as a working fluid.

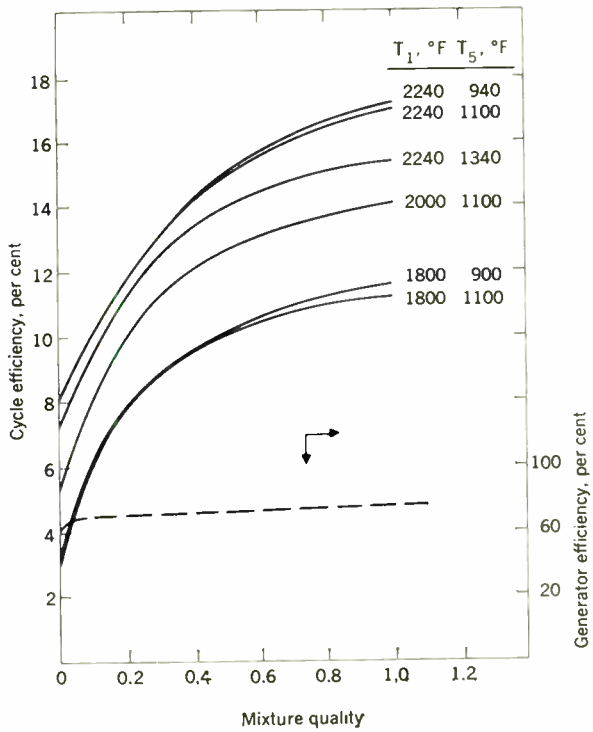
Fig. 9 (below). Efficiency of the two-phase one-component cycle for the potassium–mercury alloy as a working fluid.

losses throughout the system, and the electrical efficiency includes end losses in the generator and the magnetization current for producing the magnetic field. To make the results of Prem's analysis comparable to the three cycles studied by the author, the efficiency data shown in Fig. 7 should be multiplied by a factor of 1.14 to account for the lower generator efficiency used. Additional improvements in the cycle can be obtained by using a regenerative nozzle and providing multiple injection for the subcooled liquid. These improvements could raise the overall cycle efficiencies by a factor of 1.21.

Two-phase one-component cycle. Typical cycle efficiencies for K, K–Hg, and Hg are shown in Figs. 8 through 10. The overall cycle efficiency is plotted vs. the mixture quality at the inlet to the nozzle. The parameters on the curves are the reactor temperature and the sink temperature; the generator efficiency is also shown. The maximum efficiency in the cycle is achieved by use of the pure vapor ($x = 1.0$) at the nozzle inlet. The maximization at $x = 1.0$ is due to the fact that the outlet velocity of the two-phase mixture emerging from the nozzle and entering the generator increases with increasing quality.

The highest efficiencies that can be realized for the one-component cycle within the boundary conditions specified is obtained with mercury as the working fluid. An overall cycle efficiency of 22.5 per cent is obtained for a source temperature of $T = 1540°F$ and a sink temperature $T = 540°F$. The efficiency of the MHD generator under these conditions is $\epsilon_g = 0.75$ and the nozzle efficiency is 80 per cent. The results for the potassium–mercury mixture are similar to those with pure mercury. The principal apparent difference is that the cycle efficiencies are ~10 to 25 per cent lower and are much more sensitive to the source and sink temperatures.

Fig. 10. Efficiency of the two-phase one-component cycle for potassium as a working fluid.



Mercury and mercury–potassium eutectic have the distinct advantage of having lower boiling points and hence operating temperature ranges that are nearer present-day technological limits. This advantage rapidly diminishes for the pure mercury system since its vapor pressure increases rapidly with temperature: at $T = 1540°F$, its pressure is 1762 psi. The mercury–potassium eutectic has a much lower vapor pressure at a specified temperature than pure mercury has. At $T = 1540°F$, its vapor pressure is 198 psi.

Of the alkali metals, the fluid which can be used most advantageously in the one-component cycle appears to be potassium. A maximum cycle efficiency of ~20 per cent seems feasible over the temperature range of 2240–1160°F. Although the cycle efficiency with cesium as the working fluid is essentially the same as for a potassium system, the poorer electric conductivity of cesium reduces its effectiveness in the MHD generator. Cesium has the lowest conductivity whereas sodium has the highest. Sodium as a working fluid however yields the lowest cycle efficiencies.

Condensing injector cycle. Data on the efficiency of the condensing injector cycle for mercury and potassium are shown in Figs. 11 and 12. The cycle efficiency and generated stagnation heads of the condensing injector at positions y and 0 are plotted vs. per cent of the maximum contraction ratio. Position y refers to the minimum cross-sectional area of the condensing injector and position 0 refers to the exit of the injector. Also shown is the subcooling at the injector exit. The maximum contraction ratio is calculated from the arbitrary stipulation that the minimum cross-sectional area in the mixing section in the injector is equal to the cross-sectional area at the exit of the inlet liquid nozzle. The decision to plot the calculated performance data against the per cent of maximum contraction ratio was based upon the fact that the contraction ratio is perhaps the most important variable affecting the condensing injector performance, and hence the cycle efficiency. Other important independent parameters shown on the curves are the inlet liquid to vapor pressure ratio B, mass flow ratio of liquid to vapor R, and source and sink temperatures. For all the computations it was assumed that the MHD generator is 80 per cent efficient and that the condensing injector performance factor is 0.80. The condensing injector performance factor is defined as

$$\eta_c = \Delta P_{act}/\Delta P_{cal} = (P_{out} - P_{liq\,in})_{act}/(P_{out} - P_{liq\,in})_{cal} \quad (1)$$

where

P_{out} = outlet pressure of the condensing injector
$P_{liq\,in}$ = inlet liquid pressure to the condensing injector

The maximum efficiencies of the condensing injector cycle are slightly higher than the maximum efficiency of the two-phase single-component cycle and, in fact, are the highest of the four cycle studies. The cycle efficiencies are listed in Table I. These are approximate values that have been taken from numerous curves, such as Figs. 11 and 12. The maximum cycle efficiencies are those values shown on the figures beyond which the second law of thermodynamics is violated in the condensing injector as denoted by heavy broad lines. It is interesting to note that for a given value of R, the maximum cycle efficiency is virtually independent of the inlet stagnation pressure ratio B for a given fluid at fixed source temperature

and sink temperature, T_1 and T_5 respectively. Also, it can be seen that the maximum cycle efficiency increases slightly with an increasing flow rate ratio R. Referring to Table I, it is apparent that mercury as a working fluid in the condensing injector cycle produces the highest cycle efficiency at 30 per cent, followed closely by cesium and potassium at 27 per cent. Sodium shows the lowest performance potential. The ratio of the cycle efficiency to Carnot efficiency is in the range of 0.4 to 0.5, which, again, is comparable to the two-phase one-component cycle.

Some interesting trends derived from the cycle analysis, which are typical and noteworthy, are: (1) The cycle efficiency increases with increasing contraction ratio of the injector, increasing inlet stagnation pressure ratio B, and increasing flow-rate ratio R. There are, however, finite limitations in the variation of these parameters imposed by considerations of the second law of thermodynamics. (2) The cycle efficiency increases slightly with increasing sink temperature and decreases with decreasing heat source temperature. (3) In order to achieve conditions at the exit of the condensing injector where the greatest percentage of the total stagnation head is due to the kinetic head, the geometry of the injector must have a contraction ratio that is greater than 50 per cent of the maximum contraction ratio. Conditions of high kinetic head and low static pressure are desirable in both injector and MHD generator design. If a substantial portion of the total stagnation head is kinetic, the MHD generator may be of the variable area type, and thus be designed for much lower pressures—which may be an important factor in large-scale commercial plants due to the elevated temperature of the system and component sizes. (4) High contraction ratio condensing injector geometries are not required to obtain the maximum cycle efficiencies. The high efficiencies can readily be obtained at low and medium contraction ratios by increasing the inlet liquid to vapor pressure ratio B. This is relevant since performance of the injector can be expected to deteriorate at the high contraction ratios.

Liquid-metal cycles for commercial application

Overall efficiencies of binary cycles. The overall efficiency and potential of a central station power system employing a liquid-metal MHD topping cycle are excellent. The system mentioned previously consists of

Fig. 11. Efficiency of the condensing injector cycle for mercury. $T_1 = 1960$, $T_5 = 860$, $R = 10$, $\eta_C = 0.8$, $\eta_G = 0.8$. Legend: curves in black show stagnation pressure (P_2) curves in color show pressure at station Y (P_Y).

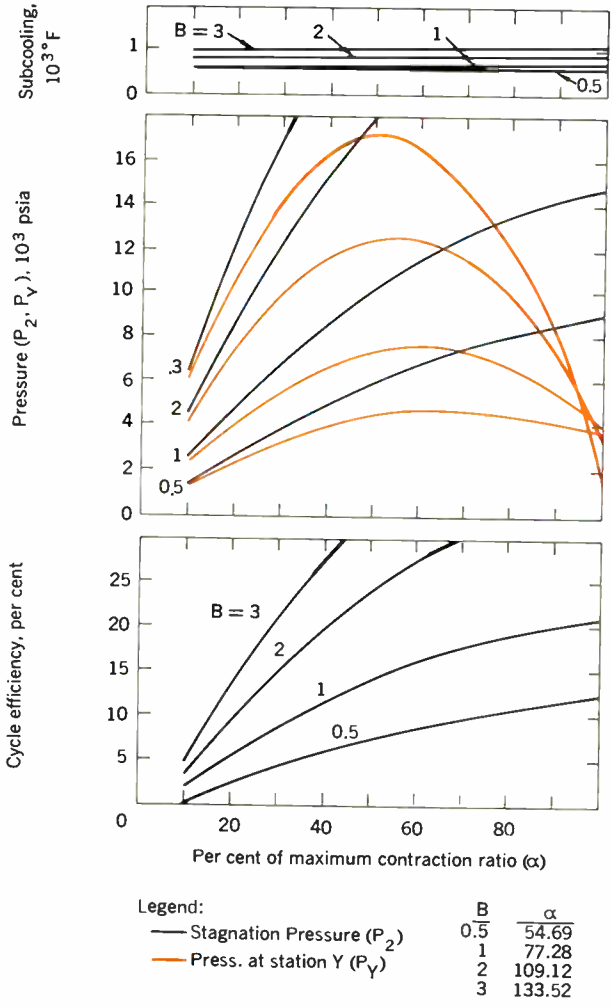
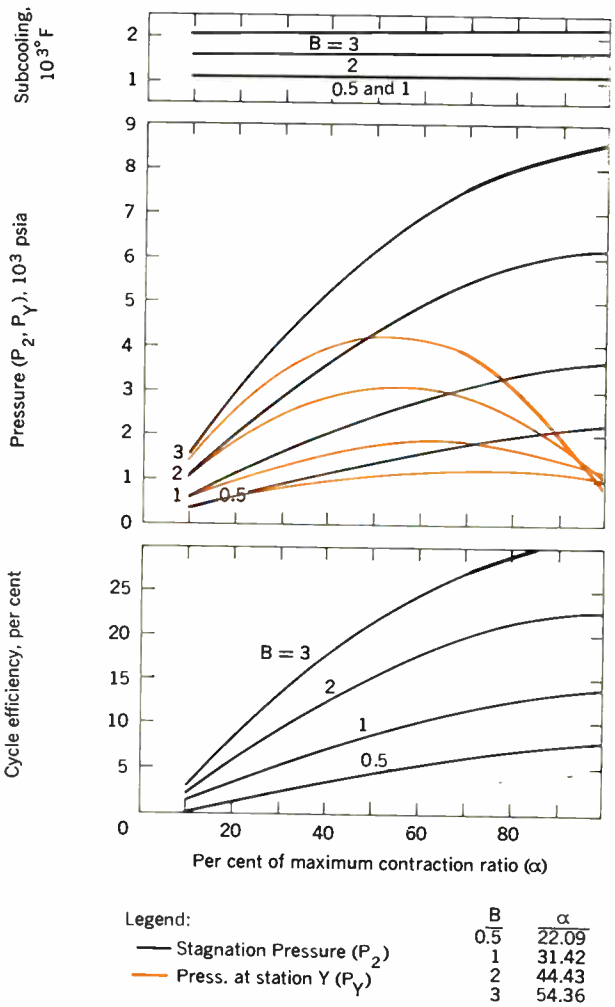


Fig. 12. Efficiency of the condensing injector cycle for potassium. $T_1 = 2700$, $T_5 = 1060$, $R = 10$, $\eta_C = 0.8$, $\eta_G = 0.8$. Legend; curves in black show stagnation pressure (P_2); curves in color show pressure at station Y (P_Y).



I. Maximum cycle efficiencies, per cent

Cesium:

	$T_1 = 2700°$ $T_5 = 1060°$	$T_1 = 2700°$ $T_5 = 1860°$	$T_1 = 2700°$ $T_5 = 1510°$	$T_1 = 2400°$ $T_5 = 1060°$
R = 3	25	7	17	24
= 5	26	11	20	25
= 7	26	13	22	26
= 10	27	14	23	26

Mercury:

	$T_1 = 2000°$ $T_5 = 1060°$	$T_1 = 1860°$ $T_5 = 860°$	$T_1 = 2000°$ $T_5 = 860°$	$T_1 = 1860°$ $T_5 = 1060°$
R = 3	—	22	22	14
= 5	16	25	24	17
= 7	19	26	26	19
= 10	21	27	30	21

Potassium:

	$T_1 = 2700°$ $T_5 = 1060°$	$T_1 = 2700°$ $T_5 = 1860°$	$T_1 = 2400°$ $T_5 = 1060°$	$T_1 = 2400°$ $T_5 = 1860°$
R = 3	22	—	13	—
= 5	24	12	13	6
= 7	26	14	14	7
= 10	27	—	17	—

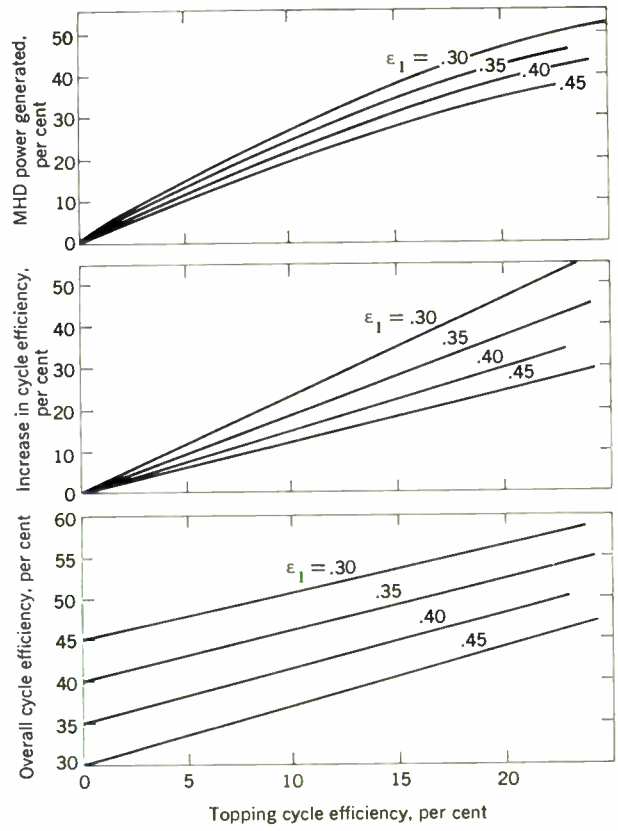


Fig. 13. Curves illustrating topping cycle performance vs. topping cycle efficiency.

a reactor heat source, MHD loop, and normal steam plant, which functions as the sink or bottoming cycle. It should be noted that a fossil-fired boiler using a liquid-metal coolant can also be utilized as the heat source. The heat transfer characteristics of the liquid metal are superior to those of water and the boiler would operate at the highest cycle temperature with a low vapor pressure.

The overall efficiencies of a binary power cycle using an MHD topping cycle were computed by combining the efficiencies of the topping cycle with steam plant efficiencies in the following manner:

$$\epsilon_C = \epsilon_{MHD} + (1 - \epsilon_{MHD})\,\epsilon_{SC} \qquad (2)$$

where

ϵ_{MHD} = efficiency of the topping cycle
ϵ_{SC} = efficiency of the steam bottoming cycle

The overall cycle efficiency percentage increase of the cycle efficiency, and the per cent of the total power generated by the MHD topping cycle are shown in Fig. 13 as a function of the topping cycle efficiency and the base steam plant efficiency. From Eq. (2) it is apparent that the overall efficiency of the binary cycle increases with increasing efficiency of the topping cycle and efficiencies greater than 50 per cent are readily obtained when the base steam plant efficiency $\epsilon_{SC} \geqq 40$ per cent and the topping cycle efficiency becomes 15 per cent or greater. The maximum overall cycle efficiency is generally reached when the secondary steam cycle is operated at maximum allowable temperature conditions. As the sink temperature is raised, the decrease in the efficiency of the topping cycle is more than offset by the increase in the steam cycle efficiency, which results in a maximization of the overall cycle. By combining Eq. (2) with the topping cycle data discussed in the previous sections, the following results can be deduced:

1. If an alkali metal is specified as the working fluid, the maximum potential overall cycle efficiency attainable is between 55 and 60 per cent and results from either a potassium– or cesium–steam binary cycle. The liquid-metal topping cycle operates in a temperature range of 2240–1100°F and produces 35 per cent of the total power. The bottoming steam plant is assumed to be the equivalent of a modern supercritical plant such as the Eddystone unit operating at 4000 psi and 1050°F. The efficiency of such a plant was assumed to be ~0.45 after upgrading the published Eddystone efficiency of 0.407. The upgrading results from the elimination of boiler inefficiency and stack losses.

2. The maximum potential efficiency of the Hg–steam binary cycle is ~56 per cent and is slightly lower than the cesium– or potassium–steam binary cycle. This is based on a source temperature of 1540°F and a condenser temperature of 1100°F. If the sink temperature is dropped to 440°F, the overall cycle efficiency drops to ~47 per cent even though the efficiency of the topping cycle increases substantially. From a thermodynamic viewpoint, mercury is the superior working fluid. However, it has several major drawbacks. Its principal drawback is its rapidly rising vapor pressure at high temperature, as mentioned previously. Other drawbacks are its (1) poor wettability and heat transfer characteristics, which require the use of additives and tend to increase operational problems considerably; (2) limited availability; and (3) relatively high cost.

3. For the medium-temperature range (1100–1600°F), mercury–potassium alloy and potassium appear to be

the most promising working fluids for the liquid-metal cycle. Overall cycle efficiencies of up to 50 per cent appear possible. The MHD topping cycle can have an efficiency as high as 15 per cent and its maximum working pressure would be below 75 psi.

Economic considerations. The attractiveness of the calculated cycle efficiencies is apparent, and so the incentive for development of the cycle must be dictated by strictly economic considerations. Although no detailed economic studies have been made thus far, results of a very preliminary study tend to be favorable. In this rudimentary study, it was assumed that for a fixed plant size the operational and maintenance costs would be unaffected by changes in the plant efficiency. The only costs that would be affected would be the capital expenditures and the fuel costs. As a result, the two major areas in which a dollar savings would accrue due to the introduction of the topping cycle are in capital expenditures for the steam-bottoming plant since its capacity is reduced in proportion to the power generated in the MHD topping cycle, and in fuel costs as a result of the increased efficiency. The dollar differential resulting from these two areas is available for converting an ordinary steam plant to the binary cycle and is, therefore, a measure of the economic incentive.

For illustrative purposes, Fig. 14 shows the total dollar differential which would result in adding an MHD topping cycle to a base steam plant as a function of the fuel costs and percentage increase in the plant efficiency. The values shown pertain to a 1000-MWe plant operating with a load factor of 0.90. The dollar differential is based on an assumed cost of $30/kW for the steam turbine–generator plant and an assumed capitalization charge on the annual fuel savings of 14.7 per cent. The capitalization charge transfers the annual fuel savings to an equivalent capital expenditure. As an example of the dollar differential involved (for an increase in the overall cycle efficiency of 20 per cent) on a base plant whose efficiency is 0.40, the break-even point is ~$25 million for a fuel cost of $0.20/$10^6$ Btu. From Fig. 13 the 20 per cent increase in cycle efficiency can be achieved by an MHD topping cycle whose efficiency is 0.125. Based on the "state of the art," such an efficiency appears realistically attainable at a maximum cycle temperature of 1600°F or less.

Comparison of liquid-metal and plasma MHD cycles

The fact that virtually the entire effort to date in the field of magnetohydrodynamics has been devoted to the plasma MHD cycle has served to pinpoint the major problem areas and limitations of this concept. A comparison of the liquid-metal and plasma MHD power cycles has been made, utilizing pertinent data on the plasma cycle reported in the literature. Gunson *et al.*[9] have described a detailed study of a binary plasma MHD–steam power plant, and a schematic and pertinent power output data from this study are reproduced in Fig. 15. Rosa and Kantrowitz[10] have presented calculated efficiencies for a closed-cycle plasma MHD–steam binary cycle; Fig. 16 is taken from their study. The detailed study by Gunson *et al.* and the cycle analysis reported by Rosa and Kantrowitz were based on a helium cycle; therefore, the efficiencies shown in Figs. 15 and 16 would be expected to be comparable. Gunson reports a net efficiency of 47.1 per cent for a plasma MHD–steam binary cycle, operating with a top temperature of 3000°F. For the same condition, Fig. 16 indicates a value of ~51 per cent. It is evident that before the efficiency of a modern steam plant is surpassed significantly, the temperature of a binary plasma MHD–steam plant must approach 3000°F. Also, to achieve overall cycle efficiencies of 55 per cent, temperatures approaching ~4000°F must be reached. By comparison, as demonstrated in the previous sections, for a liquid-metal MHD–steam binary cycle overall efficiencies of 50 per cent are possible at temperatures of ~1600°F, and efficiencies of ≧55 per cent appear possible with temperatures below 2200°F. It is also interesting to note that the pure regenerative plasma MHD cycle also shown in Fig. 16 must approach temperatures of 4000°F to achieve an efficiency of 50 per cent. In recent studies of the open-cycle combustion-fired plasma MHD cycle, Way and Young[11] indicate that the overall efficiencies of a binary cycle are from 43 to 50 per cent for a maximum temperature in the cycle of 4500°F. The highest efficiency was achieved with a supercritical steam plant as the bottoming cycle.

Regardless of the temperature range specified or the type of cycle, the potential overall cycle efficiency of a liquid-metal MHD–steam binary cycle appears to be at least equal and, in many instances, superior to the plasma MHD binary cycle. When the temperature ranges are taken into consideration, the attractiveness and potential superiority of the liquid-metal MHD cycle are enhanced even further. One of the major factors which tends to reduce the efficiency of the plasma MHD cycle is that it is essentially a Brayton cycle and so a large fraction of the power generated is required for the compressors circulating the gas. As an example, in the study made by Gunson *et al.* about 25 per cent of the total electric power generated was used for the compressors.

In addition to its attractive potential efficiency, the liquid-metal MHD cycle possesses distinct advantages. From electric conductivity and velocity considerations it can be shown that the power density achievable in a liquid-metal generator is approximately an order of magnitude higher than that attainable from the plasma generator. The size of the liquid-metal MHD generator therefore is very much smaller for a specified power output, and thus the cost of the magnet or coil (superconducting or normal) is substantially less than that for a plasma generator. Since the magnet or coil is a major cost item in an MHD plant, the potential savings in this area alone are very important.

To achieve an adequate conductivity of the plasma in the plasma closed-cycle system, it appears that either or both seeding techniques and ultrahigh-temperature heat sources (>2000°K) are required. Although the development of a reactor to operate at a temperature of 2000°K for the extended periods of time necessary in a commercial power plant appears possible, this achievement is not just around the corner. It would require a tremendous concentrated effort to attain this goal. The development of a reactor to operate for extended periods of time at temperatures of 2500°K is not in the foreseeable future. These statements are based on data such as shown in Fig. 1. The nuclear rockets being developed at present are designed for extremely short time periods of operation and therefore the technology is not directly applicable. Referring to Fig. 1, it can be seen that it is a gigantic step to extrapolate the technology to produce a long-life-

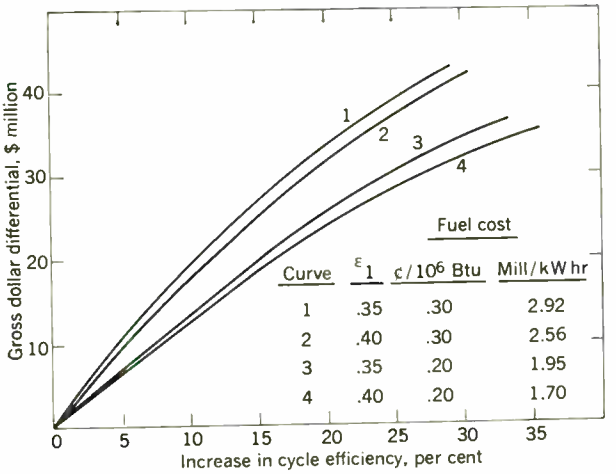


Fig. 14. Dollar differential resulting from addition of MHD topping cycle to a base steam plant.

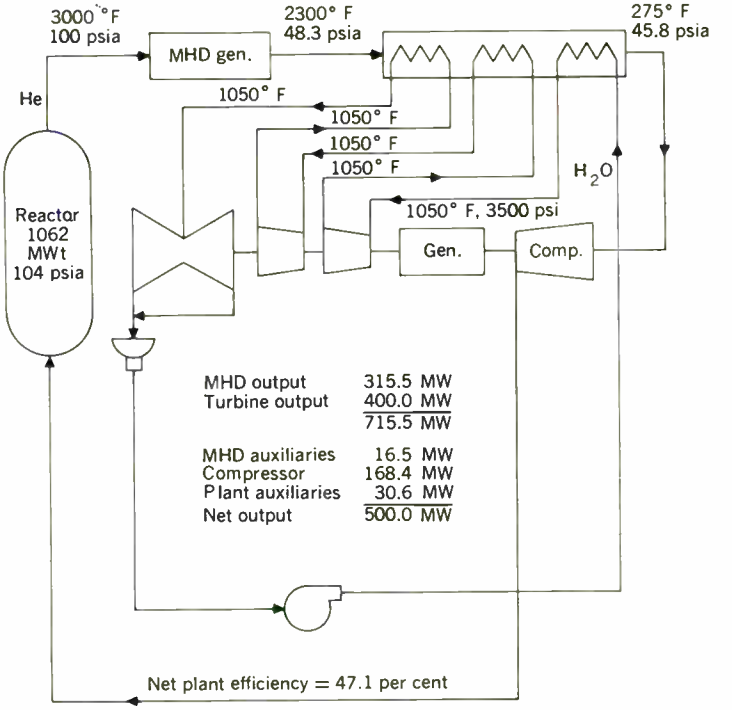


Fig. 15. Schematic diagram and data taken from a study of a binary steam–plasma MHD cycle reported by Gunson et al.

time high-temperature system. For long-term operation there are many unanswered questions in regard to fuel behavior, radiation damage, etc. It should also be noted that an ultrahigh-temperature reactor that could be used in a plasma MHD closed-cycle system would be a very costly item because of the exacting materials requirements.

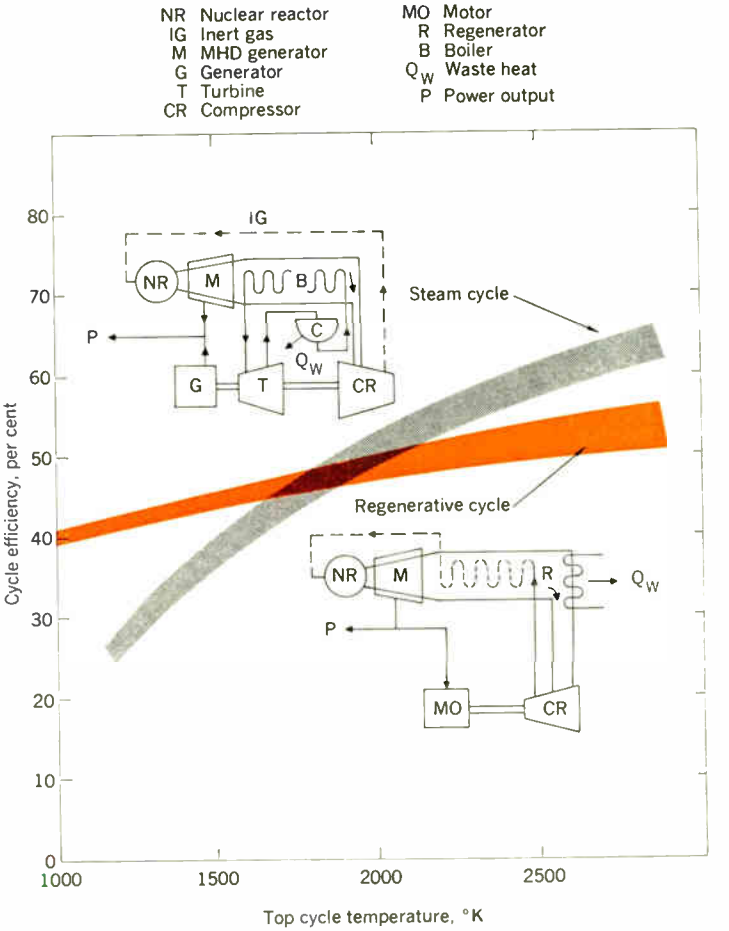
Still another important technical consideration pertains to the direct generation of ac power. The transmission and utilization of electric power is generally done under ac conditions. This consideration, coupled with the fact that dc–ac converters are expensive, makes the possibility of direct ac generation especially attractive. Jackson and Pierson[12] have shown that the operating characteristics of an MHD induction generator can be deduced from the magnetic Reynolds number based on the wavelength of the field structure and the velocity difference between the wave and the fluid. Ionized gas flows in the conductivity range at present attainable have very low magnetic Reynolds numbers, $\sim 10^{-2}$, and therefore direct ac generation with such gases does not appear to be feasible. On the other hand, a characteristic magnetic Reynolds number for a liquid-metal generator would be $\sim 10^{-1}$, and the probability of producing ac power directly is much more promising here.

State of the art

Areas of uncertainty. The calculated cycle efficiencies were based on component efficiencies evaluated from data currently available. Because of the range and extent of the cycle studies, large areas were encountered where extrapolation of the data was necessary and thus there are areas of uncertainty in each of the liquid-metal cycles described. In addition, there exists the materials problem, which is common to all approaches.

The feasibility and true potential of the condensing injector liquid-metal MHD cycle will be determined primarily by the performance of the condensing injector itself. The condensing injector (also called a condensuctor, condensing ejector, and jet pump) is not a new device, having been developed extensively as a boiler

Fig. 16. Efficiency of a closed-cycle plasma MHD-steam binary cycle reported by Rosa and Kantrowitz.[10]



feed-water pump. It has been stigmatized as a very-low-efficiency device, but the apparently low efficiencies may be attributed to the manner in which the efficiency is defined. Recent studies,[13–16] both experimental and analytical, have indicated that the performance of the injector can be very high; more important, analysis predicts that the injector can generate very large stagnation pressures—a mandatory requirement for the liquid-metal MHD cycle. The performance characteristics of the condensing injector were evaluated from the data presented in the above-mentioned studies. The data introduced in the works of Brown, Rose, and Miguel and Brown appear to be consistent in that the ratio of actual performance to predicted performance is comparable. The performance of the condensing injector as mentioned previously is arbitrarily defined by Eq. (1). This performance factor is not an efficiency but a measure of accuracy with which the actual injector pressure performance can be predicted from calculations based on continuity, momentum, and energy equations within the limitations of the second law of thermodynamics. The performance factors calculated from the data of Rose and Miguel and Brown ranged from 70 to 85 per cent for the range of parameters investigated. The data of Hays exhibit much more scatter but are also comparable over certain ranges. The geometries, fluids, and parameter ranges that are of direct application to the liquid-metal MHD cycle have not, however, been covered in these investigations. Therefore, experimental verification of the calculated performance parameters used in the cycle analysis is needed.

To achieve efficient operation of the "drift tube" in the cycle proposed by Prem (Fig. 3), the supersonic vapor must be condensed rapidly with a minimum momentum loss. It has been proposed that this be accomplished by the introduction of the subcooled liquid in stages. The theoretical efficiency of this device must also be verified experimentally.

The feasibility of the two-phase single-component cycle is dependent on the momentum transfer process, flow characteristics, and losses in the MHD generator. To achieve the assumed high-efficiency generator, the momentum of the liquid drops emerging from the nozzle must be transferred with minimum loss to the film, where the electromagnetic forces are encountered. Condensing the vapor phase on the rapidly moving film by cooling the bottom side of the generator is also being studied. If this proves feasible, power generation and condensation would occur simultaneously in one component. The efficiency of the generator will be primarily a function of the degree of separation which occurs and the skin friction losses. The elimination of the electrical shunt losses at the entrance of the generator tends to enhance the efficiency considerably. The major unknown in regard to generator performance, which is being studied experimentally, is the stability of the fluid film.

The materials problem in the liquid-metal MHD topping cycle, as in all high-temperature systems, is a serious one, and is accentuated by the use of liquid metal as a coolant and working fluid. Major problem areas exist relating to liquid-metal chemistry, design of heat-exchanger equipment, materials compatibility, and component development. Although very substantial strides have been made in the development of liquid-metal technology, it still remains essentially empirical in nature. There are, therefore, limitations on the extension of the technology to new fluids, new materials, and higher temperatures. It should be noted that the technology is predominantly based upon sodium because of its preeminence as a reactor coolant.

Liquid-metal MHD topping cycles operating in the 1000–1600°F temperature range appear to be within a reasonable extrapolation of today's state of the art The temperature limitation of the current liquid-metal technology is generally acknowledged to be ~1200°F The austenitic stainless steels have been operated for prolonged periods of time with sodium at temperatures to 1200°F. It appears that the temperature can be raised to the 1500–1600°F range if the system is properly engineered to account for the low strength of the stainless steels at elevated temperatures. The higher temperature operation has been demonstrated; as an example, a sodium loop at ANL logged more than 7000 hours in the temperature range of 1200–1500°F during intermittent operation over a period of four years.

The materials problem for the very-high-temperature (>1600°F) high-efficiency liquid-metal MHD cycles appears to be somewhat more formidable. The behavior of the alkali metals, which look promising as working fluids, is essentially a question mark at elevated temperatures. However, an extensive program is under way that is being sponsored by the United States Atomic Energy Commission for the purpose of developing and demonstrating of an advanced Rankine-cycle reactor power system operating at temperatures to 2000°F. This effort is generally referred to as the "Systems for Nuclear Auxiliary Power" (SNAP). The SNAP-50 program falls in an area that is beyond any prior experience. The program expenditures are expected to be between $1 and $2 billion over an extended period of time. The program is essentially oriented toward the development of advanced high-temperature materials and the key components of the power cycle. (It should be noted that major concern has been expressed about the reliability of the turboelectric power-conversion system.) Another program, the Advanced High Temperature Materials Program, is essentially a support effort for SNAP-50. Because of the magnitude of effort involved, it is apparent that the feasibility of the very-high-temperature (>1600°F) liquid-metal MHD power cycles will be dictated by the sucess or failure achieved in the two programs.

Current programs. Liquid-metal magnetohydrodynamics is essentially in the embryonic stage of its development. In comparison with the plasma MHD effort, the liquid-metal programs are small but appear to be expanding. To the author's knowledge there are five groups currently working in this area: Elliott and co-workers at the Jet Propulsion Laboratory; Prem and associates at Atomics International; Jackson and Brown at the Massachusetts Institute of Technology; Radebold and colleagues at Allgemeine Elektricitats–Gesellschaft in Germany; and a group at the Argonne National Laboratory. A brief description of recent accomplishments and planned programs of each of these efforts follows.

The Jet Propulsion Laboratory program to develop a space power source is perhaps the most comprehensive effort in liquid-metal MHD research that has been reported to date. Extensive nozzle, separator, and diffuser tests have been made utilizing water and nitrogen, in addition to tests on the MHD generator with NaK

as the working fluid. It is reported that the nozzle delivered 90 per cent of isentropic exit velocity, the separator had about 1 per cent liquid loss and 20 per cent velocity loss, the diffuser had about 75 per cent efficiency, and the generator operated at about 50 per cent efficiency. Potential lifetime limitations due to erosion and to insulator loss were also investigated in water-flow and lithium capsule tests, and results indicated negligible material loss up to at least 1000 hours.[17]

The development program at Atomic International is oriented toward studying (1) simultaneous mass heat transfer and momentum transfer process, which occurs in their "drift tube"; (2) the electrical resistivity of two-phase mixtures under varying conditions of pressure, temperature, void fraction, and impurity; and (3) generator types and performance. (Analytical and research work on the development of an MHD electric generator suitable for utility applications is being carried out under sponsorship of the Edison Electric Institute.) The only results that have been reported thus far pertain to the acceleration and momentum transfer process.[18] Studies with sand and air indicate that the acceleration is accomplished very rapidly, reaching 60 per cent of the air velocity at a distance of 30–40 cm from point of injection. To demonstrate the overall power conversion process, a potassium test loop facility has been designed and fabricated and will be placed in operation.

The relatively recent program at Argonne National Laboratory is directed toward studying the performance characteristics of liquid-metal generators and an evaluation of the condensing injector over parameter ranges that are of direct interest to the condensing injector cycle. The generator studies are designed to provide data on the efficiency of the generator utilizing a dispersed two-phase fluid as the entering working fluid. Various generator geometries, angles of incidence of the entering fluid, and phase distributions will be studied. A second study of the MHD generator, to be carried out with a single-phase fluid, has three objectives: (1) to confirm end losses in variable area generators and to establish maximum generator efficiencies; (2) to develop means for minimizing end losses; and (3) to investigate the fluid mechanics of the generator. These investigations are designed for providing sufficient data that the actual efficiencies of the condensing injector and two-phase one-component cycles can more clearly be defined.

Jackson and Brown have concentrated their effort on the development of the condensing injector and the ac generator. Brown has shown that a large pressure rise across the condensing injector is possible providing that a convergent area mixing section is used. More detailed analysis and experiments are planned. Jackson and coworkers[12,18] have developed the theory and outlined the performance limitations of the ac induction generator. At present an experiment has been designed and is being fabricated to obtain extensive data on the performance characteristics of the ac generator.

The program at Allgemeine Elektricitats-Gesellschaft, although it gives indication of being an extensive effort, has not been reported in detail.

Upon completion of these experimental programs, sufficient data and information should be available to make a more thorough appraisal of, and positive judgments on, the merit of the proposed liquid-metal MHD power cycles.

Conclusions

The potential of the liquid-metal MHD power cycle has been found to be excellent, and overall efficiencies of a binary cycle approaching 60 per cent may be feasible. The liquid-metal MHD cycle appears to be superior to the plasma cycle from several technical and economic viewpoints. There are, however, several major areas of uncertainty pertaining to the liquid-metal MHD concept that can only be resolved through experimentation. Upon completion of current experimental programs, sufficient information should be available to enable a more thorough analysis of the true potential of the liquid-metal MHD concept and to demonstrate the economic incentive for its development.

The author wishes to express his appreciation to David Elliott, Jet Propulsion Laboratory; W. D. Jackson, Massachusetts Institute of Technology; and L. L. Prem, Atomics International, for their criticisms, comments and for their cooperation in making recent data and information available for incorporation in this article. The author is also indebted to M. Sims, Argonne National Laboratory, for preparation of the figures, and to Mrs. M. Luebs for typing the manuscript.

REFERENCES

1. "An Evaluation of Systems for Nuclear Auxiliary Power," TID-20079, U.S. Atomic Energy Commission, Jan. 1964.

2. Elliott, D. G., "Two-Fluid Magnetohydrodynamic Cycle for Nuclear-Electric Power Conversion," *ARS J.*, June 1962.

3. Prem, L. L., and Parkins, W. E., "A New Method of MHD Power Conversion Employing a Fluid Metal," *Paper no. 63*, Internat'l Symposium on MHD Elec. Power Generation, Paris, France, July 1964.

4. Petrick, M., and Lee, Kung-You, "Performance Characteristics of a Liquid Metal MHD Generator," ANL-6870, Argonne National Laboratory, Argonne, Ill., July, 1964.

5. Jackson, W. D., and Brown, G. A., "Liquid Metal Magnetohydrodynamic Power Generator Utilizing the Condensing Ejector," Patent disclosure, M.I.T., Cambridge, Mass., Oct. 1962.

6. Prem, L. L., private communication.

7. Petrick, M., and Lee, Kung-You, "Liquid Metal MHD Power Cycle Studies, ANL-6954, Argonne National Laboratory.

8. Elliott, D. G., and Cerine, D. J., "Liquid MHD Power Conversion," in "JPL Space Programs," Quarterly Rept., vol. III, summary no. 37-17, Deep Space Instrumentation Facility, Pasadena, Calif., Oct. 1962.

9. Gunson, W. E., *et al.*, "MHD Power Conversion," *Nucleonics*, vol. 21, July 1963.

10. Rosa, R., and Kantrowitz, A., "MHD Power," *Intern. Sci. and Tech.*, Sept. 1964.

11. Way, S., and Young, W. E., "The Feasibility of Large Scale MHD Power Generation," *Paper no. 98*, Internat'l Symposium on MHD Elec. Power Generation, Paris, France, July 1964.

12. Jackson, W. D., and Pierson, E. S., "Operating Characteristics of the MHD Induction Generator," Magnetoplasmadynamic Elec. Power Generation, Conf. Rept. Series no. 4, IEE, London, England, 1963.

13. Brown, G. A., "An Analysis of Performance Data from the NUOS Condensuctor Test Facility with a New Theory for the Variable-Area Condensuctor," Rept. 44, Joseph Kay and Co., Inc., Cambridge, Mass., 1961.

14. Miguel, J., and Brown, G. A., "An Analytical and Experimental Investigation of a Condensing Ejector with a Condensable Vapor," *Paper no. 64-469*, 1st AIAA Annual Meeting, Washington, D.C., June 29–July 2, 1964.

15. Hays, L., "Investigation of Condensers Applicable to Space Power Systems—Part II, Jet Condensers," Rept. 1588-Final II, Elec. Optical Systems, Nov. 1962.

16. Rose, R. P., "Steam Jet Pump Analysis and Experiments," Tech. Rept. WAPD-TM-227, Bettis Atomic Power Laboratory, Pittsburgh, Pa., June 1960.

17. Elliott, D. G. Cerini, D. J., and Weinberg, E., "Investigation of Liquid MHD Power Conversion," presented at Aerospace Power System Conf., Philadelphia, Pa., Sept. 1946.

18. Jackson, W. D., Pierson, E. S., and Porter, R. P., "Design Considerations for MHD Induction Generators," *Paper no. 61*, Intern. Symp. on MHD Elec. Power Generation, Paris, France, July 1964.

Worst-case circuit design

Worst-case design aims for reliable circuit operation even under extreme conditions, including components degrading to end-of-life limits. Examination of pertinent factors is based on logic switching circuits but is applicable to all circuits

J. B. Atkins *IBM Components Division*

The philosophy of worst-case design is a way of life for many circuit designers. Its purpose is reliable circuit operation to meet input–output speed and noise-rejection requirements even under extreme (worst-case) conditions—including circuit components degrading to end-of-life limits. Such factors as temperature extremes, noise levels, power supply variations, maximum power dissipation, and maximum device ratings must all be considered in worst-case circuit design.

In this article, the aforementioned factors are examined on the basis of a design of logic switching circuits, which, however, is applicable to any type of circuit. Specific illustrations relating to several switching circuits are used to emphasize the importance of the various factors and suggestions for computer-programmed design and analysis will be discussed.

Finally, reasons for using a worst-case design philosophy and some means by which this philosophy may be supplemented are set forth.

A circuit that has been worst-case designed is guaranteed to meet all performance requirements, even under extreme worst-case conditions. These conditions, however, do not include such catastrophic failures as short-circuited diodes, open resistors, or malfunctioning power supplies. What then are the performance requirements and what design factors must be considered?

Performance requirements

Fan-out capability. Under worst-case conditions, the fan-out of a given circuit is defined as the number of inputs of like circuits that can be driven at the output. (Note that actual load currents are not important under this definition.) A typical number of required fan-outs ranges from three to ten logic circuits.

There may be situations where the fan-out requirement is stated as a certain minimum current. An example is a circuit that must drive a terminated transmission line and must have a certain fan-out when it is not driving a line.

Fan-in capability. Most systems requirements include a minimum fan-in limit. Fan-in is simply the number of logical circuit inputs. In some circuits, the number of inputs may limit the fan-out that can be obtained and determine the noise-rejection level of a circuit, which is usually a direct function of the number of inputs. In general, circuits with diode or transistor input gates are insensitive to fan-in when compared with circuits with resistor inputs.

Maximum delay or switching time. Worst-case design is primarily concerned with dc conditions. However, usually there are some dc conditions that must be satisfied to meet specified limits for delay. The most obvious requirement is to maintain some specified OFF drive current to control the storage delays of a saturated transistor. This requirement may take the form of a maximum limit for the ratio of ON drive to OFF drive. Another typical requirement is for an overdrive of base current sufficient to meet dc fan-out requirements plus any capacitive current that must be driven to maintain a required output voltage transition.

Minimum noise rejection level. This requirement is the most difficult to specify or satisfy because of the unpredictable nature of noise generation, both in amplitude and time duration. There are several sources of noise; the more important and typical are outlined in the following.

Coupled noise. Coupled noise that arises when switching lines are adjacent to input lines is a vital source of input noise. The amplitude and/or pulse width must be carefully controlled or allowed for in the circuit design. Both of these factors are important because a circuit with finite switching delays will be insensitive to noise pulses of relatively short duration.

For long-line cases, this problem can often be solved

only by the use of coaxial cables or some other form of shielded line.

Transients on the power supply lines. The power supply tolerance is stated typically in terms of dc variations only. Transients may appear on the lines as a result of current surges. For circuits with delays measured in nanoseconds, the transients may become quite serious. These transient variations also include shifts in the ground level, which can be significant on large systems with several different and separated subassemblies.

Logical noise. This source of noise is often overlooked but at times may be very significant. Figure 1 shows a typical example: a simple diode–transistor AND–INVERT logic circuit. Initially, both inputs are at a down level with diode D_1 conducting all of the input current, which is possible with variations in V_{CE} saturation. A rise in the input voltage to diode D_1 to $+V$ volts should not change the voltage of the collector of T_1. But consider what can happen if diode D_1 has a poor diode recovery characteristic and a high junction capacitance. This allows a pulse of current (much greater than I_{IN}) to be coupled through diode D_1 in the reverse direction. If the line length L and the associated inductance between diode D_2 and the driving collector are appreciable, the voltage at point A will rise and current will flow to the base of T_1. If this current has the necessary time duration and amplitude, transistor T_1 will turn on and a noise pulse will appear at the collector of T_1. The direction of the reverse current flow and the resulting noise are shown by dashed lines in Fig. 1.

The aforementioned noise sources can present a problem even when considered separately. When combined, these effects can be overwhelming. In any event, it is obvious that the dc design of any circuit must include a careful consideration of noise sources and noise rejection level.

Design factors

To guarantee circuit performance, the circuit designer must consider all design factors. There are two points to keep in mind: (1) good circuit performance can usually be achieved quite easily if tight tolerances are maintained on all design factors; and (2) tight tolerances lead to expensive circuits. Thus, the circuit designer must be aware of the economic trade-offs of performance vs. cost.

The following are the various design factors that generally must be considered.

Resistor tolerances. The resistor tolerances may vary with age, temperature, humidity, and sometimes may also vary with voltage. All of these effects must be taken into consideration.

Power supply tolerances. The dc tolerance on a power supply voltage as delivered to the circuit in question is usually divided into two parts: the variations a the supply terminals, and distribution losses between the power supply terminals and the circuit. As noted previously, consideration must be given to transient variations as well as to dc variations.

Semiconductor device tolerances (transistors and diodes). *Transistor dc beta* (β). Beta is typically the single most important parameter to be considered in any circuit design. This is especially true of single transistor circuits. The β of a transistor usually varies as the operating points vary and as the temperature varies. These variations may be drastically different from one unit to the next. In addition, the β variation for a given type of transistor may be quite extreme for any single condition. Maximum limits of five to ten times the minimum limits are not uncommon.

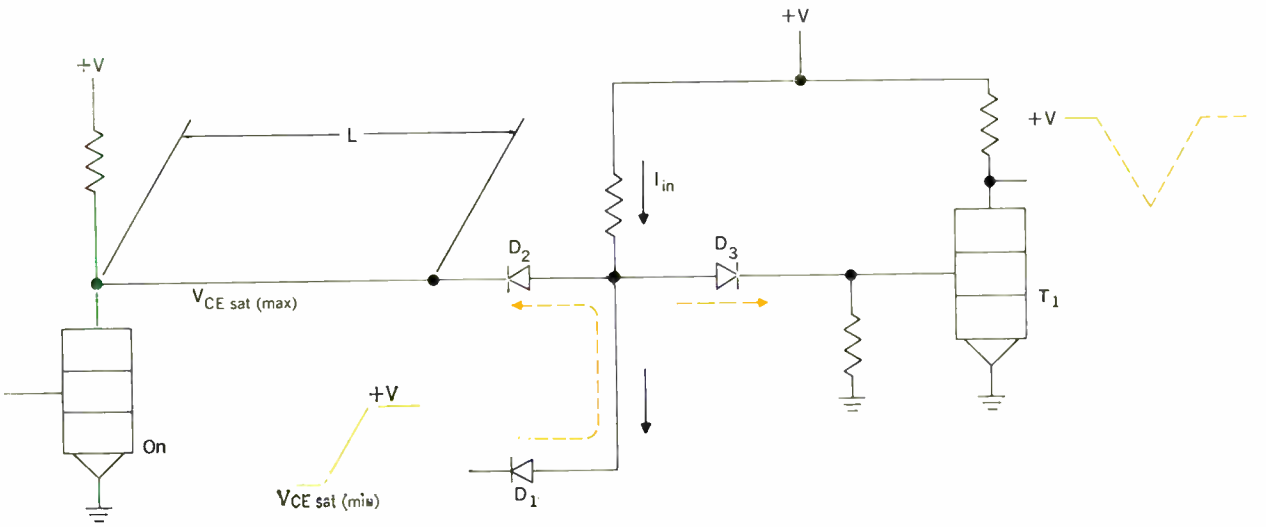
V_{BE} vs. I_c. V_{BE} is a direct function of temperature for silicon transistors. The variation is approximately -2 mV per degree centigrade, but this factor is not constant with large variations in current and temperature. In addition to temperature variations, V_{BE} variations for one transistor type at any current level may exceed 50 mV.

V_{CE} saturation. This parameter will change with operating point as well as with temperature, and the variations may be significant. Any resistance such as collector contact resistance effectively in series with the collector will cause an increase in V_{CE} as the collector current increases.

Diode forward voltage. The comments concerning V_{BE} are also applicable in this case for silicon diodes.

Transistor and diode leakage currents. These factors

Fig. 1. Logical noise situation.



are usually not important for silicon devices, but should always be checked.

Factors relating to transient performance. Factors such as storage time, transistor cut-off frequency, diode recovery times, and junction capacitances are not directly applicable for worst-case design. However, they may indirectly influence the dc design.

Ambient temperature extremes. The range of operating temperatures must be considered because some parameters vary with temperature. The temperature differentials between circuits may be of more importance than the absolute temperature.

Noise levels. Consideration must be given to all sources of noise in line with the performance requirements.

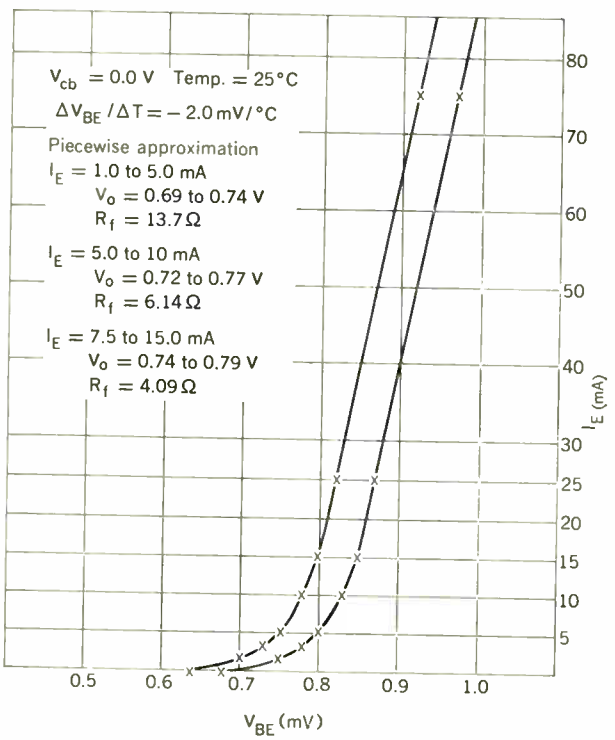
Maximum power dissipation per circuit. As circuit density increases, the power dissipation per circuit must be considered carefully. Power dissipations that are too high may lead to the design of expensive cooling schemes, or very high operating temperatures if cooling is not provided.

Maximum junction temperature. This factor is related to the maximum ambient temperature and the maximum power dissipation. A device dissipating a high average power may have a junction temperature significantly higher than ambient temperature. Reliability ratings for devices generally dictate careful control of junction temperatures.

Number and magnitude of power supplies required and/or available. If the circuit designer has any choice, he must carefully consider the cost vs. performance trade-offs involved.

In addition to the factors previously mentioned, the designer must also be aware of all maximum ratings for devices being used to which the circuit design must then conform. Typical ratings are for maximum current, reverse voltage, and power dissipation.

Fig. 2. Typical silicon transistor plot.



Design examples

In the foregoing, the typical performance requirements as well as the design factors that must be considered have been compiled. The following specific examples will serve to illustrate the importance of the various design factors. It is presumed that these examples will indicate how a typical design problem might be attacked.

In Figs. 2 and 3, which show diode forward drops and V_{BE}, the curves are for silicon devices. Note that the trace for V_{BE} vs. I_E is plotted with $V_{CB} = 0$. Thus, the transistor is not saturated, which will normally be satisfactory since most circuits are usually somewhat overdesigned.

Example 1. The first example covers a simple *RTL*

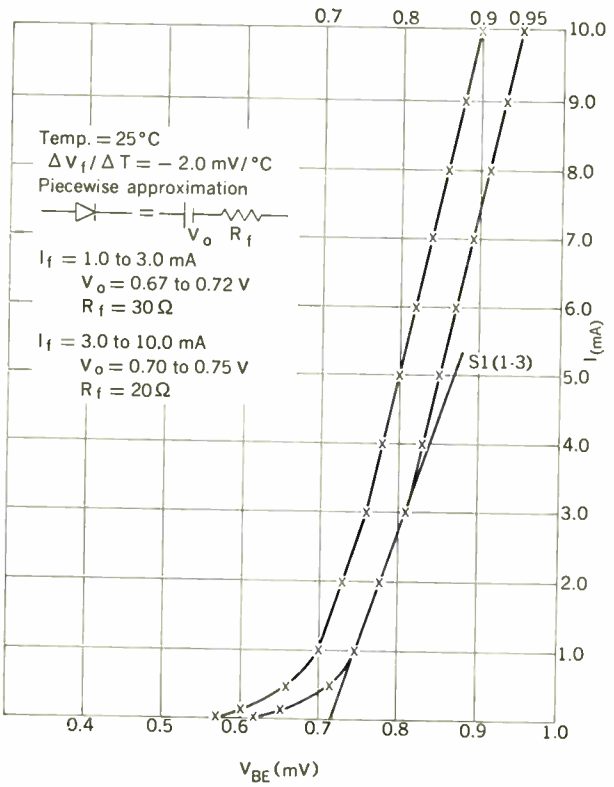
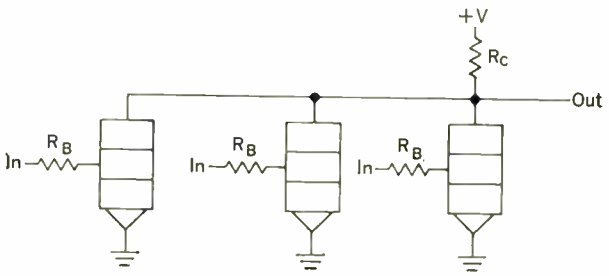


Fig. 3. Typical silicon diode plot. Composite curve plotted from a group of diodes. Resulting curves slightly altered to provide linearity.

Fig. 4. Basic RTL circuit.



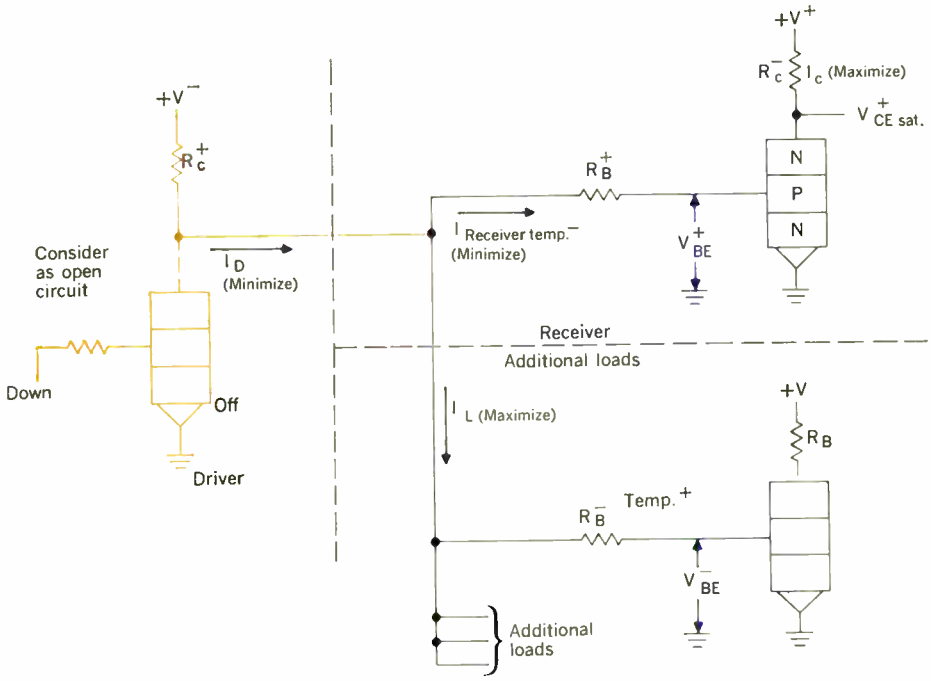


Fig. 5. Circuit for assigning worst-case values. The + and − superscripts indicate how to assign tolerances.

circuit, a basic circuit that is readily understood and can be used to illustrate the importance of several design factors (see Fig. 4). An up-voltage level at any input will cause the output to go to a down-voltage level; consequently, the output will be at an up-voltage level only if all inputs are at a down-voltage level. The number of inputs is not critical if silicon transistors are used, since leakage currents are low. Fan-out is limited by several design factors and is especially sensitive to beta. The circuit will be analyzed to find what ratio of R_c to R_B will allow a minimum beta, starting with nominal conditions and proceeding in a cumulative fashion as various worst-case limits are considered.

However, first let us consider how to assign worst-case tolerances to the various design factors. In Fig. 5, which is broken into three segments, the first segment is the driver circuit where only $+V$ and R_c are of importance since all input transistors are off and we will assume negligible leakage currents. The remaining segments are the receiver circuit and additional load circuits where almost all factors are important and are included. The rules for assigning worst-case values can be stated as: (1) minimize driver current; (2) maximize collector (and base) current required by receiver; (3) maximize factors that impede current flow to receiver; and (4) maximize current drained by additional loads.

The first example will assume operation with a collector current of about 10 mA, which will allow constant values to be used for V_{BE}, greatly simplify the analysis, and impose no limit on the example. A suggested piecewise approximation scheme for handling variable V_{BE}'s will be discussed later.

The design requirements for the example (see Fig. 6) are (1) minimum fan-out of 5, and (2) overdrive factor for base current of 1.2. This 20 per cent extra base current is assumed to be needed to provide noise tolerance.

Example 1-A. All factors are nominal. R's = ±0.0 per cent; $+V$ = 6.0 V; V_{BE} = 0.83 V at 25°C; temperature = 25°C; delta temperature = 0°C; V_{CE} saturation = 0.3 V.

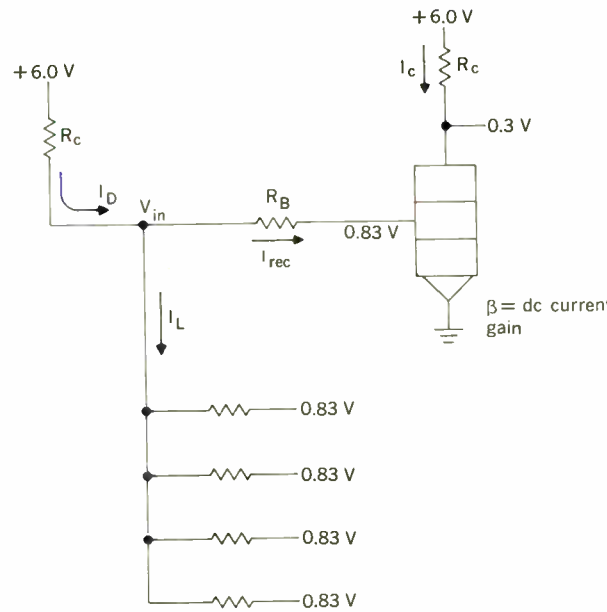


Fig. 6. Basic circuit for Example 1.

$$I_c = \frac{(6.0 - 0.3)}{R_c} \tag{1}$$

$$I_{REC} = \frac{(1.2)\,(I_c)}{(\beta)} = \frac{(1.2)}{(\beta)} \cdot \frac{(5.7)}{(R_c)} \tag{2}$$

where (1.2) is the overdrive factor

$$V_{IN} = 0.83 + (I_{REC})(R_B) \tag{3}$$

$$V_{IN} = 0.83 + \frac{6.84\,R_B}{\beta R_c} \tag{4}$$

Equation (4) for V_{IX} is derived by working back from the collector of the receiver circuit.

Next we write the equation for V_{IX} based only on consideration of the driver circuit and additional loads.

$$V_{IX} = \frac{\frac{6.0}{R_c} + (5) \cdot \frac{0.83}{R_B}}{\frac{1}{R_c} + (5) \cdot \frac{1}{R_B}} \tag{5}$$

Equating (4) and (5) and solving for β yields

$$\beta = \frac{34.2 + \frac{6.84\ R_B}{R_c}}{5.17} = 6.61 + 1.32 \frac{R_B}{R_c} \tag{6}$$

(See Table I for β values for R_c/R_B ratios.)

Example 1-B. All factors are the same as in Example 1-A except resistor tolerance, which is ±5 per cent (worst case, end-of-life limit). This yields

$$\beta = 7.96 + 1.465 \frac{R_B}{R_c} \tag{7}$$

(See Table I for β values for R_c/R_B ratios.)

Example 1-C. Same as Example 1-B except that the power supply tolerance is ±10 per cent. This yields

$$\beta = 9.41 + 1.7 \frac{R_B}{R_c} \tag{8}$$

(See Table I for β values for R_c/R_B ratios.)

Example 1-D. Same as Example 1-C except that V_{BE} has a spread of 50 mV from minimum to maximum. This yields

$$\beta = \frac{43.2 + 7.96\ R_B/R_c}{4.36 - 0.21\ R_c/R_B} \tag{9}$$

(See Table I for β values for R_c/R_B ratios.)

Example 1-E. Same as Example 1-D except that the temperature varies from 0 to 125 °C. This yields

$$\beta = \frac{43.2 + 7.96\ R_B/R_c}{4.31 - 1.16\ R_c/R_B} \tag{10}$$

(See Table I for β values for R_c/R_B ratios.)

Example 1-F. Same as Example 1-E except that $+V = +3$ V. The ±10 per cent variation is included. This yields

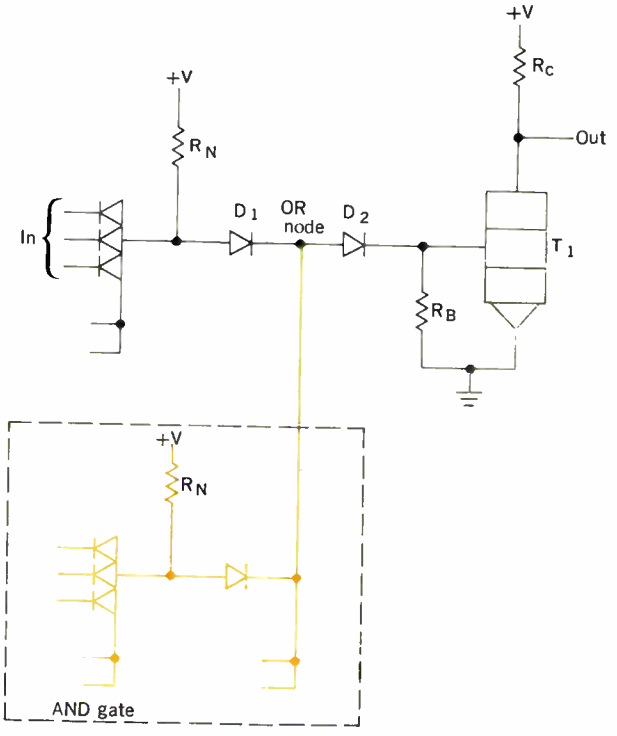
$$\beta = \frac{22.6 + 4.17\ R_B/R_c}{1.73 - 1.16\ R_c/R_B} \tag{11}$$

(See Table I for β values for R_c/R_B ratios.)

I. β values for R_c/R_B

	R_c/R_B	β
Example 1-A:	0.10	19.81
	0.20	13.21
	0.25	11.99
	0.50	9.25
	1.00	7.93
	1000.00	6.62
Example 1-B:	0.10	22.61
	0.20	15.28
	0.25	13.82
	0.50	10.89
	1.00	9.42
	1000.00	7.97
Example 1-C:	0.10	26.41
	0.20	17.91
	0.25	16.21
	0.50	12.81
	1.00	11.11
	1000.00	9.41
Example 1-D:	0.1	28.3
	0.2	19.2
	0.5	13.9
	1.0	12.3
	2.0	11.98
	2.5	12.05
	5.0	13.5
	10.0	19.4
	15.0	36.0
	20.8	∞
Example 1-E:	0.10	29.2
	0.20	20.3
	0.50	15.9
	1.00	16.2
	2.00	33.7
	3.00	57.4
	3.72	∞
Example 1-F:	0.10	40.0
	0.20	28.9
	0.25	27.3
	0.40	26.0
	0.50	26.9
	1.00	47.0
	1.25	92.5
	1.49	∞

Fig. 7. Basic DTL circuit.



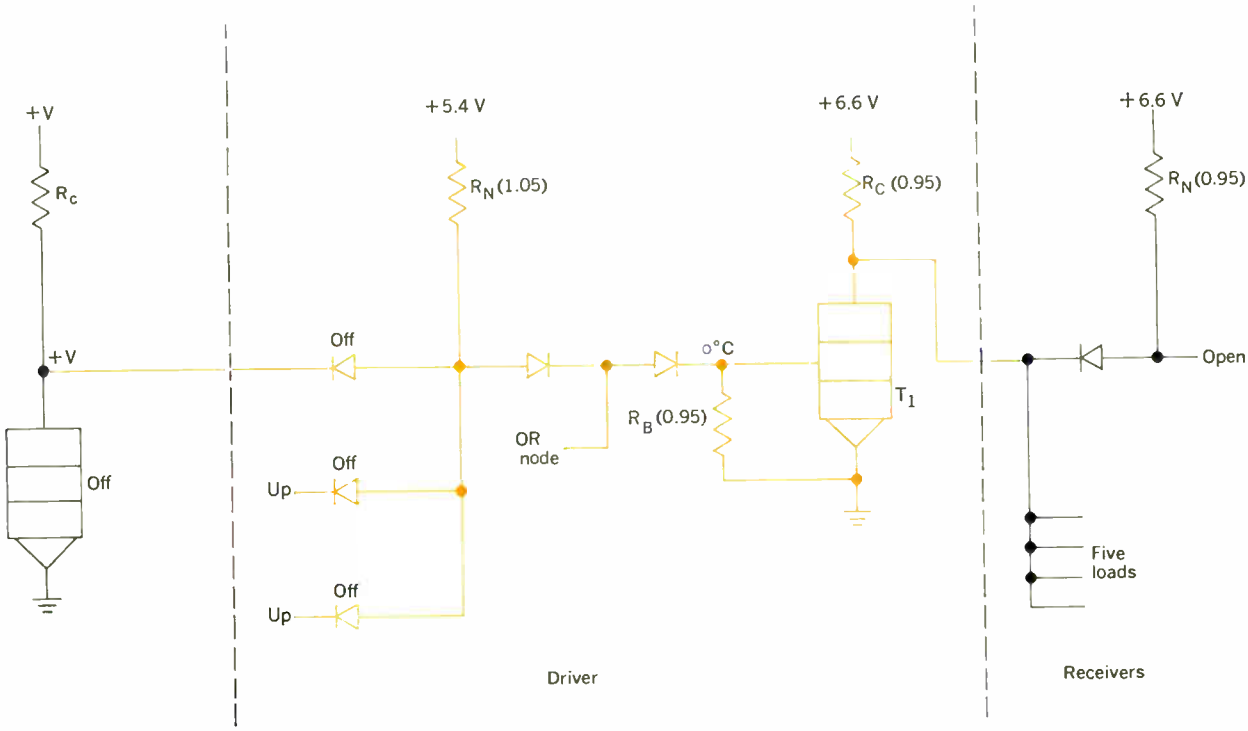


Fig. 8. Schematic drawing of basic circuit for assigning worst-case values.

Example 1-A shows that for nominal conditions the minimum β required is only 6.62 as the ratio of R_c to R_B approaches infinity. This is to be expected since the fan-out of 5 and the overdrive factor of 1.2 will require a minimum β of 6.0 even under ideal conditions. The minimum β required would go lower if $+V$ were higher than +6 V so that V_{BE} and V_{CE} saturations become less significant.

Examples 1-B through 1-F illustrate what happens to the minimum β requirement as additional design factors are considered. Note that not only does the minimum β increase but that the ratio of R_c/R_B becomes increasingly more critical, hence failure to consider all design factors could quickly lead to disaster.

Example 2. The circuit of Fig. 7 is logically more powerful than that of Example 1 since use of diodes as gating elements allows an AND-OR-INVERT function by using only one transistor. In cases where diodes are less expensive than transistors, this circuit may be much more attractive than the circuit that is shown in Fig. 1.

Let us examine this circuit in the same manner as that of Example 1. Before starting calculations, note that diode D_2 is necessary to maintain a suitable noise-rejection level (this will be justified later) and that resistor R_B is required for off bias since diodes D_1 and D_2 will be reverse biased when transistor T_1 is off.

As in Example 1, the design requirements will be (1) minimum fan-out of 5, and (2) overdrive factor of 1.2. Normally, considerations for delay require that R_c and R_B conform to certain conditions. We will not specify these initially, but instead will examine their effect.

All design factors are comparable to Example 1-E: resistors ±5 per cent; power supply = +6 V ± 10 per cent; diode spread = 50 mV; V_{BE} spread = 50 mV; temperature = 0°C to 125°C (ΔT = 125°C); and V_{CE} saturation = 0.3 V.

Example 2-A. First let us see what minimum β is required if R_B and R_c are not considered. Figure 8 indicates how worst-case tolerances have been assigned. The rules for assigning these tolerances are derived from the requirement that T_1 turn on when all inputs of any AND gate are at an up level; consequently, only one of the AND gates that might be present at the input of the driver need be considered. The rules can be stated as follows:

1. Minimize drive current going to T_1.
2. Maximize load current that T_1 must switch. Note that the AND gate being considered might not be at the same location as T_1. Thus, we take two different tolerances on the $+V$.

Writing and solving all necessary equations yields

$$\beta = \frac{34.8\ R_N + 3.20}{2.81\ R_N + 0.53} \qquad (12)$$

A close inspection of Eq. (12) shows that for minimum β the value for R_N must approach infinity. This is not a surprising result in view of what is known about current sources. The actual minimum value for β is then 12.4. This approaches the minimum $\beta < 12.0$ for the *RTL* circuit of Example 1-E.

Example 2-B. Let us examine the bias resistor R_B and the collector resistor R_c. R_c can be handled by considering it as a fractional or multiple fan-out, and thus it can be included as part of the load current. The equations must be rewritten to consider R_B. For this exercise, let R_c equal one fan-out, and let R_B be a variable. This yields

$$\beta = \frac{41.74}{\left[(0.95\ R_N + 0.31)\left(\frac{2.96\ R_N - 0.074}{R_N\ (R_N + 0.092)} - \frac{0.84\ R_N + 0.208}{R_B\ (0.95\ R_N + 0.031)}\right)\right]} \quad (13)$$

$\beta = 20.3$ when $R_B = 2.0$ kΩ and $R_N = 1.0$ kΩ

This compares with a requirement of $\beta = 13.4$ for like conditions of Example 2-A.

Example 2-C. Next, calculate what the noise rejection level is for the worst-case situation shown in Fig. 9. The noise-rejection level can be defined in terms of unity gain for an infinite chain of similar circuits. There is, of course, a different level for positive- or negative-going noise. For this particular circuit, it is obvious that the noise-rejection level is lowest for positive-going noise pulses.

At point *A*, we wish to find what value of V_{IN} results in generating an identical voltage to that at point *B* to satisfy the condition for unity gain. Then, subtracting V_{CE} saturation from this V_{IN}, we will obtain a value for the worst-case noise-rejection level for positive-going noise. Note that additional logic inputs or temperature differentials that may exist between diodes in the same circuit are not under consideration. These factors would be detrimental.

For calculations use: $R_N = 1$ kΩ; $R_c = 1$ kΩ; $R_B = 2$ kΩ; resistor tolerance = ±5 per cent; power supply = +6 V ± 10 per cent; V_{BE} spread = 50 mV, $V_{BE}/\Delta T = -2$ mV/°C; diode V_F spread = 50 mV, $V_F/\Delta T = -2$ mV/°C; $\beta = 20.0$ to 100.0; V_{CE} saturation = 0.3 V maximum; temperature = 0°C to 125°C.

The solution is arrived at by an iterative process that starts at point *B*. Let us begin the iterations by assuming that $V_{OUT} = 1.0$ V, which is the magic level.

Working backward from point *B*, the first iteration yields a value of 0.76 V for V_{IN}. Thus, the first guess of 1.0 V is slightly high, but the second guess for V_{OUT} will be 0.76 V. This method of attack insures a rapid convergence of the iterations. In actual practice, however, two iterations are usually sufficient, the results of the second iteration giving $V_{IN} = 0.76$ V. Thus, the assumption of 1.0 V was a good starting value.

Subtracting the V_{CE} saturation of 0.3 V from the V_{IN} of 0.76 V results in a noise-rejection level of +0.46 V. The removal of one of the two series diodes would reduce the noise-rejection level by 0.435 V, resulting in a noise-rejection level of only 0.025 V, normally not acceptable. So, the use of the two series diodes has been justified solely on the basis of maintaining an adequate noise-rejection level.

Example 3. The circuit[1] shown in Fig. 10 performs the same logical function as the circuit of Example 2. Diode D_3 provides feedback from the input to the output that prevents transistor T_2 from saturating. Transistor T_1 is used to provide current gain for the input network. A closer examination of this circuit reveals that with the exception of diodes and V_{BE} voltage spreads, there are no critical tolerances for this circuit because of the inherent high gain of a cascaded transistor circuit and the nonlinear feedback action. This statement holds true for transient considerations since the storage time of the transistors is not important.

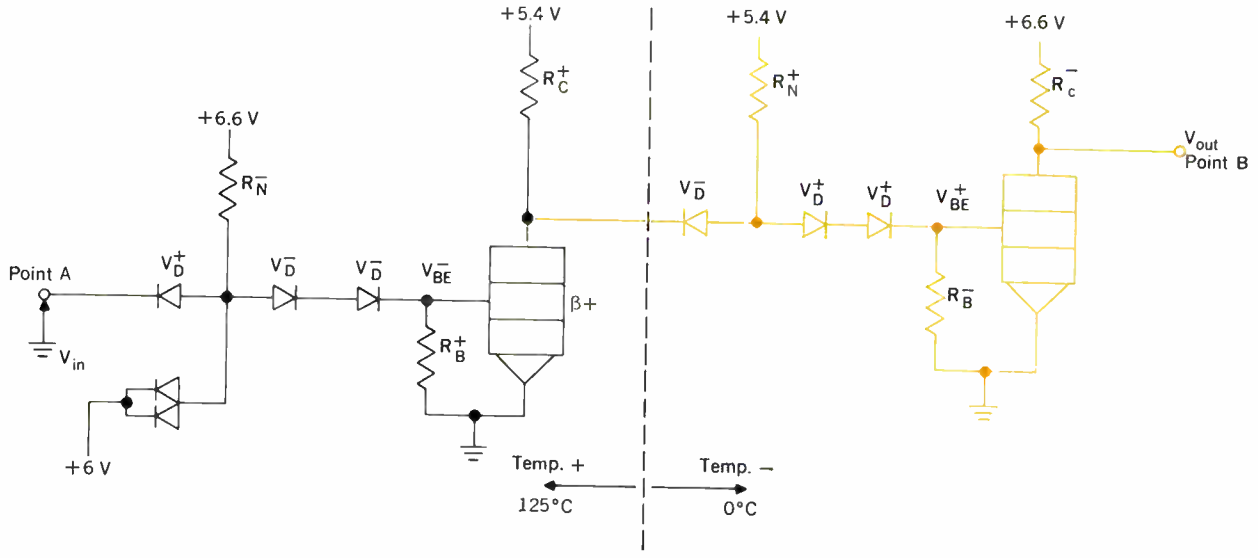
The choice of resistor and power-supply values for this circuit is based on considerations of required circuit delay times; therefore, at this point, it would be meaningless to go through a dc design procedure. Rather, some arbitrary resistor values will be chosen to illustrate two factors that must always be considered: total circuit power dissipation and maximum transistor junction temperature. The chosen resistor values and all design factors correspond roughly to those used in Example 2. The only exception is that now a fan-out of ten will be considered, as shown in Fig. 11.

The power calculations will be made using tolerances which allow transistor T_2 a maximum power dissipation corresponding closely to the condition for maximum total circuit dissipation. Let us begin by noting that V_{OUT} can readily be obtained by an iterative process. For the first guess, observe that

$$V_{OUT} = V_{BE}{}^{+}T_2 + V_{BE}{}^{+}T_1 - V_F{}^{-}D_3$$

This allows a first guess at the value of V_{OUT} by quickly

Fig. 9. Circuit for calculating noise-rejection level.



estimating various currents. Thus

$$I_{E_2} \approx 60 \text{ mA}, I_{E_1} \approx 4 \text{ mA } (\beta = 20)$$

$I_{D3} \approx 1.0$ mA and $V_{OUT} = 0.94 + 0.79 - 0.70 = 1.03$ V

Assuming $V_{OUT} = 1.03$ V, then

$$I_{c_2} = \frac{6.6 - 1.03}{0.95} + (10)\frac{(6.6 - 1.03 - 0.70)}{0.95 + 0.02} + I_{D_3}$$

$$I_{c_2} = 5.86 + 50.2 + I_{D_3}$$

Then neglect I_{D_3} in comparison with the total current

$$I_{E_2} = 56.06\frac{(\beta + 1)}{(\beta)} = 58.9 \text{ mA}$$

$$V_{BE}T_2 = 0.94 \text{ V}$$

$$I_{E_1} = I_{B_2} + \frac{0.94}{0.95} = 2.8 + 0.99 = 3.79 \text{ mA}$$

$$V_{BE}T_1 = 0.79 \text{ V}$$

$$VA = V_{BE}T_2 + V_{BE}T_1 = 1.73 \text{ V}$$

Now calculate the currents at node A based on this voltage and observe if the summation of currents is zero.

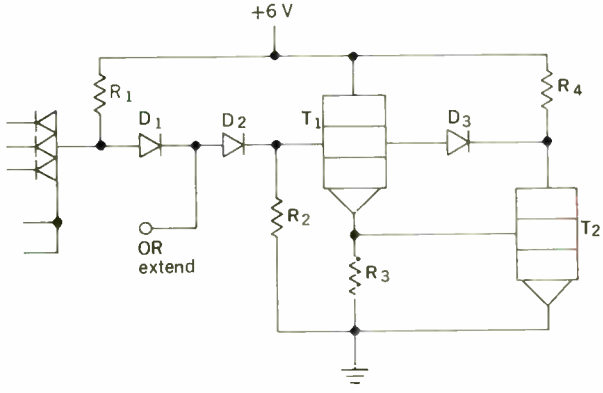


Fig. 10. Nonsaturating DTL circuit.

When the diode forward voltage is 0.67,

$$I_{IN} = \frac{6.6 - 2(0.67) - 1.73}{0.95 + 2(0.03)} = \frac{3.53}{1.01} = 3.5 \text{ mA}$$

$$I_1 = 0.91 \text{ mA}$$

$$I_{B_1} = 0.2 \text{ mA}$$

$I_{D_3} = 1.0$ mA (see minimum diode curve at $V_F = 1.73 - 1.03 = 0.70$ V)

If V_{OUT} was guessed correctly, then

$$I_{IN} - I_1 - I_{B_1} = I_{D_3}$$

Inserting the values just calculated yields $3.5 - 0.91 - 0.2 = 2.39$ mA $\neq 1.0$ mA. This is greater than the value of 1.0 mA calculated for I_{D_3}, thus $V_{OUT} < 1.03$ V.

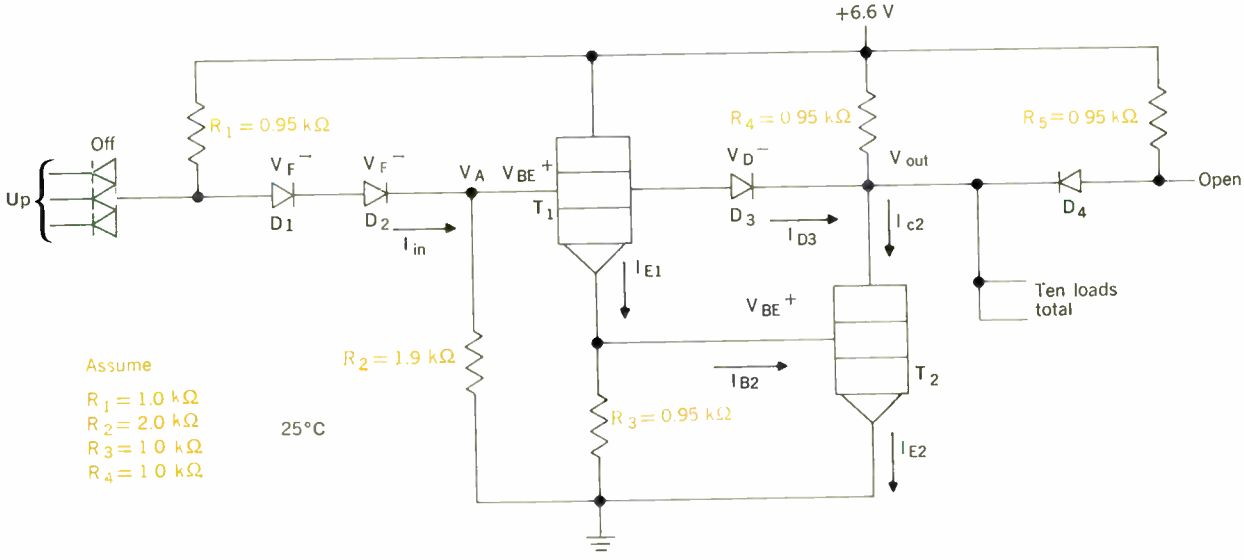
Let us next assume a decrease in V_{OUT} of 0.04 V since our first guess for V_{FD_3} was off by that amount. Thus, $V_{OUT} = 0.99$ V is assumed and another iteration begins. The second iteration shows that $V_{OUT} > 0.99$ V.

The third iteration will yield $V_{OUT} = 1.0$ V as the final value. Using this value and the calculated currents, we obtain the power dissipation of the circuit. By components, this results in

R_1, D_1, D_2	=	16.95 mW
D_3	=	1.78
R_2	=	1.60
R_3	=	0.95
R_4	=	33.20
T_1	=	23.03
T_2	=	62.25
Total	=	139.76 mW (on case)

The total power dissipation is quite high, which could be a prohibitive factor when many circuits are packaged in a small space. Thus, because of high total power dissipation, this circuit will require special consideration for maintaining a maximum ambient temperature. The junction temperature of T_2 and T_1 must also be considered closely since their power dissipation is much higher than that of transistors in any normal saturating circuit. A high thermal resistance would result in a

Fig. 11. Circuit for power dissipation calculations.



junction temperature that is several degrees above the ambient temperatures.

Computer-programmed design and analysis

Thus far, various factors in worst-case circuit designs have been discussed, and some pertinent examples studied. Unless the designer initially has a thorough knowledge of the circuit, the first phases of any circuit design should probably be done manually because of the insight that can be gained by actually plugging through a number of iterations on the design. It is also helpful to breadboard the initial design to gain understanding of the circuit. These initial steps should be followed by the writing of a computer program to handle future iterations, especially if the circuit design is critical, circuit usage is expected to be widespread, and various design factors are subject to change.

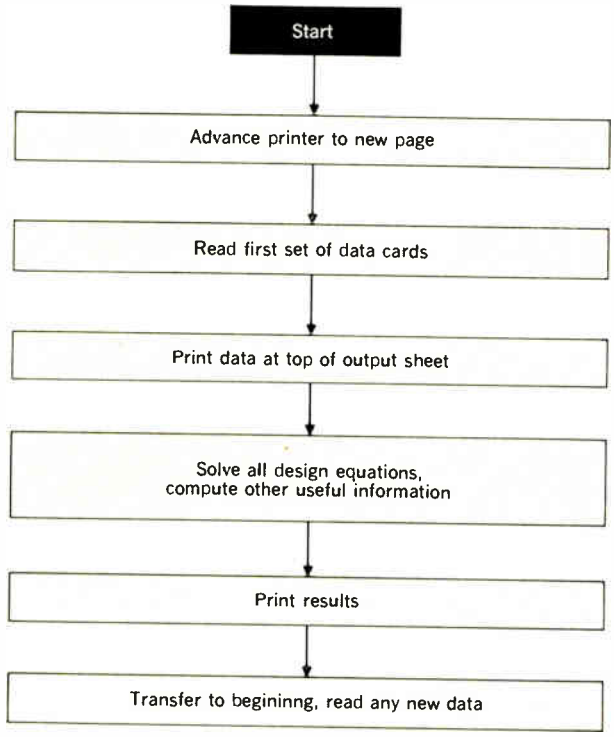
The design program should be tackled in much the same manner as the manual design. It is best not to try to solve too many simultaneous equations, since blindly writing and solving simultaneous equations can lead to very complicated problems. The net result may obscure some very important intermediate results. Note that this example assumes that the program is to be written in IBM FORTRAN. This assumption is only for illustration purposes and anyone with knowledge of another programming system will follow the program with no difficulty. The use of FORTRAN or a similar compiler greatly simplifies the programming and encourages a more sophisticated and comprehensive program. The suggested steps for the design program are:

1. Make a list of all design factors that can conceivably be variables and assign a symbolic name to each. Within FORTRAN limitations, the symbols should be assigned in a meaningful manner, sticking to conventional designations if possible. These variables will be entered by data cards, which facilitates changing the data without changing the program.

2. Write and solve all necessary design equations in terms of the variable factors. After the resistor values have been found, write and solve any and all equations for such things as power dissipation, drive currents, load currents, and maximum fan-out. It is relatively easy to analyze everything required. With a hand-wrought design and possibly a breadboard circuit one should possess the necessary insight to guarantee that the equations will cover all future contingencies.

3. The last step is to write the FORTRAN program. A flow chart of the FORTRAN program is shown in Fig. 12. There should be a minimum amount of time required for the actual FORTRAN coding and debugging; this is the advantage to be gained by using such a formula translation language. Once the carefully planned program is written and debugged, one need only punch new data cards to evaluate completely a change in any or all of the design factors. The ease with which the changes can be evaluated allows for quick evaluation of trade-offs between different factors.

Fig. 12. Flow chart of FORTRAN program.



A reference was previously made to a piecewise linear approximation for nonlinear elements such as transistor base-emitter voltages and diodes—a particularly effective method, if used carefully. The suggested technique involves a prior knowledge of the current that will operate the element, which means each element in the circuit may require individual attention. Knowledge of the approximate current range allows the nonlinear element to be replaced with a single resistor and power supply. This technique is usually quite accurate for relatively large excursions about the approximated operating point and simplifies the writing and solving of most design equations.

Philosophy of worst-case design

As stated before, a circuit that has been worst-case designed is guaranteed to meet all circuit requirements, even under extreme worst-case conditions. In many applications, particularly military, this is not only a desirable, but a necessary, goal. Failure of one circuit may cause a complete system to malfunction. In some cases, redundant circuit elements, or even redundant circuits, are used to guard against the possibility of catastrophic failures. Thus, reliability needs may completely override any cost considerations.

In commercial systems, the need for guaranteed circuit performance can be stated in terms of dollars. A circuit failure leads to the following:

1. Down time. No matter how small or large, a system that is not working is unprofitable for the customer.

2. Repair costs. The more complex systems may require hours or even days to locate the circuit that is failing. This is particularly true in the case of a marginally failing or intermittently failing circuit.

3. Replacement cost. The replacement cost may be quite significant. In many cases, the defective circuit may be packaged so that several other circuits will have to be replaced at the same time. This problem increases rapidly as density and functional sophistication increase. The "throwaway costs" are quite often the deciding factor in choosing a packaging system.

4. Loss in sales. A system with a poor performance record is not going to sell. Furthermore, the poor performance of just one system type may lead to a bad reputation for the entire product line of the company.

Thus, although the basic reason for using a worst-case design philosophy is one of economics, it is easy to show that at times this philosophy is not economically justified. In other words, the component specifications required to satisfy the circuit requirements may increase the price of the circuits to the point where the down time and repair costs become secondary. Moreover, a system's price may increase to the point where it is not competitive. Anyone who has used a worst-case design philosophy extensively is well aware of the fact that it can be wasteful. It is usually difficult to make a circuit fail under laboratory conditions. Mistakes such as having a 10-kΩ resistor instead of a 1-kΩ resistor or having a power supply at twice or one half its nominal value quite often go undetected by casual observation. Another typical example is the search for minimum β transistors needed to laboratory check a circuit that is dependent on β. After convincing the user that a higher β specification would be impossible, the transistor manufacturer quite often finds it difficult to supply minimum β units for test purposes. A more convincing demonstration can be obtained by calculating circuit performance under nominal or best-case conditions. These calculations usually indicate that worst-case design is indeed wasteful.

Thus, the need arises for a design method that offers a compromise between the requirements for unconditionally guaranteed circuit performance and unrealistic, costly component specifications. The obvious answer is a statistical approach that can take advantage of the fact that all components may not simultaneously be at worst-case end-of-life limits. For this approach to be most successful, the circuit designer must know the distributions of tolerances and parameters. Furthermore, he must have a guarantee that these distributions will not change with time. This can be a major problem since most manufacturers are reluctant to write distribution specifications, especially in the development stages. But even without guaranteed distributions, there may be much to be gained by statistical analysis.

A statistical approach to circuit evaluation need not be hopelessly complicated to be useful. The most popular and widely used approach is the "Monte Carlo" method, in which a random number generator selects values for all components according to specified distributions. The circuit is then evaluated and its performance calculated. This process is repeated many times (5000 to 10 000) until a distribution of various performance ratings, such as fan-out and fan-in, is obtained. These distributions will often allow for definite improvement in performance ratings, as compared with the worst-case calculations, obtained directly as a function of what exposure to circuit failures can be tolerated.

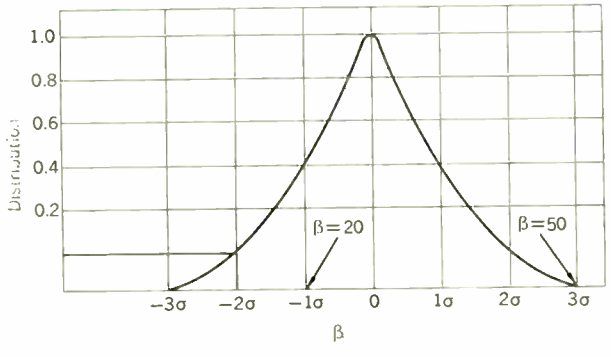
Generally, the more components per circuit the more there is to be gained from statistical evaluation. Significant improvement in performance may often be gained by exposure to 1 to 10 failures per 10 000 circuits.

Previously, it was pointed out that the distributions of tolerances and parameters must be specified to obtain the greatest benefits from statistical evaluation but that benefits may often be obtained without guaranteed distributions. The approach to use is one of statistical evaluation using worst-case distributions. As an example, suppose there is a minimum β specification of 20 with no maximum limit specified. By measuring samples of the transistors, it may become apparent that only a few transistors have a β of less than 30, with typical values greater than 50. A worst-case distribution based on this information may then have a maximum of 50, a minimum of 20, and an average value skewed toward the low end. This example is diagrammed in Fig. 13 using an assumed normal distribution of the product.

The normal distribution curve is suggested since almost all processes will naturally yield a normal distribution. The unanswered question is, "What does testing do to the distribution?" For this example it is assumed that the minimum β limit is at the -1.0 sigma point, which corresponds to a loss of slightly more than 15 per cent due to β's < 20. From the known facts about the β distribution, the assumed worst-case distribution certainly cannot be described as optimistic. But if one uses this distribution, instead of the minimum of 20, a significant improvement in circuit performance ratings will undoubtedly result. Again, the calculation of nominal and best-case performance ratings dramatically illustrates what is to be gained by a statistical evaluation.

The idea of worst-case distributions should then be extended to other design factors, such as resistor and semiconductor device tolerances. Generally, the statistical approach should not be used for such design factors as power supply tolerances and temperature since these factors may be at worst-case limits for long periods of time and thus must be handled in a true worst-case fashion.

Fig. 13. Example of worst-case distribution for β.



Summary

Worst-case circuit design is not a cookbook approach that allows good circuit designs to be arrived at without thinking. Neither is it an approach that requires an inordinate degree of sophistication and engineering skill. The success of the worst-case design approach depends entirely on a thorough knowledge and understanding of all circuit requirements and design factors that are involved.

REFERENCE

1. "Integrated Circuit Is Modified DTL," *Electronics*, Feb. 15, 1963, p. 106.

The optical properties of metals

Extensive studies are being made of the optical properties of various metals, such as aluminum, silver, copper, and nickel, and their interpretation in terms of band structure and plasma oscillations

Henry Ehrenreich *Harvard University*

The distinction between metals and insulators is made on the basis of dc electrical properties. Upon the application of a steady electric field, a metal conducts but an insulator does not. This distinction becomes increasingly vague when the applied field is frequency dependent, and particularly when it falls into the optical range of the visible and ultraviolet. In this range the electrical, or more appropriately, the optical properties of metals and insulators have a great deal in common. Thus, at the outset of this discussion, which ostensibly deals only with metals, we must recognize the necessity of understanding something about both metals and insulators.

The earliest realistic model for a metal was proposed by Drude at the turn of the century. He supposed the electrons to form a gas of free particles which could respond, for example, to dc fields by producing current. He completely neglected the ionic cores present in the solid, and knew of no statistics on the subject other than those of Boltzmann. It was later discovered that, through Sommerfeld's substitution of Fermi statistics for Boltzmann statistics, Drude's picture could be transformed into a very realistic and useful model for a metal. However, even in its original form, Drude's model is already adequate for the discussion of many aspects of the optical properties of metals.

An almost equally simple picture of an insulator was proposed by Lorentz. He suggested that in this case the electrons be envisioned as tied to certain equilibrium positions by springs. Since the electrons were not free to wander away, the crystal could have no dc conductivity. On the other hand, polarization effects in the presence of an ac field were quite possible.

In the light of the sophisticated modern developments of solid-state physics, it is perhaps remarkable that such simple ideas still have relevance in the interpretation of the optical properties of metals. It is, however, true qualitatively that if one can understand how electrons described by these models respond to light, or equivalently to an ac electric field, then an elementary synthesis and generalization of these ideas will come close to a true description of the state of affairs in a real solid.

Classical oscillators

The response of one of the electrons in a Lorentz insulator to a uniform oscillating electric field, $E(t) = E(\omega)e^{-i\omega t}$ having a single Fourier component corresponding to frequency ω, is described by the elementary equation of motion

$$\ddot{x} + \gamma\dot{x} + \omega_s^2 x = \frac{eE(\omega)}{m} e^{-i\omega t} \qquad (1)$$

Here $\ddot{x} = d^2x/dt^2$ is the acceleration term, $\gamma\dot{x}$ is a damping term proportional to the velocity and characterized by a strength γ, and $\omega_s^2 x$ is associated with Hooke's law describing the restoring force of a spring whose natural frequency is ω_s. The right-hand side of the equation represents the driving term due to the electric field. Since the equation is linear in the displacement x, we may assume that $x(t)$, and, in fact, all electromagnetic field quantities, have the same time dependence as $E(t)$;

that is, $x(t) = x(\omega)e^{-i\omega t}$. We may note in passing that the addition of a term having the form αx^2 to the preceding equation would give rise to typical nonlinear electromagnetic effects, which have been recently studied by means of very intense laser beams. The substitution $x(t) = x(\omega)e^{-i\omega t}$ transforms (1) into a linear algebraic equation for $x(\omega)$ that can be readily solved.

The vibrations of all of the electrons in a unit volume of the insulator result in an induced polarization given by

$$P(t) = P(\omega)e^{-i\omega t}$$

whose amplitude, $P(\omega) = enx(\omega)$, is proportional to the displacement $x(\omega)$ and the concentration of particles n. We therefore find

$$P(\omega) = enx(\omega) = \frac{ne^2E(\omega)/m}{\omega^2 + i\omega\gamma - \omega_s^2} \qquad (2)$$

The frequency-dependent dielectric constant $\epsilon(\omega)$ describing the response of the system to an external field $E(\omega)e^{-i\omega t}$ having frequency ω may be defined in the standard way (cgs units):

$$P(\omega) = \frac{1}{4\pi}[\epsilon(\omega) - 1]E(\omega) \qquad (3)$$

Combining (2) and (3) yields the dielectric constant for the Lorentz insulator

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma - \omega_s^2} \qquad (4)$$

where $\omega_p = (4\pi ne^2/m)^{1/2}$ is the plasma frequency associated with the electrons. We note that this frequency, whose physical meaning will be made clearer later on, appears quite naturally in this formulation. Also to be emphasized are the facts that $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$ is complex and that the real and imaginary parts, ϵ_1 and ϵ_2, are simply related to the index of refraction n and the extinction coefficient k, which are more commonly used to describe the optical properties of systems. As shown in standard texts,

$$\epsilon_1 = n^2 - k^2 \qquad \text{and} \qquad \epsilon_2 = 2nk$$

Thus (ϵ_1, ϵ_2) and (n, k) contain the same physical information. For purposes of theoretical descriptions, the dielectric constant formulation turns out to be more convenient.

The behavior of the function described in (4) is shown in Fig. 1. For an insulator, in which the spring frequency is finite, the dielectric constant is that of an oscillator. On the other hand, for a metal we can recover the Drude model simply by equating the spring frequency to zero, meaning that the electrons are free to wander through the crystal. The characteristic behavior for the real and imaginary parts of the dielectric constant is now as shown in Fig. 1(B); ϵ_2 and ϵ_1 become positively and negatively large, respectively, with diminishing frequency. Thus we may alternatively characterize the electrons in a Drude metal as oscillators of zero frequency.

Analysis for a real metal

Let us now turn to a more explicit examination of the relationship of the preceding results to the dielectric constant for a real metal. Figure 2 is a greatly simplified diagram of a portion of the band structure along two directions k_1 and k_2 in the Brillouin zone. In a solid, energy levels are generally described in terms of a wave number k and a band index. The wave numbers corresponding to independent solutions of the wave equation are all contained in the first Brillouin zone, which is the fundamental unit cell in reciprocal space and results from the periodicity of the lattice in ordinary space. The Brillouin zone boundaries are indicated by the vertical dashed lines on the diagram. Although Fig. 2 represents the energy levels for a single electron only, one can, apparently to good approximation, regard the electrons in a real solid as independent and apply the diagram to such systems. In this independent particle model the electrons only know of one another's existence through the exclusion principle, which provides that no more than two electrons, one of each spin, can occupy a given state. This means that at low temperatures all levels below the Fermi level, indicated on the diagram by the dashed horizontal line, are filled and all others are empty.

Optical excitation processes

We may now describe the three relevant optical excitation processes in metals in terms of the diagram. First, let us consider the so-called intraband transitions, which are analogous to the optical excitations in a

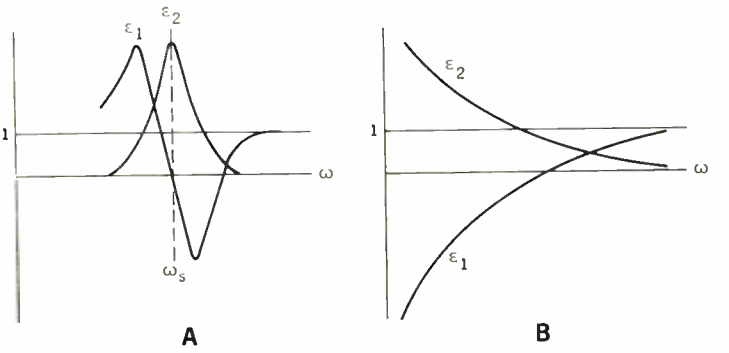
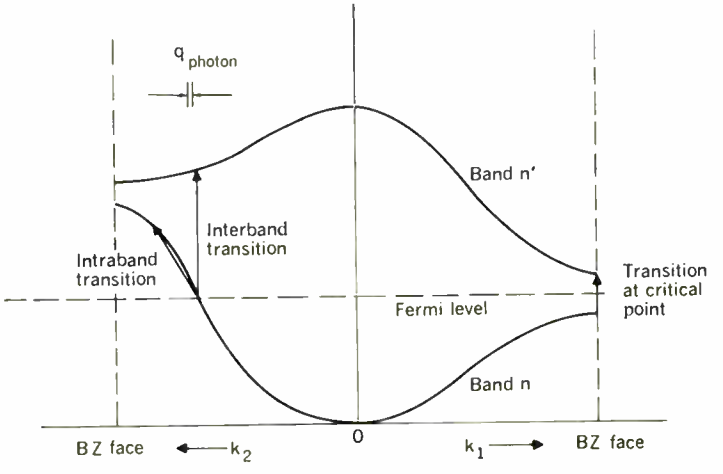


Fig. 1. Behavior of function described by Eq. (4). A—Insulator ($\omega_s \neq 0$). B—Metal ($\omega_s = 0$).

Fig. 2. Schematic band diagram for metal.



free-electron metal. As indicated, they involve the excitation of an electron by a photon within the same band. Because of the exclusion principle, the electron must go from an initial filled state to a final empty state. These transitions therefore can only occur in metals, in which some of the bands are partly filled, but not in insulators, in which all bands are either completely full or completely empty. By contrast, the second kind of excitation, the interband transition, which involves two bands, is common to both insulators and metals. This process has its analog in the optical excitation of a Lorentz insulator and, in fact, reduces almost precisely to that for the very artificial case of a solid consisting of a widely separated array of two quantum level atoms whose wave functions do not overlap appreciably. In the absence of any other interactions involving the electrons, the interband transitions can be shown to be essentially vertical in k space, as a result of the fact that the total crystal momentum of electron plus photon must be conserved during the transition. For wavelengths that are in the visible and ultraviolet regions, the momentum of the light wave q_{photon} (Fig. 2) is very much smaller than the diameter of a Brillouin zone, and for this reason the k of the electron is essentially unchanged.

Comparison of the result of a proper quantum mechanical derivation of the dielectric constant of a solid with the elementary results we have obtained here shows that it is legitimate to visualize this quantity as a sum of contributions from both intraband and interband transitions. These transitions can be described essentially as in a Drude free-electron metal plus a contribution of oscillators at each k connecting vertically filled and unfilled states.

Since they are possible everywhere in the Brillouin zone and for all energies beyond a certain threshold, interband transitions might be expected to present relatively structureless contributions to the dielectric constant. The observation first made in connection with semiconductors by J. C. Phillips—that very large contributions arise from certain small regions in k space, which result in pronounced structure—was very important, since it led to a new branch of an old discipline, which one might call "band-structure spectroscopy." These regions, which contain a so-called "critical point," are characterized by very rapidly varying, often large densities of states for optical transitions. For the most part these occur at points of high symmetry, points at which the band structure of a given material is most likely to be known. Thus, given a rough idea of the band structure, it is possible to use optical techniques to obtain some important band gaps, thereby providing a basis for better calculations.

The third kind of elementary excitation in a solid is the so-called plasma or collective oscillation, which results from the high density of electrons that are present in a metal and the fact that they can act cooperatively due to the Coulomb interaction among them. This cooperative motion may be regarded as a fluidlike vibration of the electrons as a whole through the solid, which produces fluctuations in the electron density. Under certain conditions plasma oscillations can represent normal modes of the entire system. This means

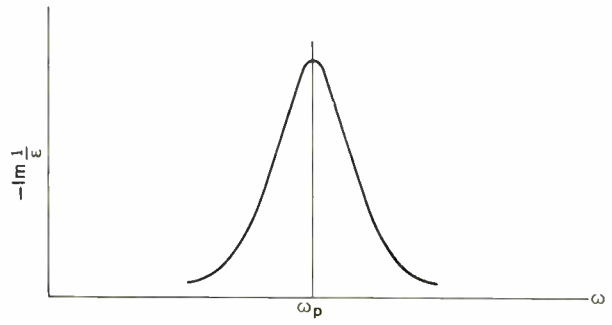
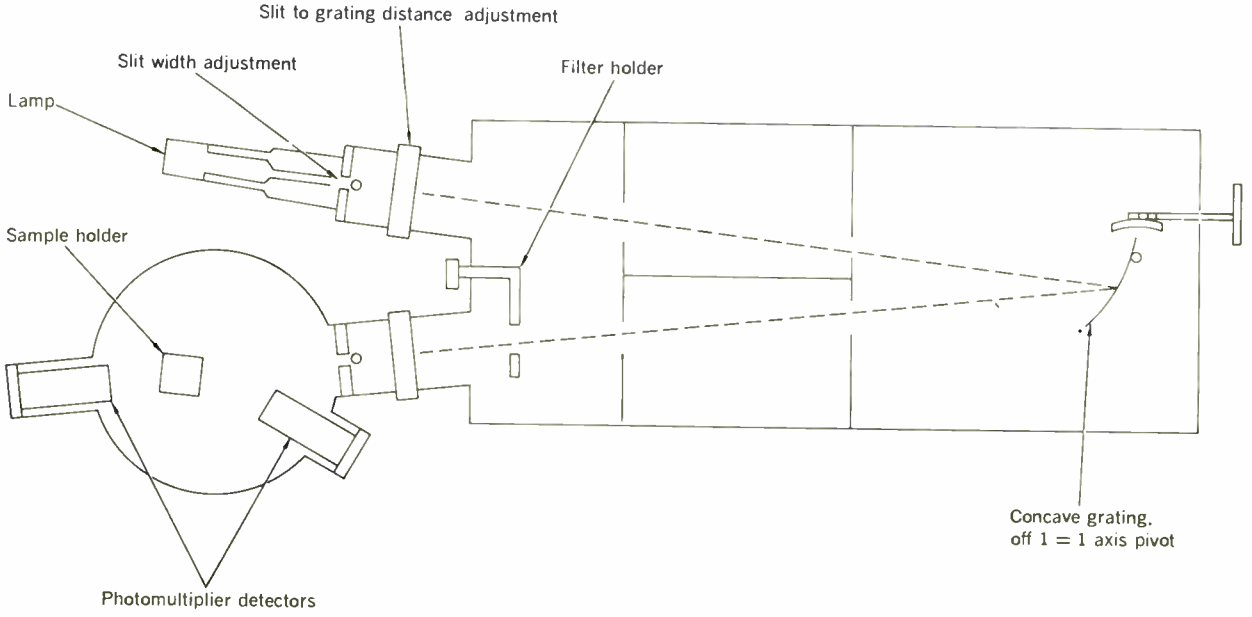


Fig. 3. Energy loss as a function of plasma frequency.

Fig. 4. Schematic drawing of experimental apparatus showing vacuum grating monochromator and sample chamber. (From Philipp and Ehrenreich[1])



that once such oscillations are excited, they do not decay in time. Since they maintain themselves, we would expect at the plasma frequency ω_p to find the presence of an internal electric field $E(\omega_p)$ in the solid, due to the oscillations in the electron density, without the presence of an external field $D(\omega_p)$. Since

$$D(\omega_p) = \epsilon(\omega_p)E(\omega_p) = 0$$

and $E(\omega_p) \neq 0$, the condition for plasma oscillations is therefore

$$\epsilon(\omega_p) = \epsilon_1(\omega_p) + i\epsilon_2(\omega_p) = 0 \qquad (5)$$

Thus, strictly speaking, plasma oscillations can exist as normal modes of the system only if both the real and imaginary parts of the dielectric constant vanish at ω_p. In practice, however, the plasma frequency will be well defined if $\epsilon_2(\omega_p) < 1$ when $\epsilon_1(\omega_p) = 0$. Since ϵ_2 represents the damping of the plasma resonance, the condition $\epsilon_2(\omega_p) < 1$ implies that the damping should be small.

The most familiar method of determining the plasma frequency has utilized the measurement of the characteristic energy loss of fast electrons traversing a thin film; see Fig. 3. The energy loss of such an electron is proportional to Im $1/\epsilon$, a function that is seen to be very sharply peaked about the plasma frequency when the conditions for plasma oscillations are fulfilled. Since the optical properties determine the real and imaginary parts of the dielectric constant, it is clear that from them one can also calculate the energy loss function.

Analysis of data

With these basic ideas in mind, let us now turn to the consideration of actual data and, by way of example, the kind of physical information that can be obtained from them. In this generally interpretative presentation, we shall not have the opportunity for an extensive discussion of the experimental technique, which is due largely to the work of Philipp and Taft. This technique is based on the recognition that the tremendous range of absorption coefficients present in a metal make their direct measurement impractical. Accordingly, the reflectance is measured at normal incidence over an extended energy range by means of the instrument shown schematically in Fig. 4. The essential ingredients are a source, a monochromatizing grating, the sample, and a photocell for measuring the intensity of the reflected light. By pulling the sample out of the way one can also directly measure the intensity of the incident beam. The reason for complications in this kind of setup is that air becomes opaque to light in the ultraviolet region. As a result, the measurements must be performed in vacuum for much of the energy range of interest.

As we already have seen, a complete description of the optical properties at a given frequency requires two optical constants, whereas the reflectance measurements described yield only one. The way out of this dilemma is provided by the Kramers–Kronig relations. These are very general equations that connect the real and imaginary parts of the logarithm of the complex reflectance

$$\ln r = \ln|r| + i\theta$$

and also the real and imaginary parts of the dielectric constant, as well as its inverse. The equation for the phase θ involves an integral that depends on the reflectance:

$$\theta(\omega_0) = \frac{1}{2\pi}\int_0^\infty \frac{d\ln r}{d\omega} \ln\frac{\omega+\omega_0}{\omega-\omega_0}\, d\omega \qquad (6)$$

This integral can be evaluated numerically to a good approximation if the reflectance is known over a sufficiently wide energy range.

There are similar relations between the pairs of quantities n, k and ϵ_1, ϵ_2, any of which will provide a complete description of the optical properties of solids.

In order to describe somewhat more specifically the

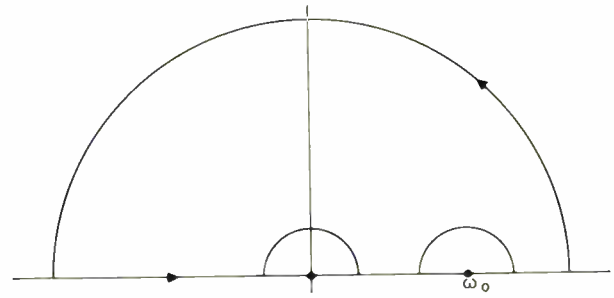
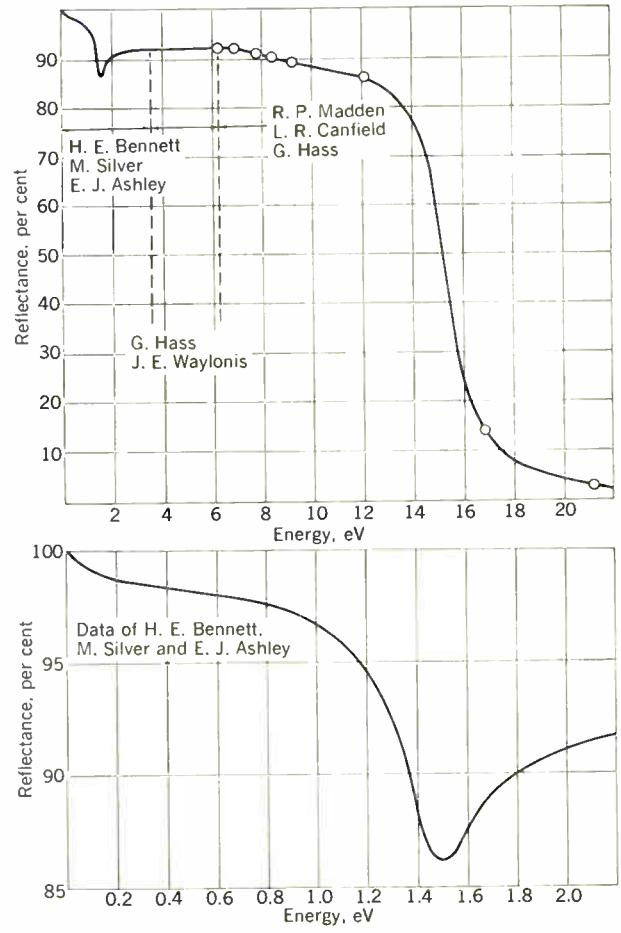


Fig. 5. Contour of integration to be used in connection with Eq. (7).

Fig. 6. Spectral dependence of the reflectance of aluminum. (From Ehrenreich, Philipp, and Segall[3])



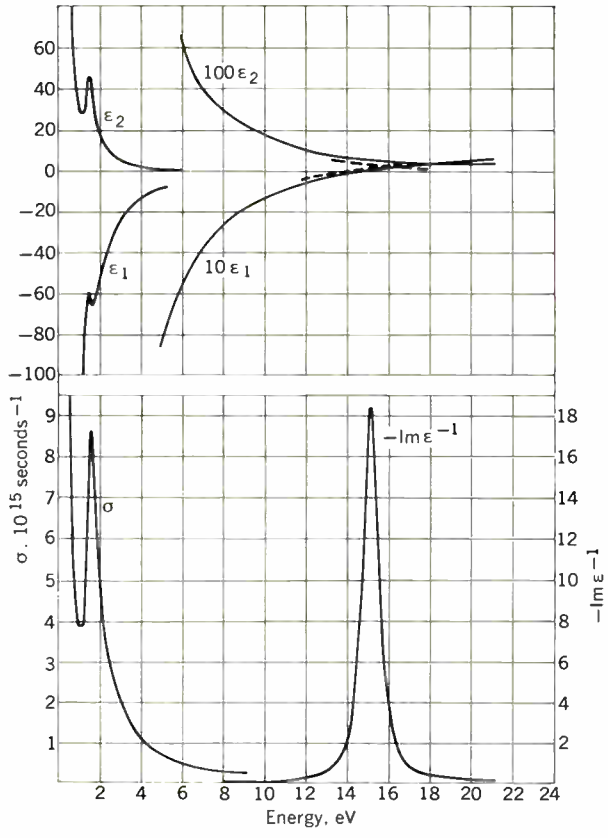


Fig. 7. Spectral dependence of the real and imaginary parts of the dielectric constant, the conductivity $\sigma = \omega\epsilon_2/4$, and the energy loss function $-\mathrm{Im}\ \epsilon^{-1}$ for aluminum obtained by Kramers-Kronig analysis of the curve of Fig. 6. (From Ref. 2)

nature of the Kramers–Kronig relationship, we might consider how it is obtained for the case of the dielectric constant appropriate to a single Lorentz oscillator. It depends essentially on the fact that $\epsilon(\omega) - 1$ is analytic in the upper half plane because the damping constant γ is such that the amplitude of vibration becomes smaller and not bigger with increasing time. Thus, the sign of γ assures that all the poles of the function being integrated occur in the lower half plane. This is simply an expression of the causal nature of the system. If we have established the analyticity of a function in the upper half plane, we can use Cauchy's theorem to obtain the following equation:

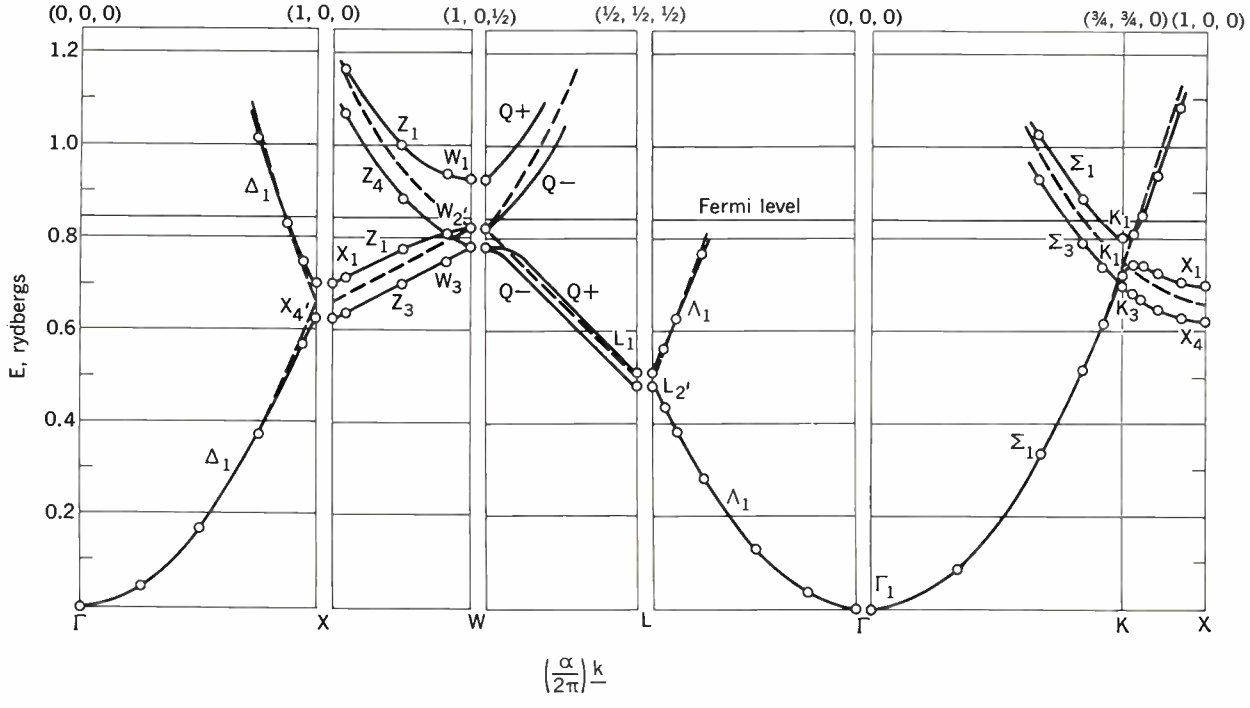
$$P\int_0^\infty \frac{\epsilon(\omega) - 1}{\omega - \omega_0} = i\pi[\epsilon(\omega_0) - 1] \qquad \text{for finite } \omega_s \qquad (7)$$

A contour of this integration process is shown in Fig. 5. Because of the presence of the factor i, Eq. (7) reduces to a relation between the real and imaginary parts of ϵ.

Aluminum

Figure 6 shows the reflectance of aluminum, which is remarkable for its especially simple optical properties. The data shown here represent painstaking work of the several investigators listed. The experimental difficulties encountered were attributable to the fact that aluminum is very reactive with oxygen and forms an oxide layer upon exposure to even traces of air. The data shown here can be considered representative of pure aluminum. The relatively simple curve illustrates the most important features present in the reflectance of a metal. For an ideal Drude free-electron metal, the reflectance stays at 100 per cent as long as ϵ_1 is negative. When ϵ_1 passes through zero at the plasma frequency, the reflectance falls off very rapidly. We note that aluminum appears to conform qualitatively with this description except for the sharp dip near 1 eV, which is due to

Fig. 8. Energy bands for the various symmetry axes within the Brillouin zone and on the zone surface. Solid curves represent calculated energies; dashed curves indicate the free-electron energy bands. (From Segall[3])



interband transitions and therefore represents a distinct departure from this model. The simplicity of the optical properties of aluminum results from the fact, to be discussed in more detail subsequently, that the interband transitions are confined to a very narrow energy range. In addition, it can be seen that the reflectance of really pure aluminum surfaces remains about 90 per cent up to photon energies of 10 eV.

The results of a Kramers–Kronig analysis of the preceding data are shown in Fig. 7. First, note that the real and imaginary parts of the dielectric constant near zero frequency have the behavior characteristic of a Drude free-electron metal. Around 1 eV we can see the effect of additive superposed structures, an effect associated with interband transitions. As expected, the real part of the dielectric constant remains negative until 15 eV, when the reflectance begins to decrease. The imaginary part is seen to be very small when ϵ_1 passes through zero. These are the hallmarks of a plasma resonance, and the energy loss function, which is shown at the bottom of Fig. 7, exhibits a very sharp peak in this region. The plasma frequency is given by the simple classical expression discussed previously for an electron concentration of three per atom and a mass equal to that of the free electron. In this frequency range the electrons are therefore behaving as though they were really free particles and not imbedded in a solid. The reason for this extremely simple behavior lies in the exhaustion of the *f*-sum rule in this frequency range, which means that all important interband transitions involving the valence band as an initial state here take place at lower energies. In somewhat more elementary terms, we could say, using our model of the Lorentz insulator, that the electrons no longer feel the springs by which they are attached to the lattice sites since in this frequency range the driving frequency is much greater than the natural frequency of the spring.

Since the reciprocal dielectric constant also satisfies the Kramers–Kronig relations, it is possible to perform a similar analysis on the energy loss function as obtained from experimental, characteristic energy loss data. The results of such an analysis, performed by LaVilla and Mendlowitz at the National Bureau of Standards, are indicated by the dashed lines in Fig. 7 and agree very well with the results obtained from the optical data.

Since intraband effects set in at a finite frequency and since the intraband contribution to ϵ_2 is confined to the low-frequency range, the two contributions can be separated. One simply fits the Drude expression for the dielectric constant—whose form, incidentally, is correct even for an interacting electron gas—to the data at low frequency, considering the effective mass and relaxation time as adjustable parameters. In this manner we can isolate the interband contribution and attempt to calculate it directly from the band structure.

Figure 8 shows the results of Segall's calculations along various directions of the Brillouin zone for the face-centered cubic lattice. The details here are entirely unimportant for our purposes. We wish to note only that the band structure is very nearly free-electron-like, as seen from the very small band gaps that are in evidence. Indeed, if the band gaps vanished entirely, the band structure shown here would be simply that of a free-electron gas. There are only two small regions of the

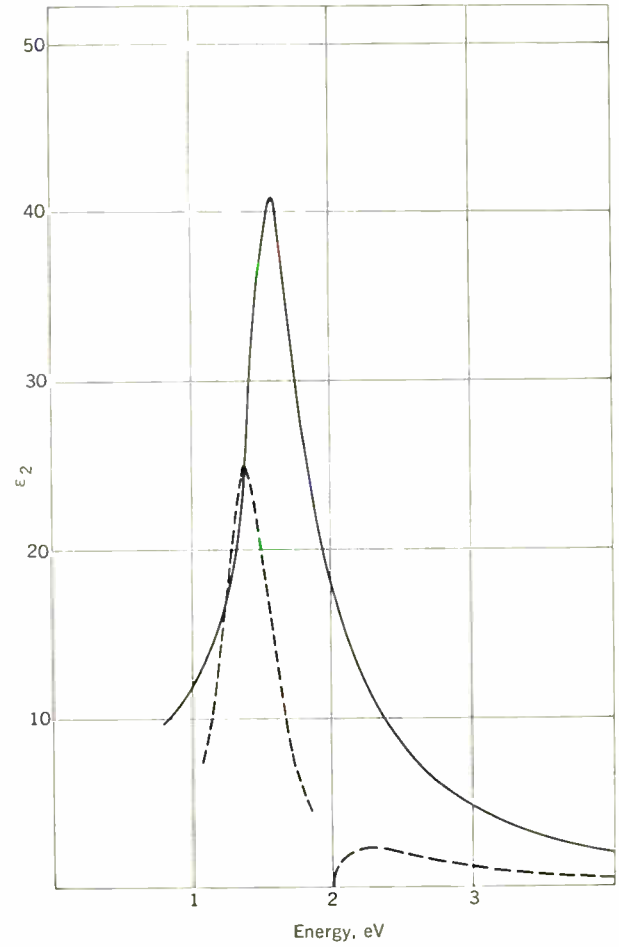


Fig. 9. Calculated contributions to the imaginary part of the dielectric constant from the interband transitions around W and Σ. The dashed peak is the sum of the contributions from the transitions between the W_2' and W_1 bands and between the Σ_3 and Σ_1 bands. The dashed curve starting at 2.0 eV is associated with the transition from the W_3 bands to the W_1 band. For comparison purposes, the experimentally determined interband part of $\epsilon_2(\omega)$ is given by the solid curve. (From Ref. 2)

Brillouin zone corresponding to critical points that contribute importantly to interband transitions, namely, *W* and the axis Σ. From the known band structure and wave functions, we can now calculate the expected interband contribution to ϵ_2 without the use of adjustable parameters. In Fig. 9 the results of this calculation are compared with the experimental interband contribution to ϵ_2. Both the shape and energetic position of the two curves are seen to agree quite well. The magnitude, however, of the calculated curve is seen to be too small. The discrepancy can be shown to be caused by many electron effects, which of course are neglected in a simple calculation based on the band approximation. However, the extent of the agreement between the experimental and theoretical results shown here indicates that these effects change the curvatures of bands but apparently do not contribute greatly to their relative separations.

Having obtained the optical constants out to 25 eV, we can, with the help of absorption data in the soft

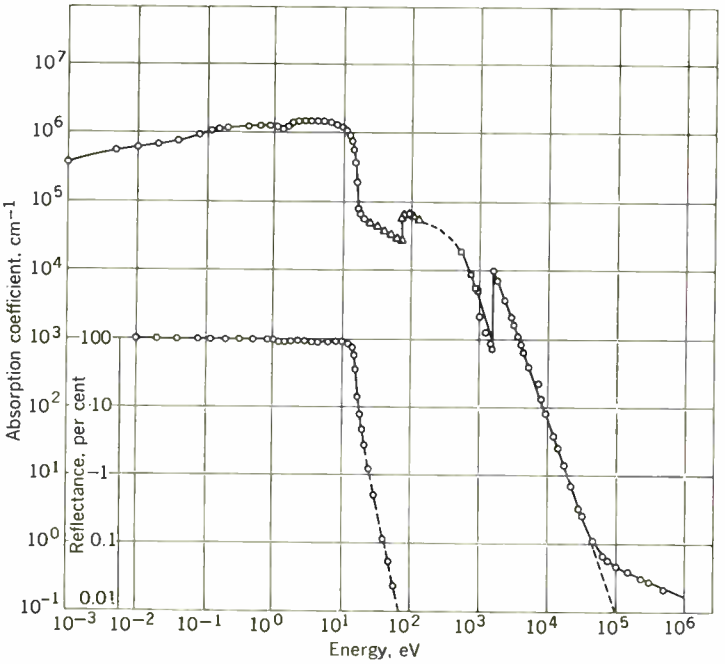


Fig. 10. Spectral dependence of the absorption coefficient for aluminum. The upper curve is a composite of data from several sources. (From Philipp and Ehrenreich[1])

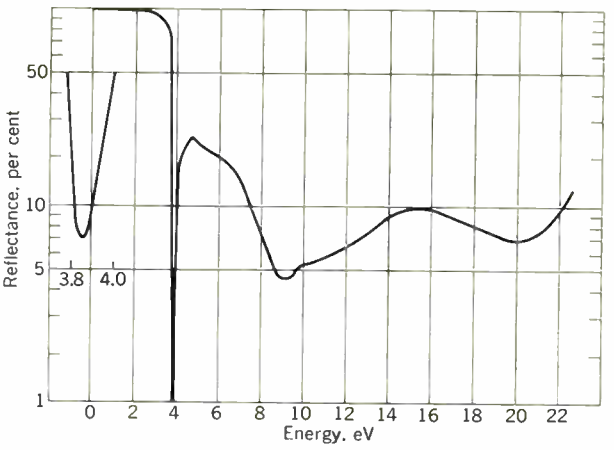


Fig. 11. Spectral dependence of the reflectance of silver. The data of Schulz[5] are plotted below 3 eV. (From Ehrenreich and Philipp[6])

X-ray range and beyond, extend the results as far as we wish. For example, as shown in Fig. 10, it is possible to construct a curve of the absorption constant for aluminum out to a million eV. (The transitions are associated with the L and K absorption edges of aluminum.) From this, in turn, we can determine both optical constants by means of a Kramers–Kronig analysis.

The noble metals

Let us now turn to a group of metals having somewhat more complicated optical properties—namely, the noble metals, silver and copper. In Fig. 11 we observe at low frequencies the 100 per cent reflectance, already pointed out for aluminum, that is characteristic of the free-electron range. However, near 4 eV we note a dramatically sharp dip, with the reflectance dropping from nearly 100 per cent to less than 1 per cent in a fraction of a volt. Such rapid variation of the optical properties is most unusual in a solid.

At energies beyond the dip the structure is characteristic of interband transitions. Here the transitions are not confined to a small energetic region, as in the case of aluminum, but rather are widely dispersed, as in most metals. In the present case, the reason is that the d bands lie close to the Fermi level and have oscillator strengths—or equivalently, for the Lorentz model, natural spring frequencies—that extend over a large energy range.

Let us return to the explanation of the dip. We have already seen that a sharp decrease in the reflectance is usually associated with a plasma resonance. Figure 12, which shows the real and imaginary parts of the dielectric constant and the energy loss function for silver, also exhibits a sharp peak in the energy loss function near 4 eV. In this range ϵ_1 goes through zero and ϵ_2 is very small. However, in the present case, the decrease in reflectance is immediately followed by an equally sharp rise.

To interpret this behavior, let us consider Fig. 13 and first examine the schematic diagram showing the behavior of ϵ_2 at low energies. After the rapid decline of ϵ_2 with increasing energy associated with free-carrier effects, there occurs a sharp rise near 4 eV which may be interpreted in terms of transitions from the d bands to the Fermi level. The correctness of this identification has been substantiated by detailed calculations using the calculated band structure.

It should be noted that it is this transition that is responsible for the color of the noble metals. In copper and gold it occurs at 2 eV, or about 6000 Å. This wavelength lies in the red part of the spectrum. As a result of this absorption, copper and gold have the characteristic red-yellowish color that we associate with them, whereas silver is different because this characteristic transition lies in the ultraviolet region.

Returning to the structure of ϵ_2, we see that there is a clear-cut separation into the free-electron part associated with intraband transitions of s electrons and the bound, oscillator-like part associated with interband transitions of d electrons. The resultant total dielectric constant consists simply of a sum of these two parts. In the low-frequency range only the free electrons contribute to the optical properties, whereas in the high-frequency range the behavior is dominated by the bound electrons. We observe that ϵ_1, which too can be decomposed into free and bound parts, also has a peak associated with this transition. Significantly, however, ϵ_1 becomes positive before ϵ_2 begins its rise and while it is still very small. This behavior results in the observed very-well-defined plasma resonance, which is quenched at very slightly higher energies by the strong interband effect associated with the rapid rise in ϵ_2. It is a competition of these two effects that is responsible for the sharp dip in the reflectance. The separation of ϵ_1 into free and bound contributions, which is indicated schematically in Fig. 13, shows that the plasma resonance occurs because the interband contribution to ϵ_1 forces the total dielectric constant through zero. As a result, we may visualize

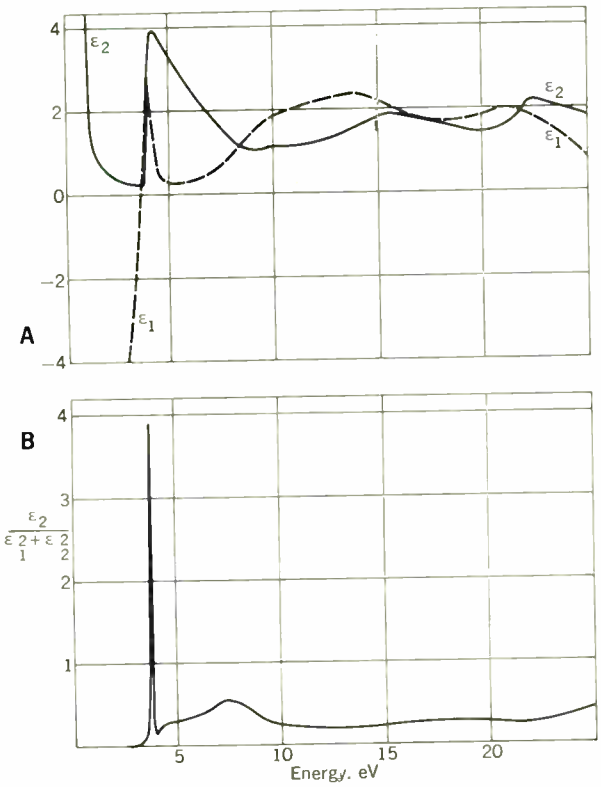


Fig. 12. Spectral dependence of (A) the real and imaginary parts of the dielectric constant and (B) the loss function for silver. (From Ref. 6)

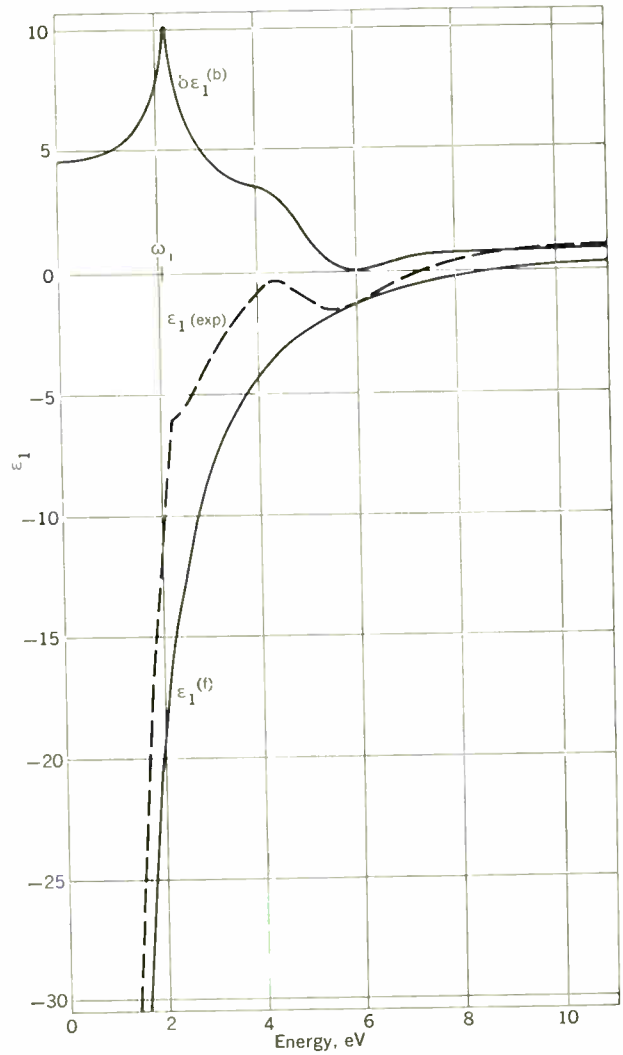
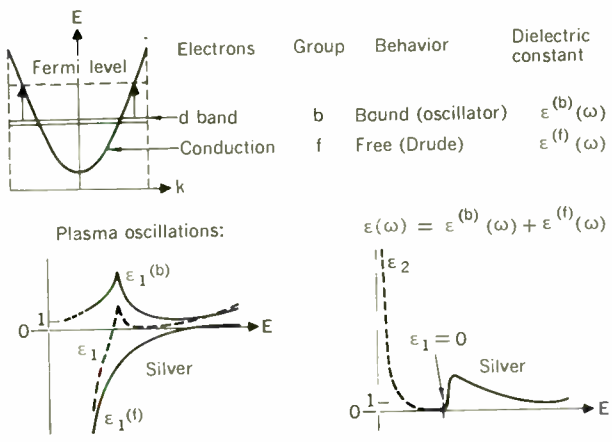


Fig. 14. Decomposition of the experimental values of ϵ_1 for copper into free and bound contributions $\epsilon_1^{(f)}$ and $\delta\epsilon_1^{(b)}$. The threshold energy for interband transitions is indicated by ω_1. (From Ref. 6)

Fig. 13. Types of optical transitions in noble metals.



this plasma oscillation as a collective process involving both the *s* and the *d* electrons.

It is interesting to note that the corresponding resonance is not observed in copper, since, as shown in Fig. 14, the interband contribution due to the transitions from the *d* band to the Fermi level, responsible for its characteristic color, occurs at sufficiently lower energies that ϵ_1 does not pass through zero and hence cannot fulfill the necessary condition for plasma oscillations.

To show how well the Drude theory fits the free-carrier part for silver and copper, Fig. 15 presents a comparison of theory and experiment for each of two closely lying values of the effective mass. The experimental results in the present case are obtained from the separation of the intraband and interband contributions; theoretical results are obtained by substitution into the simple Drude formula discussed earlier, regarding the effective mass as an adjustable parameter. The conductivity relaxation time for scattering, which also appears in the Drude formula, is entirely unimportant in this frequency range. The shape and magnitude of the results are seen to be very well represented by the theory for reasonable values of the effective mass.

Transition metals

In a similar way we can try to understand even more complicated metals such as the transition metals. These materials, which are frequently ferro- or antiferromagnetic, have *d* bands that intersect the Fermi level. Since these bands are very narrow and of high multiplicity, we would expect interband transitions for the transition metals to set in at even lower energies than in copper, in which the *d* bands are still appreciably

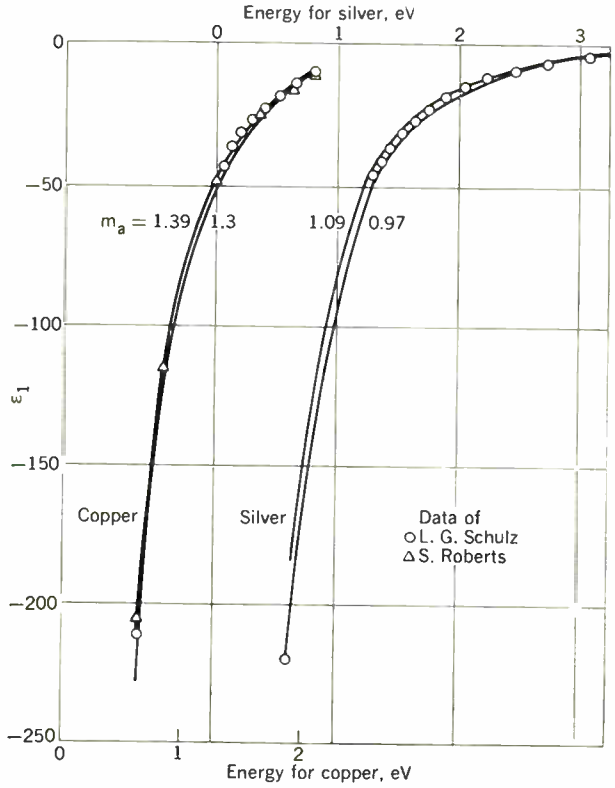
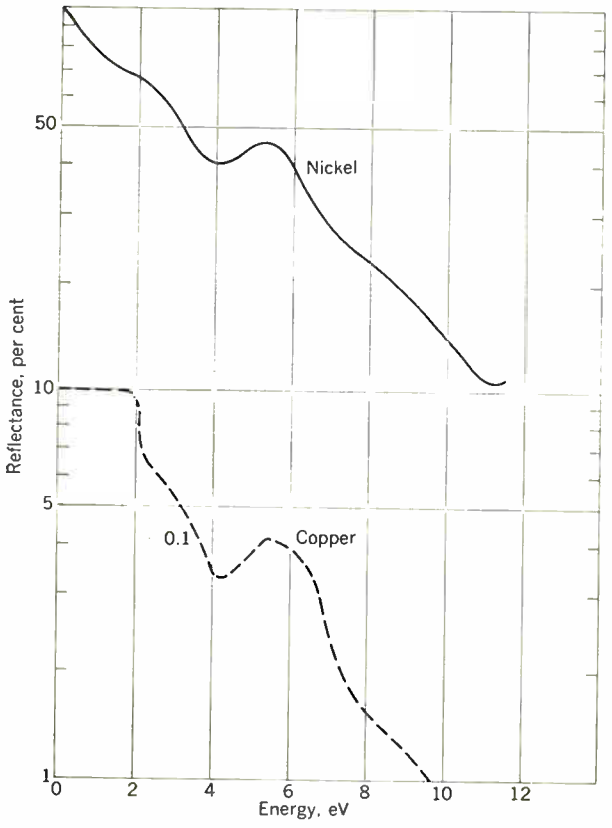


Fig. 15. Comparison of experimental and theoretical values of ω_1 for silver and copper in the free-electron region for several values of optical mass m_a. Points correspond to experimental data of Schulz[6] and Roberts.[7] (From Ref. 6)

Fig. 16. Spectral dependence of the reflectance of copper[6] and nickel. (From Ehrenreich, Philipp, and Olechna[8])



below the Fermi surface. Since we have seen that the onset of interband transitions is effective in reducing the reflectance below the 100 per cent value that we associate with free-carrier effects, we would expect the reflectance of the transition metals to decrease at somewhat lower energies.

That this is the case can be seen by comparing the reflectance for nickel and copper (Fig. 16). It can be shown that the interband transitions in nickel set in already at 0.3 eV. As a result of this, the reflectance has almost no flat plateau near 100 per cent as in the case of copper. These low-lying interband transitions are appreciably masked by free-carrier effects and therefore are not conveniently studied through the ordinary optical properties. Recently we observed that the ferromagnetic Kerr effect is not obscured by free-carrier effects and therefore is able to give precise information about the interband transitions in this range. A study of these transitions is very important, since it provides clues as to the ferromagnetic splitting of the *d* bands and other interesting aspects of the band structure of ferromagnetics.

The preceding applications to actual metals illustrate the great usefulness of optical measurements over an extended energy range in furthering the fundamental understanding of solids. It seems clear that investigations of this nature will continue to be as useful as they have proved to be for many years.

REFERENCES

1. Philipp, H. R., and Ehrenreich, H., "Optical Properties of Semiconductors," *Phys. Rev.*, vol. 129, Feb. 15, 1963, pp. 1550–1560.

2. Ehrenreich, H., Philipp, H. R., and Segall, B., "Optical Properties of Aluminum," *Ibid.*, vol. 132, Dec. 1, 1963, pp. 1918–1928.

3. Segall, B., "Energy Bands of Aluminum," *Ibid.*, vol. 124, Dec. 15, 1961, pp. 1797–1806.

4. Philipp, H. R., and Ehrenreich, H., "Optical Constants in the X-Ray Range," *J. Appl. Phys.*, vol. 35, May 1964, pp. 1416–1419.

5. Schulz, L. G., "Experimental Study of Optical Properties of Metals and Relation of Results to Drude Free Electron Theory," *Phil. Mag. (Suppl.)*, vol. 6, Jan. 1957, pp. 102–144.

6. Ehrenreich, H., and Philipp, H. R., "Optical Properties of Ag and Cu," *Phys. Rev.*, vol. 128, Nov. 15, 1962, pp. 1622–1629.

7. Roberts, S., "Optical Properties of Copper," *Ibid.*, vol. 118, June 15, 1960, pp. 1509–1518.

8. Ehrenreich, H., Philipp, H. R., and Olechna, D. J., "Optical Properties and Fermi Surface of Nickel," *Ibid.*, vol. 131, Sept. 15, 1963, pp. 2469–2477.

BIBLIOGRAPHY

LaVilla, R., and Mendlowitz, H., "Optical Constants of Aluminum in Vacuum Ultraviolet," *Phys. Rev. Letters*, vol. 9, Aug. 15, 1962, pp. 149–150.

Philipp, H. R., and Taft, E. A., "Optical Constants of Germanium in the Region 1 to 10 eV," *Phys. Rev.*, vol. 113, Feb. 15, 1959, pp. 1002–1005.

Philipp, H. R., and Taft, E. A., "Optical Constants of Silicon in the Region 1 to 10 eV," *Ibid.*, vol. 120, Oct. 1, 1960, pp. 37–38.

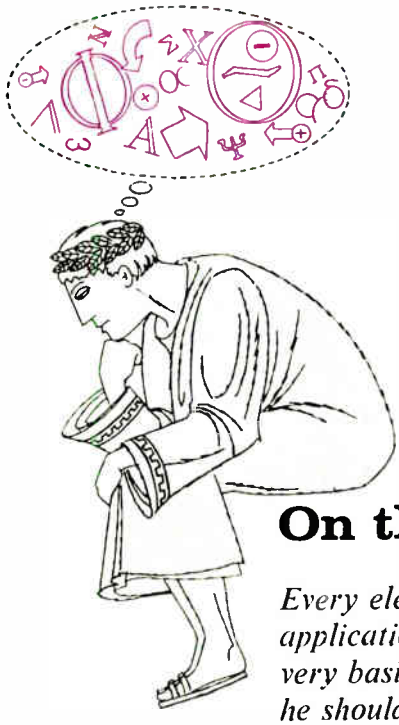
Philipp, H. R., and Taft, E. A., "Optical Properties of Diamond in the Vacuum Ultraviolet," *Ibid.*, vol. 127, July 1, 1962, pp. 159–161.

Phillips, J. C., "Optical Absorption in Germanium," *J. Phys. Chem. Solids*, vol. 12, 1960, pp. 208–209.

Pines, D., *Elementary Excitations in Solids*. New York: W. A. Benjamin, Inc., 1963, chap. 4.

Seitz, F., *Modern Theory of Solids*. New York: McGraw-Hill Book Co., Inc., 1940, chap. 17.

Stern, F., "Elementary Theory of the Optical Properties of Solids," in *Solid State Physics*, vol. 15, F. Seitz and D. Turnbull, eds. New York: Academic Press, Inc., 1963.



On the nature of the electron—Part I

Every electrical engineer is familiar with the applications of the electron, since its activities provide the very basis for his specialty. However, he should also understand something of the actual nature of this fundamental particle of matter

J. L. Salpeter *Adelaide, South Australia*

The electron was discovered in 1897 by J. J. Thomson. By deflecting cathode rays in electric and magnetic fields he could determine the ratio of charge to mass (e/m = "specific charge") of the particles constituting cathode rays, and he found that this ratio remained the same, no matter what gas had been used to fill the cathode-ray tube or what material constituted the electrodes. Assuming that the charge e is the same as in electrolytic experiments, the mass m of the cathode-ray particle could be determined. It was a great surprise to find that this mass was about 1/1800 of the mass of the lightest atom—hydrogen. The conclusion drawn from these results was that electricity has an atomistic structure; i.e., there exists a smallest amount of electric charge and any charge found in nature is an integral multiple of this elementary charge. This elementary charge, however, is inseparably combined with mass, as found by Thomson in the new particles, the electrons. Apart from being the atom of electricity, the electron is a constituent of every "atom" of every element. Normally it is bound to the atom, but by special processes it can be liberated from it and become a "free electron." Actually the "bound electron" was discovered slightly earlier, in 1896, by Zeeman, a Dutch physicist, who found that it is possible to split the spectral lines of a light-emitting element if this element is placed in a magnetic field. The interpretation given by Lorentz to the "Zeeman effect" was briefly this: the emitted light is generated within the atom by an oscillating electric particle, endowed with charge and mass, which later proved identical with the particle found by Thomson. The first one was in the "bound" state, the latter in the "free" state.

With the discovery of the emission of electrons by hot metals the "electronic age" was inaugurated. It is not necessary to relate here what part the free electrons played and are playing in vacuum tubes, radio valves, rectifiers, etc. With the discovery of transistors and conduction by "holes" the bound electron (valency electron) became interesting to the electronic engineer too. However, in this article we are not going to discuss the engineering applications of free and bound electrons. We are going to meet the electron not professionally, but, as it were, in its private life. What does it really look like? How does it behave when nobody looks, and what is its social life (if any)?

Admittedly, the answers to questions of this kind are not necessary for the efficient use of vacuum tubes and other electronic devices. But, after all, man does not live by bread alone and at all times philosophy has drawn heavily on science on respect to questions concerning the nature of our surroundings. Knowledge of the nature of the fundamental particles may help us to obtain an answer to the eternal question, "Who are we?" Science is often regarded as a means to help us invent more and better gadgets, but scientists themselves rather wish science to be appreciated for its own sake, and for the sake of its philosophic consequences. Of all people who are not physicists themselves, the electrical engineer is the most familiar with the electron, although not on a personal footing. The intention of this article is to arouse his scientific curiosity.

Essentially full text of an article appearing in the December 1957 issue of the PROCEEDINGS OF THE IRE, and also scheduled for publication in the March issue of the IEEE STUDENT JOURNAL. Part II, previously unpublished, will appear in the April issue of SPECTRUM and the May issue of the STUDENT JOURNAL.

Atomic structure of charge, matter, energy

The 19th century saw great successes in the physical sciences. Indeed, the successes were so great that some sciensists believed that everything that could be discovered had been discovered, and in the future it would be only necessary to consolidate the achievements. The disciplines of mechanics, hydrodynamics, elasticity, electrodynamics, optics, and so on, were all rounded off and perfected, mainly in the language of partial differential equations. If the Creator of the world was a mathematician, as some maintained, He must have been very fond of partial differential equations. These equations are admirably suited to describe phenomena in continuous media and "fields," and their precision and beauty induced some students to adhere to the idea of continuous matter longer than was warranted. Still, at the end of the century there were some prominent scientists, such as the great chemist W. Ostwald, who regarded the atomic structure as an unproved hypothesis. Today we still treat space and time (or rather space–time) as continuous entities, but everything that happens and proceeds in space–time does so atomistically. We no longer believe that God is particularly fond of partial differential equations, and even the partial differential equation on which wave mechanics is based serves rather to conceal than to reveal the actual state of affairs.

"Atom" is a Greek word meaning indivisible, and the hypothesis of atomism is of Greek origin. The Greeks had no physical science to speak of and the theory of atoms was built not on the basis of observation and experiments, but on pure speculation. Some maintain that it was pure coincidence that their theory proved correct after more than 2000 years and would deny any merit to Democritos, who is regarded as the father of the atom. However, Schroedinger tries to vindicate the ancient philosophers by showing that it is possible to arrive at the atomic hypothesis by pure speculation, based only on the primitive observation that matter can be condensed and rarified. Suppose now we regard matter as continuous and let us try to imagine what happens when we condense it. We could start by mentally subdividing the piece of matter into, say, 1000 pieces and condensing each of the 1000 pieces. But then we should have to subdivide each of the 1000 pieces again into 1000 pieces, and continue doing so ad infinitum. Subdividing matter into finite pieces is of no use if matter is continuous, and so let us single out a finite number, although a very large one, of geometrical points within the piece of matter and watch how those points come nearer to each other when the matter is condensed. Yet, since we started with a finite number of points, we left out some points between the selected ones. If we make a second finite selection, there will be still some more points omitted, and the points in the condensed piece will never come into contact with each other.

Let us think about the situation. The paradox encountered in this mental exercise can be illustrated by the following geometrical consideration. Let us visualize two straight lines, one 1 inch long, the other 10 inches long. We are going to establish a one-to-one correspondence between the points of the two lines. To begin with, the two end points of the 1-inch line correspond to the two end points of the 10-inch line. Then let us select arbitrarily any point on the 1-inch line, measure its distance from the left end, say, multiply the distance by 10, and mark on the 10-inch line a point at that distance from the left end. Conversely, we can select any point on the 10-inch line, measure its distance, divide it by 10, and mark it correspondingly on the 1-inch line. To each point of the 1-inch line corresponds a point on the 10-inch line and, conversely, to each point of the 10-inch line corresponds a point on the 1-inch line. The 1-inch line is as rich in points as the 10-inch line, which is obviously paradoxical but uncontrovertible. These paradoxes usually crop up when we deal with infinities and we have seen that the concept of continuous matter involves infinities. A good example of what happens if we treat infinities seriously is this. Let us imagine a hotel with an infinite number of rooms and let the hotel be fully occupied. Despite the fact that every room of the hotel is occupied, it is possible to accommodate many more guests in this hotel, in fact, an infinite number of additional guests. All we have to do is to move the guest in room no. 1 to room no. 2, the guest in room no. 2 to room no. 4, the guest in room no. 4 to room no. 8, and so on. This process can be repeated without limit. We should not be afraid that some people at the end of the hotel will be turned out, because there is no end.

All those disabilities connected with infinities are avoided if we conceive of matter as consisting of a finite number of ultimate, indivisible particles; i.e., atoms. This reasoning is of course no substitute for the experimental proof of the existence of atoms, but we can give the ancient Greeks credit for arriving at the concept of the atom by a reasoning that today still has an appeal for us. This is not to say that the concept of the ultimate, fundamental particle does not involve philosophical difficulties of its own, but in dealing with them we are assisted by Nature itself, since all experimental evidence speaks in favor of the existence of ultimate, no longer divisible, particles.

Fundamental particles

The term "atom" is today a misnomer. We know that the atom consists of a positive nucleus with a number of negative electrons moving around it. The electrons that perform such useful functions in vacuum tubes for us were originally constituents of the atoms of the cathode. The atom is then no longer indivisible, but we can at least detach some electrons from it. Even the nucleus (with the exception of hydrogen) is not indivisible. The nucleus of hydrogen is a fundamental particle—the "proton"—while the nuclei of all heavier elements consist of protons and neutrons. The neutron is a fundamental particle of approximately the same mass as the proton, but without charge. There are a number of other fundamental particles, but we shall be concerned here solely with the negative electron, and mention only that there exists a positive electron, termed "positron" (or antielectron), but this does not normally occur in matter. We may also mention that an antiproton and an antineutron (i.e., a negative proton and a neutron with a reversed magnetic moment) are conceivable and have been discovered.

How is the electron to be imagined? In the early days of electron theory it was an acceptable procedure to assume some shape of, and charge distribution within, the electron, for instance, sphere or ellipsoid, to regard it as rigid or deformable, and so on, and to calculate its behavior accordingly. Today we would regard this procedure as naive. We might as well imagine the electrons

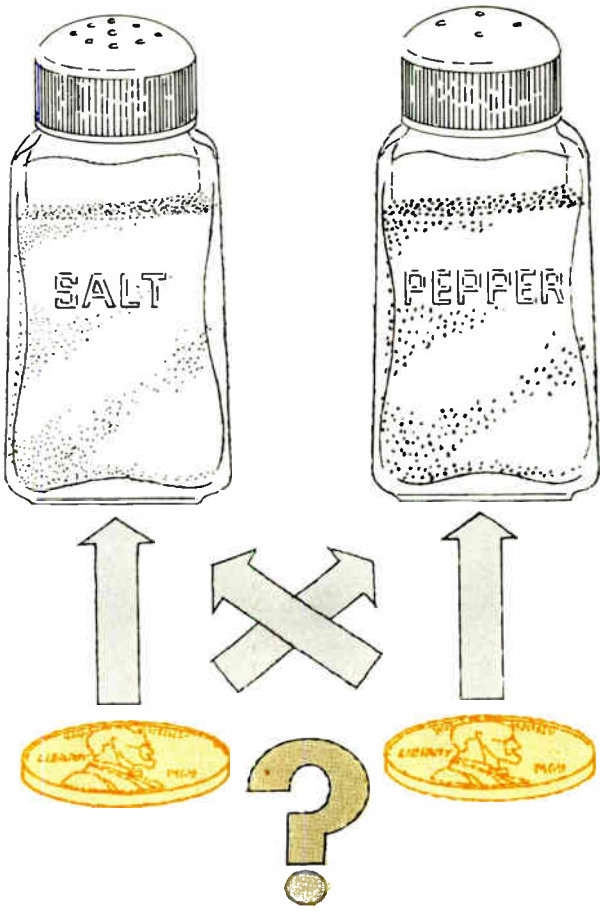
as tiny hard billiard balls, painted blue and red, with blue standing for negative, red for positive charge. If we consider charge distribution within the electron, the question obtrudes itself, how does this charge manage not to explode (or are the charge elements within the electron perhaps glued together with some cement)? Today we are not encouraged to try to explore the structure of the electron itself. An ultimate particle has no structure, and we have to stop somewhere exploring structure. To begin with, there is charge and mass connected with the electron, but we are not to picture charge and mass as separate entities embraced by the electron. We should, rather, regard charge and mass as properties of the electron. Needless to say, even if we could see the electron, we could not discern its color, because its diameter would be so much smaller than the wavelength of visible light that it could not possibly reflect any.

Although the electron is one of the constituents of visible and tangible matter, it differs from ordinary matter by virtue of its indivisibility. This indivisibility is not only to be accepted as a fact, but even in our imagination we are not to try to subdivide the electron or we shall be confronted with difficulties similar to those that arise, for example, when we deal with infinities. As Dirac puts it, the smallness of an ultimate particle is absolute, not relative. In the process of subdividing matter, once we reach the ultimate particle we must be prepared to encounter behavior and properties quite unlike those connected with ordinary matter.

Properties of the electron

The electron is characterized by three properties: charge, mass, and spin. We already have discussed charge and mass. Apart from these, the electron is endowed with spin; i.e., it rotates around its diameter with some specific unchangeable velocity. By no means is it possible either to speed up or to retard this rotation. Spin was postulated by Uhlenbeck and Goudsmit in order to explain spectral lines in accordance with quantum theory. Later Dirac derived spin by applying the principle of relativity to the theory of the electron, and has shown the spin of fundamental particles to be a consequence of the symmetry of the universe with respect to space and time. Although nobody has ever observed the electron in the act of spinning, the spin seems well founded in basic theory. It is responsible for the magnetic behavior of the electron; in fact, due to the spin, the electron, apart from incorporating elementary charge, represents an elementary magnetic dipole.

So far we have been discussing a single electron and its three properties. As soon as we consider an aggregate consisting of two or more electrons, a property becomes apparent that has no meaning for a solitary electron. This property is the "indistinguishability" of electrons. One electron is exactly like its fellow electron and it is absolutely impossible to tell one from another. This is a very remarkable property with some very important consequences. We have warned before that in the process of subdividing matter, as soon as we arrive at the ultimate particles we are bound to encounter behavior that is very different from the behavior of macroscopic, visible, and tangible matter. Some of those new features, such as the dual aspect (wave particle aspect), are very difficult to visualize, but others, such as indistinguishability, although not occurring in our routine experience, are



possible to comprehend by the process of going to the limit. In our everyday experience there are no two objects equal in every respect. Among the 2½ billion inhabitants of the earth, there are no two individuals absolutely alike. This would be the case even if we allowed for the differences in age or if we took into consideration all the people who ever lived on earth. The greatest similarities we encounter are among twins, but even with twins who cannot be told apart by their own mother, we would detect differences by careful examination (fingerprints, weight, size). Or take balls for ball bearings as an example. They may look alike, but by weighing them on a sensitive scale, we would most certainly detect small differences.

Let us see what classical physics had to say in this matter. Even in classical physics it was assumed that mass and charge of all electrons were equal, but there was no emphasis on this equality. Within the errors of measurement, mass and charge proved to be equal for all electrons and there was no phenomenon known which would call for fluctuations in the values of charge and mass as an explanation. It will be realized that only a theory can postulate equality, because measurements alone, no matter how accurate, are encumbered by unavoidable errors and we could always assume—if we wished—that there are small fluctuations, which however are masked by the errors. No measurements could ever prove absolute equality. But in classical physics the matter was of no importance, because there were no effects known that could be explained either by accepting or rejecting the notion of absolute equality.

It is different in quantum mechanics. Here the indistinguishability of electrons generates a new kind of

interaction, a new kind of nondynamic force which explains such things as chemical valency, cohesion of crystals, and ferromagnetism in iron, nickel, and cobalt.

Indistinguishability also throws some light on the nature of fundamental particles. We remember the story of the little girl whose mother gave her two pennies, asking her to buy salt with one penny, and pepper with the other. After a while the girl returned with the pennies because she had forgotten which of them was meant for the salt and which for the pepper. For her the pennies were distinguishable and she did not understand that the values incorporated in the two pennies were indistinguishable. (In anticipation of such difficulties, the term "pennyworth" was introduced into the English language, but this has not been done in other languages.)

In the realm of energy we have no difficulty in realizing that two equal units of energy are indistinguishable—we could hardly tell the kilowatthour we consumed yesterday from the kilowatthour we consume today. And yet, according to Einstein, mass and energy are mutually convertible into each other. It should not be surprising then that entities that were indistinguishable in the form of energy remain indistinguishable in the form of matter. However, we shall see that indistinguishability can be regarded as confirmed by experience.

Perfection and imperfection

The indistinguishability of fundamental particles also could be expressed by saying: They are perfect. There are only a few numbers that characterize an electron or a proton, but there are absolutely no deviations whatsoever from those numbers. It is different in the realm of macroscopic objects or objects of our everyday experience. "Nothing on earth is perfect," we often say. The human body is nominally symmetrical, but we know that the organs within the body are not arranged symmetrically. Even the face is never quite symmetrical. If we cut a photograph of a face in half vertically and replace one half by a mirror image of the other, we obtain a symmetrical face, but such a face, surprisingly enough, looks unnatural.

Perhaps we have a better chance to find perfection among nonliving objects. A diamond crystal may look like a near approach to perfection. A man in the street values in the diamond its optical properties and its hardness. The physicist is more impressed by the regularity of the arrangements of its atoms in a space lattice. The surfaces of a crystal may have a pretty appearance and shape (configuration of the external surfaces). We rather appreciate the crystalline nature of an object by inspecting its internal arrangement in form of a space lattice of perfect periodicity. But is this periodicity really perfect? Since the discovery by Laue of X-ray diffraction by a space lattice, thousands and thousands of crystals have been investigated regarding their space lattice, but none of them was perfect. Some years ago, a symposium was held on the subject, "Imperfections in Nearly Perfect Crystals." This title implies a regret and resentment that but for a few incidental imperfections the crystals would have been perfect. However, the imperfections were not incidental. They are in fact unavoidable, and the extent of any kind of imperfection can be determined thermodynamically.

The most common imperfection in a crystal is chemical impurity. Minerals found in nature are always contaminated but we can purify them by chemical and other methods in the laboratory. Methods are now available (for instance, the melted zone method of purifying germanium) which achieve a purity that would have been considered impossible a few years ago. However, an absolute absence of foreign atoms is impossible, not just for practical reasons but in principle.

Briefly, the reason is this: Let us say, we have one impurity atom per cubic millimeter of the crystal. This impurity atom may occupy one of the sites occupied ordinarily by one of the atoms of the substance of which the crystal is composed. Since there are roughly 10^{19} atoms in one cubic millimeter there are 10^{19} ways in which the contamination of the crystal can be realized. On the other hand, there is only one way in which a perfectly pure crystal can be realized. Hence the probability of the occurrence of a contaminated crystal is many orders of magnitude larger than for a perfect crystal. However, the impurity atom within the crystal lattice may involve more energy than an atom of the host substance and so a compromise will eventuate between the requirements of a minimum energy and maximum entropy. This is essentially the basis of the thermodynamic reasoning which leads to a quantitative evaluation of the extent to which contamination is to be expected.

Macroscopically nothing is perfect; on the level of fundamental particles everything is perfect.

Pauli's exclusion principle

The engineer is familiar with the free electron; i.e., the electron outside the atom. We know that the density of free electrons in vacuum or in an ionized gas never exceeds, say, 10^{15} per cm^3—at least in terrestrial laboratories. This means that the mutual distance of free electrons is on the average never less than 10^{-5} cm, which means 1000 interatomic distances in a solid. At these distances the only interaction to be considered between the electrons is the electrostatic repulsion due to the charges of the electrons. It is not even necessary to be aware of the electron spin; we shall see later why there is little chance to observe spin of a free electron. Likewise, the indistinguishability of electrons or otherwise does not play any prominent part with free electrons.

Within an atom (or molecule or crystal) the electron density is of the order of 10^{23} to 10^{25} per cm^3 and the mutual distances are of the order of 10^{-8} cm or less. At these distances the electrons no longer behave simply as ordinary particles of small size; they show features characteristic of fundamental particles. One of them is a peculiar mutual dislike of electrons of the same sign spin. We deliberately avoid the term "force" or "repulsion," because it is more than a repulsion. According to Pauli's exclusion principle, we shall never find two electrons of the same sign spin in the same orbit of an atom. Now, it would not be correct to say: let us place two electrons of the same spin in the same orbit and we shall see that due to an enormous short-range mutual repulsion one of them will leave the orbit. This statement would not be correct, because, to begin with, no power on earth would be able to place two equivalent electrons in the same orbit. We shall discuss this curious impossibility later, but let us first describe more fully this fundamental exclusion principle.

The principle is best described in its historical con-

text. We mentioned previously that the Zeeman effect (splitting of spectral lines in presence of a magnetic field acting on the emitting atom) had been interpreted by Lorentz on the assumption that an electric particle bound to the atom by a quasi-elastic force is oscillating around its equilibrium position. But what does this particle (electron) do when it is not oscillating (i.e., when the atom is not emitting light)? It could not be at rest, because nothing would prevent it then from falling into the nucleus. The planets do not fall into the sun because they have sufficient kinetic energy to enable them to cruise around the sun on a circular or elliptical orbit. Electrons are attracted to the positive nucleus by Coulomb's forces (inverse square law) just as the planets are attracted to the sun by gravitation and, accordingly, if endowed with sufficient kinetic energy, they can avoid falling into the nucleus by cruising around it. However, unlike the planets, an electron following a curved orbit emits electromagnetic radiation and so a cruising electron would continually lose energy, the diameter of its orbit would decrease, and eventually it would fall into the nucleus.

Whether we assume the electron at rest or in motion, the classical theory saw no way of preventing the collapse of the atom. At this stage Niels Bohr proclaimed his revolutionary theory of the atom according to which the electron can move around the nucleus without radiating and losing energy, provided its total energy is one of a set of discrete energy levels, characteristic for the element concerned. There exists for each chemical element a set of permitted energy levels, the lowest of which belongs to the "ground state." In the ground state the electron or electrons are nearest the nucleus. If the atom absorbs a certain amount of energy, the electron can be lifted to the next higher level and the atom is said to be in an "excited state." On dropping from an excited to the ground state, the electron loses energy, which is radiated as light. No intermediate energy levels are permitted. This is at complete variance with the behavior of a free electron and with the classical theory of electromagnetism. In classical theory we can endow an electron with any amount of energy and according to classical theory any acceleration of the electron is accompanied by radiation. (Motion on a curved orbit involves acceleration.) These two rules were broken by Bohr's theory, but the theory proved successful in explaining the numerical relationships of spectral lines and was accepted enthusiastically by physicists.

The hydrogen atom is the simplest of all. It consists of a positive nucleus (proton) of unit charge and with one electron moving on an orbit of radius 0.53 Å. The next in the periodic system of elements is helium with two electrons moving on an orbit of radius 0.30 Å. The third element is lithium with three electrons for which the orbit shell should have a radius 0.20 Å. (The radii become smaller with increasing charge of the nucleus, because the attraction becomes greater.) However, it was known that the diameter of the lithium atom must be much larger than 0.4 Å because the "atomic volume" of lithium was known to be large. For this reason it seemed unlikely that all three electrons were located on the same shell. It is here that Pauli's exclusion principle came into its own. This principle postulated that no electron shell shall contain more than one pair of electrons of the same state of motion—exactly as in Noah's Ark there was only one couple of each kind of animal. Only, in the case of electrons the couples are not distinguished by sex but by spin. According to this principle the third electron in the lithium atom has to go to a new shell (of 1.50 Å radius). The beryllium atom has four electrons and the fourth electron goes to the second shell, the third and fourth electrons having opposite spins. The second shell is hereby closed and the fifth electron in the boron atom goes to a new shell. As we progress in the periodic system of elements, new electrons (shells and subshells) are formed, and so we can say that Pauli's exclusion principle is responsible for the electron shell structure of the chemical elements. Without this principle all electrons would go to the first shell and it is impossible to say how the world would look in that case.

The exclusion principle is generally expressed in the following way: No two electrons of the same spin shall be in the same space and have the same state of motion. The "same space" refers, in the case of atoms, to the electron shells, but otherwise it refers to a system in which the electrons can interact. In old textbooks of physics we can find a principle of impenetrability—no two bodies can occupy the same place. The exclusion principle is less demanding; it requires only that no two electrons of the same spin shall occupy the same place and be in the same state of motion. On the other hand, the words "the same place" are not to be taken literally. The electrons occupy the same place, if they can interact.

With the advent of wave mechanics the exclusion principle has been incorporated in it and has been reformulated in a much more precise and quantitative manner. We are now going to discuss wave mechanics to the extent necessary for the understanding of the exclusion principle and its consequences.

Wave mechanics

The 20th century revolution in physics is best shown in the wave or quantum theory. The experimental basis on which wave mechanics was founded is electron diffraction by a crystal, say, of mica. Let us direct an electron beam onto a crystal plate, and place a photographic plate behind this crystal plate. After exposure and development of the photographic plate, we notice first of all a black circle corresponding to the cross section of the electron beam, but apart from this circle, we notice around it a regular diffraction pattern, quite similar to the diffraction pattern obtained by X rays. Diffraction is a clear demonstration of the wavelike character of the electrons, and yet their origin in the electron gun points to their particle aspect. The diffraction pattern does not depend on the intensity of the cathode ray. We can use a beam of very low intensity and expose the plate for a very long time and obtain the same pattern as we did with high intensity and short time. We can make the electron beam intensity so low that only one electron passes through the crystal at a time. How do the electrons know which point of the pattern each of them has to hit? There is only one answer to this question—each electron is itself a wave and each electron creates the whole pattern, but energy can be absorbed only in finite quanta and so the electron hits the photographic plate at one point only, the distribution of those points being given by a probability function. We are then faced with the following situation. The electron originates in the gun as a particle and is being absorbed by the photographic emulsion as a particle, but in between it behaves like a wave. This dual aspect of the electron "wave-particle" caused great concern and was

the subject of numerous discussions and controversies. (A suggestion has been made to call the electron a "wavicle," but the term alone, of course, does not solve any problems.)

In this article we wanted only to discuss a few properties of the electron, in particular "indistinguishability," because this is not beyond common sense apprehension and yet has far-reaching consequences. Indistinguishability is closely connected with the exclusion principle, and this in turn is best expressed in the language of wave mechanics.

We have seen that the phenomenon of electron diffraction suggests the wave nature of the electron. What exactly is it that vibrates in the electron wave? The answer is quite different from the answers to the same question for any other kind of wave. The intensity of the electron wave indicates the probability that we shall find the electron at a particular point x, y, z in space. As with any other kind of wave, the intensity is equal to the square of the amplitude and the amplitude is a function of the space coordinates, but the amplitude itself has no physical meaning. It is only an auxiliary quantity, whose sole purpose is to enable us to compute the intensity.

Apart from the amplitude, the shape of the wave is an important characteristic, and particularly the wavelength of the fundamental. In wave mechanics the wavelength λ is connected with the observable electron velocity v by the following relation:

$$h/\lambda = mv$$

where h stands for Planck's constant and m for the electron mass.

Let us now see how the wave nature of the electron helps us to understand the structure of the atom. As a preliminary experiment let us put the electron into a box, not literally a box with walls of cardboard or steel, but a box whose walls are potential barriers. The potential walls are high enough to prevent the electron from escaping, which means in the language of wave mechanics that the probability of finding the electron outside the box, or even at the walls, is zero. Consequently, the wave representing the electron is a standing wave within the box. Assuming the length of the box to be L, then the wavelength of the standing wave will be $2L$, or L, or $2L/3$, and so on, or generally

$$\lambda = 2L/n$$

where $n = 1, 2, 3, 4, \ldots$ Remembering the relation connecting λ with the electron velocity v, we see that this velocity can have only one value of a set of discrete values. This is a very strange and very important result. In classical mechanics an electron confined in a box can have any value we want to endow it with; we can change the velocity continuously. In wave mechanics only certain velocities are possible and if we want to increase or decrease these velocities we have to do it in finite jumps. What we have said about velocities applies, in our case, equally well to the energies, because the energy of the electron in the box is made up entirely of its kinetic energy. We arrive at the result that the energy of an electron in a box is capable of assuming only certain values of a set of discrete levels given by the length of the box.

To some rough approximation we can regard the atom as a kind of a potential box in which the electron is held captive by the attraction of the nucleus and we can transfer the results just mentioned to the atom. In this way we arrive at the postulates of Bohr's theory of the atom, but with this difference: while in Bohr's theory the postulates of discrete energy levels appear suddenly from nowhere, they are here the natural outcome of identifying the electron with a standing wave. However, as our aim was to express Pauli's exclusion principle in a more rigorous manner, let us return to this principle.

Symmetrical and antisymmetrical wave functions

The original wave equation, derived by Schroedinger in 1926, refers to a single electron. We are not going to discuss its mathematics; we will only describe the meaning of the wave parameters. We have mentioned how the wavelength is linked with the electron momentum (product of mass and velocity), while the frequency is proportional to the total energy of the electron. The quantity corresponding to the refractive index in case of light waves is determined in the case of electron waves by the electric potential at a given point (x, y, z). The wave amplitude has no physical meaning, but the wave intensity indicates the probability of finding the electron at a given point. Originally the wave intensity was interpreted as charge density, and this is still true today if we consider the time average of the charge density over an appropriate period of time. If, for instance, we let a weak electron beam penetrate a thin crystal of mica and watch blackening of a photographic film placed behind the mica, the pattern of blackening will be given by the wave intensity.

The wave equation for a single electron has been applied successfully to the hydrogen atom, where we have a single electron in the field of the positive proton. However, in the next element, helium, we have two electrons, in lithium three electrons, and so on (and in uranium as many as 92 electrons). The first rule to be observed in case of a system containing more than one electron is Pauli's exclusion principle—no more than one electron of the same spin in one orbit. Pauli's principle is quite independent of wave mechanics, and in fact overrides it; the wave equation for two electrons has to be formulated in such a way as to exclude automatically the simultaneous presence of two electrons of the same spin in the same place.

If $\zeta_1(x_1, y_1, z_1)$ and $\zeta_2(x_2, y_2, z_2)$ are the two wave functions of the two electrons, it would be tempting to write the wave functions for the case of two electrons as the sum of $\zeta_1(x_1)$ and $\zeta_2(x_2)$:

$$\psi = 0.71[\zeta_1(x_1) + \zeta_2(x_2)]$$

where $\zeta_1(x_1)$ and $\zeta_2(x_2)$ stand as abbreviations for $\zeta_1(x_1, y_1, z_1)$ and $\zeta_2(x_2, y_2, z_2)$. The factor 0.71 will be explained presently. The function $\psi = \zeta_1 + \zeta_2$ would take account nicely of interference and diffraction phenomena of electron waves and would be in strict analogy to other waves encountered in physics, such as optical and accoustical waves. However, we have said before that the wave intensity expresses the probability of finding the electron at a given point. Similarly, the intensity of a two-electron wave function should stand for the probability of finding electron 1 at point 1 and electron 2 at point 2. With the requirement of representing the probability, the functions ζ_1 and ζ_2 are as a rule written in "normalized" form, i.e., provided with such a factor as to make the integral of ζ_1^2 and ζ_2^2, respectively, over the whole space equal to unity. A probability of one means certainty, and since we

are certain to find the electron *somewhere* in the whole infinite space, the integral (the sum of all probabilities) must equal unity. Let us now similarly integrate ψ^2 over the whole space. If we form the square of ψ we obtain

$$\psi^2 = 0.5(\zeta_1^2 + 2\zeta_1\zeta_2 + \zeta_2^2)$$

The integral of ψ^2 over the whole space will then be equal to unity plus the integral of $\zeta_1\zeta_2$. Now by analogy with optical waves we realize that the integral of $\zeta_1\zeta_2$ over the whole space must be zero. In optics the wave intensity stands for the energy and if we let two waves interfere we know that whatever happens the energy will be conserved; i.e., integral of $(\zeta_1 + \zeta_2)^2$ must be equal to integral of $\zeta_1^2 + \zeta_2^2$, or the integral of $\zeta_1\zeta_2$ must be zero. With electron waves it is the charge that is being conserved, which is satisfactory, but what about probability? The sum $0.5(\zeta_1^2 + \zeta_2^2)$ stands for the probability of finding *either* electron 1 at point 1 *or* electron 2 at point 2, but this is not what we were after. We wanted an expression for the probability of finding electron 1 at point 1 *and* electron 2 at point 2. The probability that two events will take place is equal to the product of the probabilities of the single events. Let us then write

$$\psi = \zeta_1\zeta_2$$

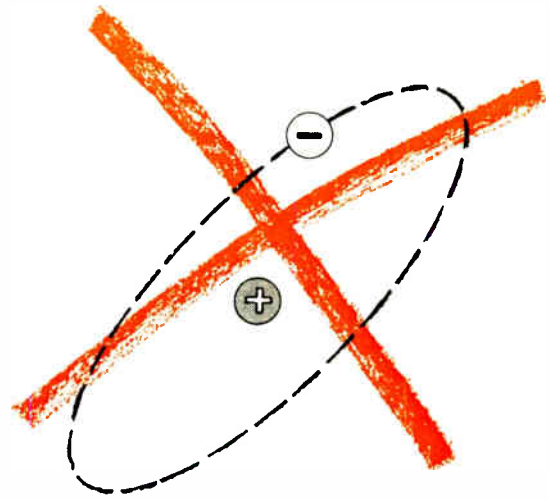
and we shall then have, as required,

$$\psi^2 = \zeta_1^2\zeta_2^2$$

At the same time we realize that the integral of ψ^2 over the whole space will be equal to unity. (First, we keep point 2 constant and integrate ζ_1^2 over the whole space thus obtaining ζ_2^2 as a result, and then we integrate ζ_2^2 over the whole space and obtain unity.) This would be satisfactory but we remember that electrons are indistinguishable and therefore it is not permissible to label electrons as we did in the reasoning leading to $\psi = \zeta_1\zeta_2$. As long as we are considering single electrons, it is a matter of course to establish a function ζ_1 for one electron and a function ζ_2 for another electron. As soon, however, as we have a system consisting of two electrons, the wave function is required to be formulated in such a manner that interchanging in the two electrons does not alter the function (or, rather, its meaning). We can arrive at such a formulation in the following way. Given $\zeta_1(x_1)$ and $\zeta_2(x_2)$ we know that $\zeta_1(x_1)\cdot\zeta_2(x_2)$ leads to the probability of finding electron 1 at point 1 and electron 2 at point 2. By exchanging electron 1 with electron 2 we obtain the product $\zeta_1(x_2)\cdot\zeta_2(x_1)$, which leads to the probability of finding electron 2 at point 1 and electron 1 at point 2 (this is really the same probability as the previous one since the electrons are not distinguishable). The wave equation is a *linear* differential equation and, consequently, if we have two solutions of the equation, any linear combination of the two solutions will be a solution again. To find a solution that is invariant in respect to the interchange of the two electrons, we put

$$\psi = 0.71[\zeta_1(x_1)\cdot\zeta_2(x_2) \pm \zeta_1(x_2)\cdot\zeta_2(x_1)]$$

We see that interchanging electrons 1 and 2 leaves ψ unaltered in the case of the plus sign and reverses the sign of ψ in the case of the minus sign; however, reversing the sign of ψ is irrelevant because only its square has significance. It can be shown that the expression written for ψ is the only linear combination satisfying the equation of



which ζ_1 and ζ_2 are solutions and making the function invariant with respect to electron exchange. The function with the plus sign is referred to as "symmetrical"; that with the minus sign is an antisymmetrical function.

What happens if the two points x_1 and x_2 coincide; i.e., if $x_1 = x_2$? In the case of the symmetrical function $\psi = \zeta_1\zeta_2$; in the case of the antisymmetrical function the wave function ψ vanishes. Let us now choose the symmetrical function for the case of two electrons of opposite spin and the antisymmetrical function for the case of two electrons of the same spin. In doing so we achieve automatic fulfillment of Pauli's principle, because $\psi = 0$ (and consequently $\psi^2 = 0$) means that the probability of finding two electrons of the same spin in the same place is zero. We see now how wave mechanics of a two-electron system has been formulated so as to embrace the exclusion principle. At the same time it must be emphasized that wave mechanics goes beyond this principle. While the exclusion principle tells us what the electrons are forbidden to do, wave mechanics tells us what they are actually doing. Since its inception in 1926 wave mechanics has made an enormous contribution to progress in physics. One is reminded of a saying of Einstein: "God may be secretive, but He is not malicious." It may be difficult to gain access to a certain region of knowledge, but once we have pierced a small hole through the curtain, we are often permitted to see more than we hoped for.

Exchange energy

One of the most interesting and important consequences of wave mechanics (including indistinguishability of electrons and the exclusion principle) is the existence of a new kind of nondynamic force: exchange forces and exchange energies. Exchange forces, being responsible for the chemical bond, form the basis of theoretical chemistry, and similarly, they underlie the phenomena of ferromagnetism quite apart from the role that exchange energy plays in the interpretation of atomic and molecular spectra. It is quite possible to obtain, even without mathematics, a qualitative understanding of the origin of exchange forces in the following way. Let us visualize two hydrogen atoms, each with a positive nucleus (proton) and an electron cruising around the proton. If the two electrons are of opposite spin, the exclusion principle permits them to be close to each other. Not only are they permitted to be close together, but since they are represented by a symmetrical wave func-

tion, the probability of where we shall find them has a maximum for $x_1 = x_2$, i.e., the same point. It should be mentioned here that the notion of the orbit of an electron around the nucleus—as we often see it pictured on the covers of popular books on atomic physics, and even in advertisements for electronic devices—is obsolete today, although the term "orbit" is still used in a loose sense for convenience.

The wave function permits us to compute the probability of finding the electron at a particular point at any one time, but does not permit us to draw a continuous orbit along which the electron travels (this will be discussed at some length in the next section). While for a single hydrogen atom the probability of finding the electron has a maximum at some particular distance (0.53) Å from the nucleus but otherwise does not depend on the direction, it is different in the case of two hydrogen atoms. If the two electrons are of opposite spin, the probability has a maximum for both electrons at the mid-point between the two protons. We then have at the mid-point a double electronic (negative) charge attracting on each side a single positive proton charge. (The charge on the proton is of the same magnitude but opposite sign to that on the electron.) The double electron charge serves as a cement between the two protons and in this way the hydrogen molecule is formed. No molecule is formed if the two electrons have spin of the same sign. We also can easily see why a hydrogen molecule does not attract a third hydrogen atom. Since the two electrons in the molecule have opposite spins, the electron of a third atom would have the same spin as one of the two electrons of the molecule and, according to the exclusion principle, would avoid the vicinity of this electron, hence the lack of attraction.

The reasoning so far is of a purely qualitative nature. In particular we disregarded the electrostatic repulsion of the two electrons of opposite spin (the magnetic interaction is so small that it can be neglected for the time being). Two electrons of opposite spin like each other despite their mutual electrostatic repulsion, but this attractiveness cannot be described in such simple terms as Coulomb's electrostatic forces. Coulomb's law was established for charges and charged particles, and is valid in wave mechanics without any modification. The new "exchange forces" however can be applied to electrons only on the basis of the wave nature of the electron. Briefly it could be said that the exchange energy is due to the fact that generally $(A + B)^2$ is not equal to $A^2 + B^2$. If A and B stand for wave amplitudes, then—as we have seen before—$(A + B)^2$ is equal to $A^2 + B^2$, if we integrate over the whole space, because the integral of AB over the whole space cancels out. The physical meaning of $(A + B)^2 = A^2 + B^2$ is the conservation of energy. In wave mechanics, however, the wave intensity stands for charge density rather than for energy density. Energy is given, as we know, by elementary electrostatics, by the product of charge and electric potential V. In this case the integral of $(A + B)^2V$ will not be equal to the sum of the integrals of A^2V and B^2V, because the integral of ABV does *not* cancel out.

Let us consider as an example the two electrons in the helium atom. The simplest case is the ground state of the atom, where the two electrons are of opposite spin and therefore both can be in the same, the lowest energy, orbit. Yet this case is not interesting for us, because the two electrons, being in the same orbit, have the same wave function, and if the wave function of a single electron is $\zeta_1(x_1)$, the wave function ψ for two electrons will be simply $\psi = \zeta_1(x_1)\zeta_1(x_2)$. The next case of higher energy is where one electron is in the ground state, the other in an excited orbit. Here the two wave functions are different. $\zeta_1(x_1)$ and $\zeta_2(x_2)$ and the wave function of the two electrons ψ will be either a symmetrical or an antisymmetrical function, depending on whether the two electrons have opposite or the same spin:

$$\psi(x_1,x_2) = 0.71[\zeta_1(x_1)\zeta_2(x_2) \pm \zeta_1(x_2)\zeta_2(x_1)]$$

Let us consider the energy of the system. There is, first, the electrostatic energy of the nucleus, then the electrostatic energies of the two electrons taken separately, and, last, the interaction energy of the two electrons. This last term is of particular interest to us. To obtain the interaction energy we multiply the charge density (or rather its time average) by the electric potential and integrate the product over the whole space. The charge density is given by ψ^2, while the electric potential is equal to e^2/r_{12}, where e stands for the electronic charge and r_{12} for the mutual distance of the two electrons. Let us put, for short, $\psi = A + B$, then we shall see that the integral of $0.5(A^2 + B^2)\cdot(e^2/r_{12})$ over the whole space includes e^2/r_{12}, which is just the Coulomb interaction energy as in classical electrostatics. Apart from that, however, we have a term $AB\cdot e^2/r_{12}$ integrated over the whole space, which does not occur in the classical theory. This term is what is referred to as "exchange energy." Let us inspect the product AB more closely; that is,

$$AB = \pm \zeta_1(x_1)\zeta_2(x_2)\zeta_1(x_2)\zeta_2(x_1)$$

We see at once that this product has a finite value only if the orbits of the two electrons overlap to some extent. If there were no overlapping $\zeta_1(x_2)$ would be equal to zero and so would be $\zeta_2(x_1)$. It follows that exchange forces are only short-range forces, which explains why there is little chance of coming across exchange forces while observing free electrons. Then we see that the exchange energy is proportional to e^2/r_{12}, i.e., to the electrostatic energy, but can occur with a plus or minus sign. We cannot say then that exchange energy is a kind of extension of electrostatic energy. We have to reconcile ourselves to the fact that exchange energy is the outcome of the wave nature of the electron (more specifically, of the interpretation of wave intensity as charge density and not energy density). The term "exchange" energy indicates the indistinguishability of electrons and hence their exchange in the formulation of the wave function ψ as the prime source of exchange energy.

Uncertainty principle

There is something incongruous about the "wave nature of an electron." There is a story about an actor who was looking for a job and when asked by the producer about his past achievements, answered that he had played a wave (a sea wave produced by rattling of a mechanical contraption). Now we can imagine how an actor can play a wave, but how can a wave play the part of an electron? Let us take the simple case of a plane sinusoidal wave, extending from $-\infty$ to $+\infty$. Frequency and wavelength (or rather wave number) define energy and momentum of an electron accurately, but

otherwise this wave gives no information about the electron and, in particular, no information about the position of the electron. The question is sometimes asked as to whether an electron is "in reality" a particle or a wave. It is doubtful whether physics is able to answer this kind of question. The job of the physicist is to devise methods of enabling us to predict physical occurrences. For free electrons the particle aspect is very well suited to describe the phenomena (such as the electron tracks in a Wilson cloud chamber or in a photographic emulsion); for an electron in an atom the wave picture is better suited.

The electron is neither a particle nor a wave; both are only modes of description. But is it possible to bridge over the span between an infinite wave and a tiny particle? Fortunately, it is. Superposing a number of sinusoidal waves gives us a "wave packet," or—as the communications engineer would rather say—a "pulse." A Fourier analysis may give us the "spectral" components of the pulse and they may be all solutions of a given wave equation.

What is the difference between an infinite wave and a pulse? Of course, the pulse is confined in space and its propagation simulates the motion of a particle. The sharper the pulse the more it resembles a moving particle. Then why not represent an electron by an extremely sharp pulse rather than by an extended wave? The answer is that in representing the electron by a sharp pulse we have to pay a price for the "sharpness." We know the sharper the pulse the broader the spectrum in the Fourier analysis of the pulse. The spectrum of the pulse shows us the distribution of the wavelengths, or wave numbers, among the components of the pulse, and we know that the wave number stands in wave mechanics for the momentum of the electron (product of mass and velocity). This means that the sharper the pulse the larger will be the range of wave numbers from which we can choose the momentum of the electron. An infinite wave gives us precisely the wave number (electron momentum or velocity) but no position of the electron. An infinitely sharp pulse gives us exactly the position of the electron but no velocity. We cannot know exactly both the position and velocity of the electron; the product of the sharpness of the pulse and the sharpness of its spectrum is constant. If we denote the range of position by Δx and the range of velocity by Δv, then we shall have

$$\Delta x \cdot \Delta v = h/m$$

where h stands for Planck's constant and m for the mass of the particle. This is the celebrated "uncertainty principle" first put forward by Heisenberg.

It follows immediately from the uncertainty principle that we cannot imagine an electron being at rest. Being at rest means $v = 0$ and accordingly $\Delta v = 0$, but if Δv is zero then Δx must be infinite, and vice versa, if $\Delta x = 0$ then Δv must be infinite and accordingly $v = \infty$ too. This gives us another explanation as to why an electron cruising around the nucleus can never fall into it. Falling into the nucleus would give the electron a definite position and therefore an infinite velocity. The actual orbit is a compromise between the large value of energy the electron would have very near the nucleus and the large energy if completely detached.

The uncertainty principle tells us, too, why there is little chance to observe the spin of the free electron directly. The spinning electron represents a magnetic dipole whose field intensity is inversely proportional to the cube of the distance. According to the uncertainty principle, there is no hope of being able to observe the spin of an electron at rest. It must be moving. Yet a moving electron generates a magnetic field of its own around its path, whose intensity is given by the reciprocal of the square of distance. Obviously the field intensity due to the spin should be larger than the "uncertainty" of the field intensity due to the motion of the electron. If we write the formulas for both magnetic field intensities, this requirement leads to $\Delta r \cdot \Delta v < h/m$, yet according to the uncertainty principle this product cannot be smaller than h/m, hence it is impossible to observe the spin of the free electron directly. The spin of electrons bound in the atom is not observed directly either, but in the case of the atom the splitting of spectral lines in presence of a magnetic field gives a good clue as to the spin of the electrons responsible for the spectrum.

Another way to regard the uncertainty principle is to consider the way we observe the electron. To observe a macroscopic object, say, a chair, we let an image of the chair be formed on the retina of our eye. Of course, we can see the chair only if it is illuminated, but we know that the chair does not suffer in any way by the illumination. An electron is so small that we could not see it, either with the naked eye or with an ordinary microscope. In an imaginary experiment we could use an X-ray microscope, because the wavelength of the X rays could be chosen small enough to be reflected from the electron. Yet this would be a pretty hard X ray, of very high energy, and as a result the electron hit by the X ray would be dislocated. The very act of measurement would disturb the quantities to be measured. A similar phenomenon occurs in psychology: for instance, if we want to watch how our own mind behaves when we fall asleep, we either fail to make the observation or fail to fall asleep. Only to some approximation is the observation possible, and the uncertainty principle states quantitatively how near we can come to the assessment of the quantities to be measured.

The uncertainty principle applies not only to the electron but to larger particles as well. However, we notice that the right-hand side of the relation is equal to h/m, which means that the uncertainty is inversely proportional to the mass of the particle. For a proton it is about 2000 times smaller than for an electron and for a tiny (just perceptible) grain of sand it is practically zero. That is the reason why in ordinary mechanics it is not necessary to consider uncertainties.

A consequence of the uncertainty principle is the impossibility of drawing the "orbit" of an electron. We can ascertain the position of the electron as often as we want, but we could not be certain that the orbit of the electron would be the same if we did not disturb it by our measurements. Perhaps it may even be meaningless to ask what the orbit would be if we did not observe the electron. Nor can we be sure that in two successive observations we observe "the same" electron, since electrons are indistinguishable. All this is philosophically disquieting; it is the first time in the history of physics that we have a theory in which the observer and the observed enter on equal footing. In all other branches of physics the phenomena are described regardless of whether there is an observer or not. In wave mechanics the phenomenon observed is partly created by the act of observation.

The uncertainty principle gives us a new angle from

which to regard indistinguishability. Let us say that two identical twins are walking on the street. We cannot tell one from another, but we could label the one on the right-hand side Tom and the other Dick. They may change places, engage in a brawl, but as long as we keep an eye on them, we can tell which one is Tom and which is Dick. With two electrons it is different, because we cannot watch them continuously and as soon as we shut our eyes they may change their identities, if they have identities.

All this is not pure speculation; we can test these considerations in the following way. We bombard a sample of matter by negatively charged, high-energy particle beams to ascertain the maximum energy that can be transferred from the bombarding particle to an atom, or, rather, to an electron of an atom of the sample. For a given energy of the incident particle a formula can be derived for the maximum transferable energy as a function of the mass of the hitting particle. It is true that we cannot vary the mass of the incident particle at will, but since cosmic rays provide us with such a variety of fundamental particles of various masses, it is not quite unrealistic to consider this transferable energy as a function of the mass of the incident particle. Starting with infinite mass for which the value of the function is zero, the function increases with decreasing mass, but at the moment the mass of the incident particle equals the mass of the particle struck, i.e., when both are electrons, the maximum transferable energy drops suddenly to half the value indicated by the formula. This happens because we could not tell which of the two electrons was hitting and which was hit, which was the donor and which the acceptor of energy, hence the factor $1/2$. This result has been confirmed by experiment.

Shooting high-energy particles at a fluorescent screen makes the effect of a single particle observable. The light output in the form of a scintillation indicates the energy of the incident particle and at the same time we can observe the spot on the screen where the particle has impinged. This would be a contradiction to the uncertainty principle, so let us consider how we do observe the spot where the screen has been hit. We may observe the screen by means of a microscope in which case the limit of resolution is of the order of 2000 Å—about 100 million times the size of a fundamental particle. We see the light generated by the impact of a single particle, but to say we also see *where* the particle has hit the screen would be tantamount to saying: "I know exactly where Mr. Smith is at the moment; he is somewhere on the surface of the earth."

The relativistic electron

We began with the stated aim of getting acquainted with the electron on a personal footing, but this proved to be impossible, because a single electron has no personality, no individuality, no "sameness." If the fundamental particles have no individuality, but macroscopic objects have, they obviously owe their individuality not to the matter of which they consist, but to the pattern of their assembly. Nineteenth century physics seemed to confirm, or even encourage, the materialistic outlook in philosophy, which attributed to matter an overriding importance. This is no longer true. It was particularly the discovery of Rutherford of the nucleus of the atom that dealt a severe blow to the materialistic outlook. Rutherford found that the nucleus in which more than 99 per cent of the matter of the atom is concentrated occupies only a minute fraction of the total volume of the atom. The atom consists mainly of empty space, while the density of the nucleus is such that a matchbox full of nuclear substance would weigh millions of millions of pounds. This is not how the 19th century man imagined matter.

If we grip with our hand a piece of steel, we may feel that the matter of our palm comes into intimate contact with the matter of the steel. Actually the best we can hope to achieve is that the electron clouds of the atoms of the palm come into touch with the electron clouds of the atoms of the steel. Since matter is mainly concentrated in the nucleus, intimate contact of matter with matter should involve contact of nuclei. While this is not impossible, it is a rare occurrence, usually involving nuclear reactions with sometimes rather conspicuous consequences. In gripping a piece of steel with our hand, "common sense" deceives us into believing that we come to grips with matter. Actually, however, the rigidity and hardness of steel are due not directly to any hardness of the fundamental particles, but to the "fields" surrounding them. The fields surrounding the particles are in a way more real than the particles themselves.

We would like to conclude with a remark about what the theory of relativity has to say about the electron. One of the postulates of the theory of relativity concerns the velocity of propagation either of a particle or a field, namely, nothing can propagate with a velocity higher than the velocity of light. Let us now visualize an electron facing the approach of an electromagnetic wave. If the electron has a finite size, one end of it will be engulfed by the wave sooner than the other and will be acted upon earlier. This means that one end will start to move while the other still knows nothing about the wave. As a result the electron should become deformed, but this is impossible for a fundamental particle. If we start imagining that a fundamental particle can be deformed, we might as well assume that a wave of sufficient strength would tear the electron apart, which would be a contradiction to the concept of an ultimate particle. We cannot assume that the electron is absolutely rigid, because that would mean that any impulse reaching the electron at one end travels to the other end with infinite velocity, yet according to the theory of relativity there is nothing that travels with a velocity higher than the velocity of light. The only way out is to assume that the electron has no spatial extension at all, in other words, that the electron is nothing but a geometrical point in which charge e and mass m are concentrated with infinite density.

The electrostatic energy of a sphere of charge e is proportional to e^2/r, where r stands for the radius of the sphere, and accordingly the energy of a pointlike electron is infinite. This is of course a great difficulty for the theory, although physicists have learned somehow to live with those infinities.

We started by pointing out that the concept of continuous matter involves infinities that can be avoided by assuming finite, ultimate, indivisible particles. We have now come all the way around the circle and are once again facing infinities, even while working with ultimate particles. It seems that we face infinities whenever we arrive at frontiers of knowledge, or perhaps at the frontiers of our potentiality for knowledge.

Progress in optical computer research

Lasers are considered basic components for optical computer logic. The use of optical rather than electric signals in digital computer circuits may offer important advantages for future information-processing systems

Oskar A. Reimann *Rome Air Development Center*
Walter F. Kosonocky *RCA Laboratories*

All-optical computer techniques

Oskar A. Reimann

High-speed electronic computer circuitry is becoming interconnection limited. The reactance associated with the mounting and interconnections of the devices, rather than the response of the active components, is becoming the main factor limiting the speed of operation of the circuits.[1] A possible approach to computer development that might circumvent interconnection limitations is the use of optical digital devices rather than electronic devices as active components.

Although it is premature to give conclusive results on the basis of research conducted so far, several points can be mentioned:

1. In optical transmission lines, the wavelength of the signals will be much shorter than any of the circuit dimensions; thus all of the reactive effects present in the interconnections could be eliminated. One might still have to be content with mismatched transmission lines, but this is not a major problem since the decay times are very short.

2. The possibility of signal connection between parts of the system without electrical or actual physical contacts is very attractive for integrated-circuit techniques. With optical signals, a totally new approach to the interconnection of digital devices is possible.

3. Laser devices show promise of very fast switching speeds. Together with optical interconnections, these devices could provide digital circuits that are considerably faster than electronic circuits.

4. Finally, new circuit techniques may give additional freedom to the computer engineer and may increase the computing power of future information-processing systems.

All-optical and optoelectronic approaches

At this point, distinction should be made between the "all-optical" laser digital devices and optoelectronic circuits. The operation of laser digital devices is based on the interaction of optical signals with laser materials; only optical signals are used as the inputs and the outputs. Therefore, laser digital circuits should be capable of taking full advantage of the very fast switching speed of laser materials in response to intense optical signals.

Optoelectronic circuits, on the other hand, require conversions between optical and electric energies. Such circuits are already established in the digital technology for very special applications. A typical example is the case wherein the transfer of signals without mechanical connection or with perfect electrical isolation is of great importance. Optoelectronic digital circuits using electroluminescent materials and photoconductors have a long history in their application to digital logic. Present components provide a minimum switching time of approxi-

mately 50 ms, with a power consumption (per element) of approximately 10^{-5} watt. This type of device is very slow and therefore has only a limited application in digital logic. It is useful in image processing and character recognition. With the advent of the laser, of efficient light-emitting diodes, and of high-speed photodetectors, interest in the application of higher-speed optoelectronic circuits to digital logic has increased. The work of Biard[2] in applying optoelectronic circuits to integrated electronics is representative of this type of effort. In general, however, optoelectronic circuits cannot compete in performance with their electronic counterparts. The high losses associated with converting the signal between optical and electric energies require high-gain electronic amplification, which slows down the operation of these circuits. This situation could be changed if the amplification were provided by a laser amplifier. Therefore, we may in the future expect to see optoelectronic circuits that will combine laser amplifiers with other high-speed semiconductor devices.

Several years ago Rome Air Development Center undertook an investigation of various optical effects that might lead to optical digital devices for use in an all-optical computer. The application of laser phenomena to all-optical processing of information was chosen as a long-range goal. Although this goal is still rather ambitious, the possible payoff for success is great enough to warrant serious effort on this research.

As part of this investigation, American Optical Company undertook to explore for RADC the use of optical fibers for performing digital functions. Initial investigations were concerned with more or less classical optical effects, among them the Faraday and Kerr effects. With the development of the glass-fiber laser by Elias Snitzer of American Optical Company, research work turned to the exploration of the use of laser fibers for performing digital functions. At approximately the same time, the neuristor laser computer was conceived at RCA, to whom a contract was awarded for feasibility studies of a laser computer employing neuristor laser components.

Fiber-laser studies

The attempt to use the Faraday or Kerr effect to make glass-fiber light switches gave marginal results.[3] Both effects are unattractive for computer application because they involve a high magnetic or electric field. Also, both effects require optical-to-electrical conversion. The experimental results also showed that the Faraday effect in glass is small at room temperature. The material with the greatest known Kerr constant is a liquid, which in itself is not very appealing for computer applications. Subsequently, the laser quenching effect was first demonstrated by C. J. Koester with neodymium-doped glass lasers.[3] This experimental result was significant, because it demonstrated that one laser signal can affect the output of another laser. In the quenching effect, the quenching beam robs energy from the quenched laser and thus is amplified, while the population inversion in the quenched laser drops below that necessary for laser oscillation. On the basis of an analysis of the laser quenching effect it was concluded that, in principle, the effect can be fast and exhibit gain.[4] The effect was also shown by Fowler of IBM[5] using GaAs (gallium arsenide) lasers, and, more recently, efficient GaAs laser quenching has been achieved at RCA Laboratories with a dual laser oscillator, as described in the second part of this article.

The work at American Optical Company also included a study of pulse propagation in fiber lasers and of resonant coupling between two fiber lasers in the same cladding. Amplifications of low-level signal pulses of the order of 10^4 were exhibited in a one-meter fiber laser. High input signals exhibited saturated amplification in which the first pulse is amplified more than the second and the first pulse approaches a steady-state pulse with varying input signals. The results of the fiber-laser-coupling experiments showed that two fiber lasers with a diameter of 5 microns and length of 10 cm in the same cladding and separated by approximately 5 microns exhibited very strong coupling (time coincidence). On the other hand, similar fibers but with an 11-micron separation exhibited no coupling.[4,6]

The neuristor laser computer

In the neuristor laser computer, conceived at RCA, all information and control signals are in the form of optical energy. Fiber-optic elements, with appropriate concentrations of active emissive ions and passive absorptive ions, are the basic components of this system. The computer is powered by being in a continuous light environment that provides a constant pump power for maintaining an inverted population of the emissive ions. Among the potentially attractive features of such a system are the freedom from power-supply connections for individual circuits, the possibility of transmission of signals without actual connections between certain locations, and a promise of high-speed operation.

A theoretical study[7,8] of the neuristor concept in the form of fiber-glass lasers showed that the fundamental requirements of a neuristor line could, at least in principle, be met with lasers. A laser traveling-wave transmission line that has a linear scattering-loss mechanism possesses inherent stability, as evidenced by the existence of steady-state pulses that propagate down the line with a constant velocity. If the line includes saturable absorber ions (in addition to the emissive ions) and linear losses, a line that propagates steady-state pulses can be obtained by a proper choice of system parameters. The line may possess a stimulation threshold such that weak incident pulses are attenuated while strong incident pulses develop into steady-state pulses characteristic only of the line itself. Following the steady-state pulse, the laser neuristor line has a refractory period or recovery time. The duration of this period depends directly on the recovery times of the ions of the emitter and the saturable absorber. In view of the present emphasis on saturable absorber material for Q-switched lasers[9–11] it is reasonable to assume that a saturable absorber will be developed for the neodymium glass laser.

The major limitation of the laser neuristor is the pump power requirement. It appears that the shortest recovery time that may be expected within the foreseeable future is of the order of 1 to 100 microseconds. Thus, optically pumped lasers could be used for digital operations only at kilocycle repetition rates and only in form of resonators and not as continuous transmission lines. Open-loop or closed-loop fiber-glass lasers could form such laser-resonator digital devices. It also appears that it should be more efficient to use these laser resonators in a bistable or

monostable circuit mode of operation rather than in the form of neuristor transmission lines.

The main result of the laser neuristor feasibility study was the conclusion that lasers are capable of satisfying all the requirements for digital devices. It was shown that, in addition to the neuristor-type logic, lasers in the form of resonators and amplifiers can have input–output characteristics that resemble those of conventional logic circuits, such as gates or flip-flops. The concept of saturable optical absorber material that can provide a threshold function for laser digital devices was suggested early in the program.[7] Saturable absorption at optical frequency was first demonstrated with an unpumped ruby crystal.[8] Spectroscopic tests of saturable absorber materials were then expanded to the studies of solutions of phthalocyanine, as described in the second part of this article.

A significant contribution of the laser neuristor feasibility study to the laser field is Wittke's analysis of pulse propagation in a laser transmission line without[12] and with[11] a saturable absorber material. This was the first correct formulation of signal-pulse shaping by an infinitely long laser amplifier. The steady-state pulse developed by such an amplifier tends to leave behind a transparent line. The steady-state-pulse duration is a function only of the loss-to-gain ratio and, for a large range of this ratio, the pulse duration T is of the order of the reciprocal of the line width of the laser transition. Thus, if a high-gain laser amplifier is used for pulse forming in laser digital circuits, most laser materials at room temperature would have to be operated with pulses of the order of 10^{-11} second in duration to exhibit a steady-state pulse.

Once the basic limitation of low repetition rates with optically pumped lasers was realized, interest shifted toward semiconductor lasers as components for laser digital devices. The basic performance characteristics will now be considered, and examples will be presented to show how semiconductor laser oscillators and amplifiers can be used as laser digital devices.

Laser digital devices

Walter F. Kosonocky

Laser digital devices may be used for general-purpose logic circuits in very much the same way that transistors are now used, except that all of the processing is done with optical rather than electric signals. Operation is based on nonlinear (saturable) interaction of intense optical signals with laser materials. The two basic nonlinear processes are (1) the quenching of gain in a laser and (2) saturation of optical absorption.

Large-signal response of laser materials

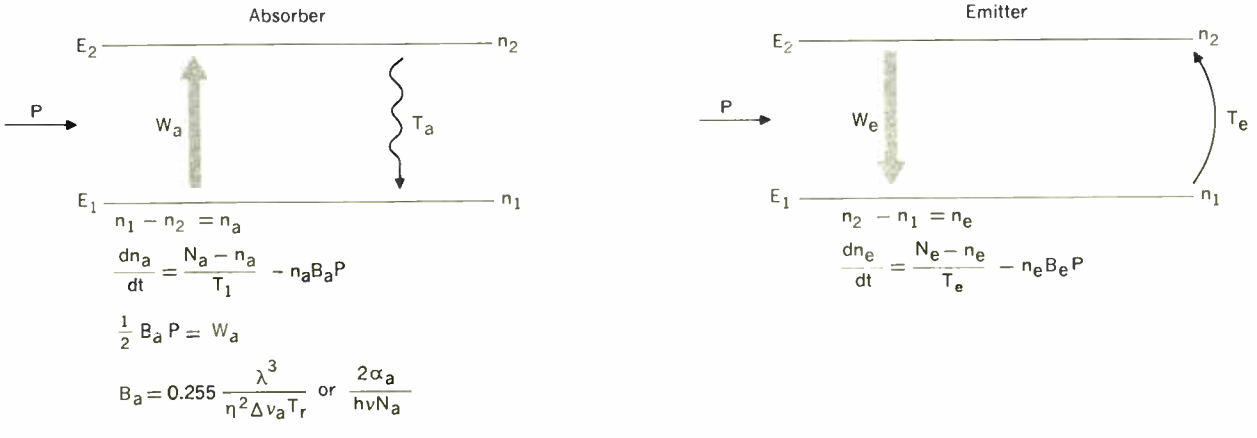
The interaction of an electromagnetic wave with a "saturable" absorber material (the absorber) or a continuously pumped laser material (the emitter) can be described by a two-energy-state model.[7,8] In Fig. 1 the stimulated transitions are represented by a transition rate $w = \frac{1}{2}BP$, where B is an interaction constant and P is the signal flux-power density in watts/cm^2. For a two-level absorber this interaction constant is equivalent to the absorption cross section σ; that is, $B = 2\sigma = 2\alpha/h\nu N$, where α is the absorption coefficient in cm^{-1}, N is the concentration of the ground state E_1 in ions/cm^3 when $P = 0$, and $h\nu$ is in joules. The interaction constant can also be expressed in terms of homogeneously broadened line width $\Delta\nu$ and spontaneous-emission time constant T_r by

$$B = 0.255 \frac{\lambda^3}{\eta^2 \Delta\nu T_r} \qquad (1)$$

where wavelength λ is in microns; η is the index of refraction; $\Delta\nu$ is in wave numbers, cm^{-1}; and T_r is in seconds.

The spontaneous transitions of the absorber toward the ground state are represented by an effective recovery time T_a. The effective recovery time of the emitter T_e represents the effective pumping rate of the laser material. Equivalent two-energy-state representations of three-level and four-level lasers are shown in Fig. 2. The responses of both materials to an applied optical signal P can be ex-

Fig. 1. Model for the interaction of an electromagnetic wave with a two-energy-state material.



pressed by similar differential equations

$$\frac{dn}{dt} = \frac{N - n}{T_1} - nBP \tag{2}$$

where the symbol n is used to represent a population difference between the two energy states; in the case of the absorber $n = n_a = n_1 - n_2$, and in the case of the emitter $n = n_e = n_2 - n_1$.

Let us consider two laser signals incident on a two-energy-state absorber material. The absorption spectrum of this material matches the emission spectrum of the emitter material that produces the input laser signals; see Fig. 3. The absorber material can be characterized by a concentration of absorptive ions per cm^3 N_a, a lifetime of the excited state T_a, and an interaction constant B_a. This constant relates stimulation transitions probability per unit time w_a to the flux-power density P of the optical signal at the resonant frequency $\nu_0 = (E_2 - E_1)/h$. (The absorber will be referred to as saturable if it can be saturated by the signal intensity that is produced by the laser source under consideration.) The effect on the applied signal due to interaction with the saturable absorber can best be shown if signal S_{A1} in Fig. 3 is assumed to have a much smaller intensity than signal S_{B1}. In addition, for reasons of geometry, S_{B1} is not as strongly affected by the absorber. Under these conditions, the following equations apply:

$$\frac{dn_a}{dt} = \frac{N_a - n_a}{T_a} - n_a B_a P \tag{3}$$

and

$$\frac{dP_A}{dx} = -\frac{1}{2} h\nu n_a B_a P_A \tag{4}$$

where $P = P_A + P_B$.

The absorption coefficient α_A can be defined as

$$\alpha_A = -\frac{dP_A}{P_A\, dx} = \frac{1}{2} h\nu n_a B_a \tag{5}$$

The steady-state solution of (3) is

$$n_a = \frac{N_a}{1 + B_a T_a P_B} \tag{6}$$

Therefore, the steady-state absorption coefficient is

$$\alpha_A = \frac{\alpha_A(0)}{1 + B_a T_a P_B} \tag{7}$$

where $\alpha_A(0) = \frac{1}{2} h\nu N_a B_a$ is the low-signal absorption

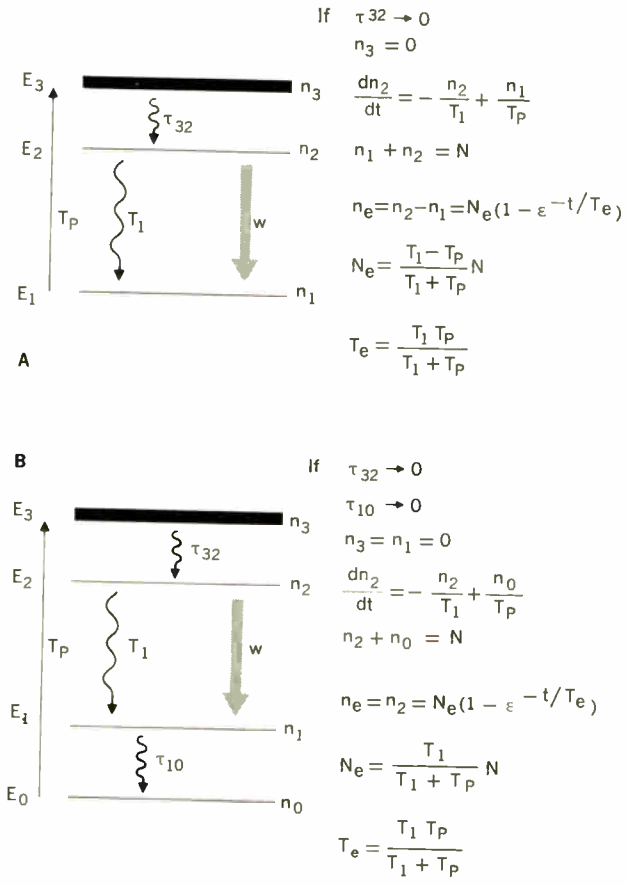
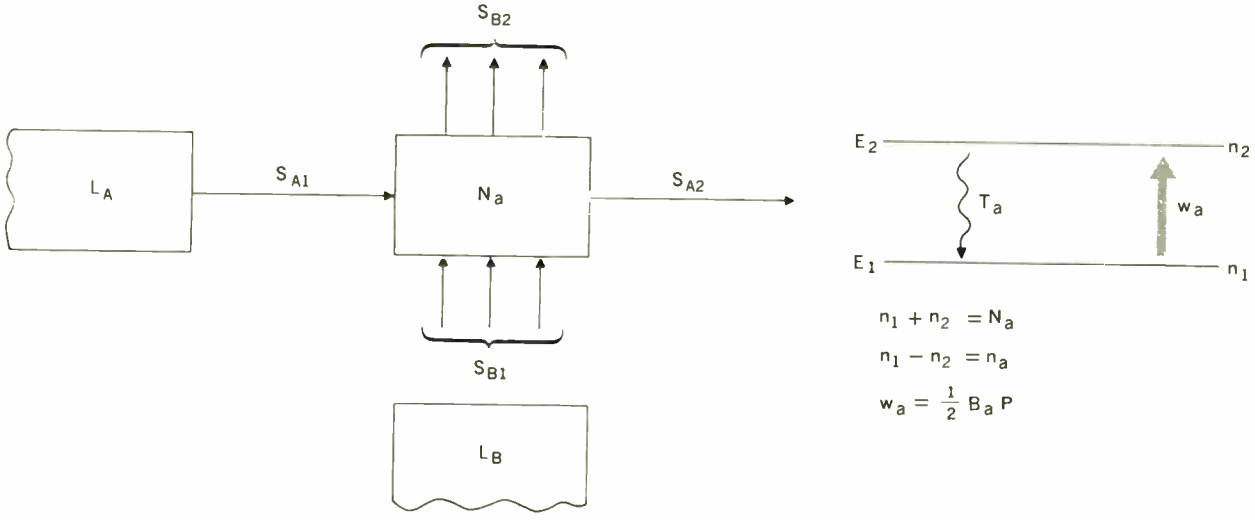


Fig. 2. Equivalent two-energy-state representations for (A) three-level and (B) four-level lasers.

Fig. 3. Schematic diagram for test of saturable absorber.



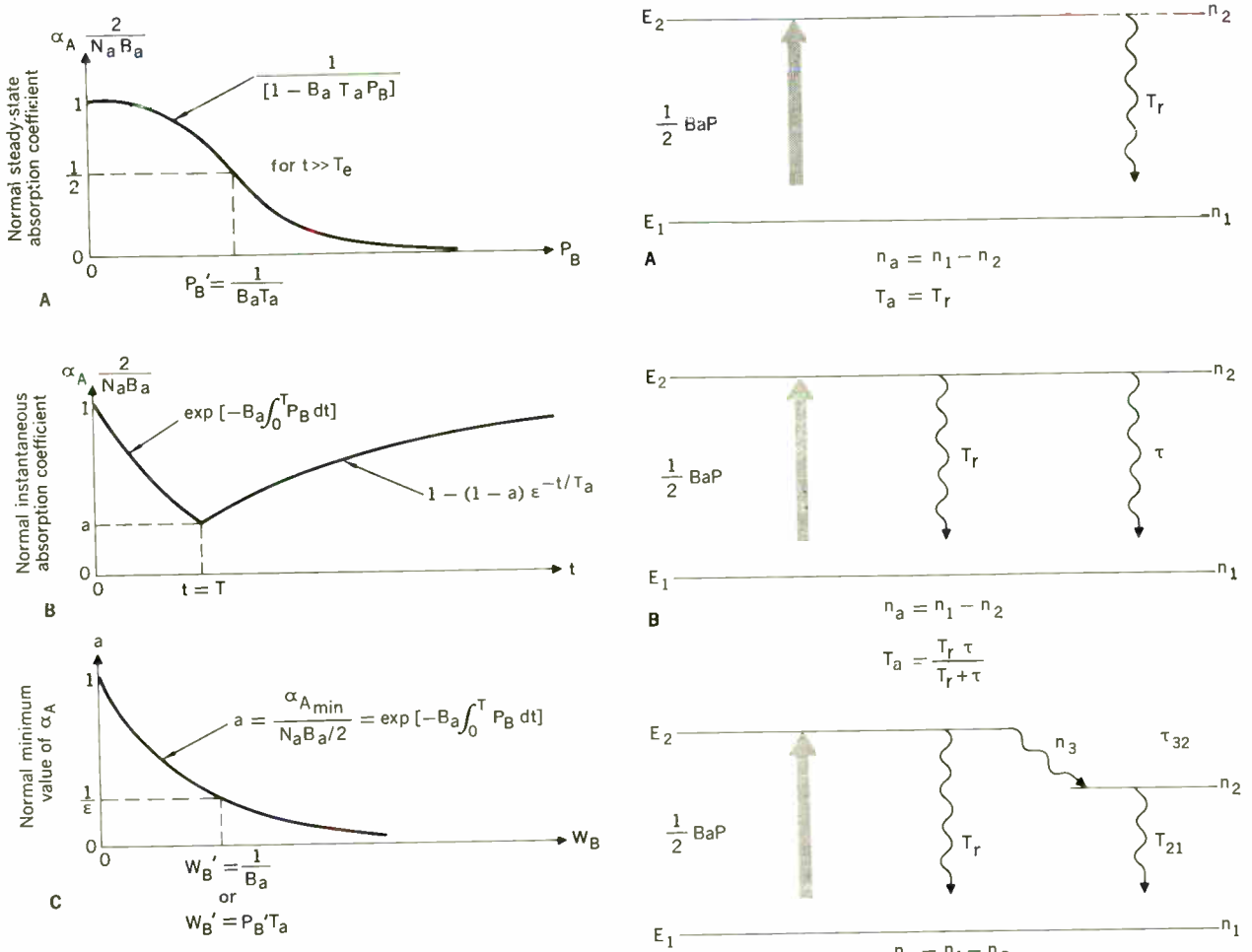


Fig. 4. Response of the saturable absorber in Fig. 3. A—Normalized steady-state absorption coefficient α_A as a function of flux-power intensity P_B of signal S_B. P_B is plotted on logarithmic scale. B—Normalized instantaneous absorption coefficient α_A as a function of time for pulse signal S_B with duration T and energy W_B. C—Normalized minimum value (a) of α_A as a function of W_B.

Fig. 5. Representation of saturable absorber. A—Two-energy-level absorber with radiative lifetime. B—Two-energy-level absorber with nonradiative recovery time. C—Three-energy-level absorber.

coefficient. The curve of α_A vs. P_B is sketched in Fig. 4(A).

Equations (6) and (7) apply for steady-state modulation of the absorption coefficients by signals with durations that are much longer than the recovery time ($t \gg T_a$). The transient effect on α_A when the signal S_B is in the form of a pulse with a duration $T \ll T_a$ can be expressed simply, from (2), by

$$n_a = N_a \exp \left[-B_a \int_0^T P_B\, dt \right] \tag{8}$$

and

$$\alpha_A = \alpha_A(0) \exp [-B_a W_B] \tag{9}$$

where

$$W_B = \int_0^T P_B\, dt$$

is the energy density in joules/cm² of the saturating pulse. The curve of α_A as a function of time for the transient response is shown in Fig. 4(B). In Fig. 4(C) the normalized minimum value of α_A is shown as a function of the energy W_B of the signal S_B. When the signal S_B is removed, the material has an exponential recovery with the time constant T_a.

Three types of saturable absorbers can be considered (see Fig. 5). Figure 5(A) shows a two-energy-level absorber with a radiative recovery time $T_a = T_r$. In Fig. 5(B) is shown a two-energy-level absorber that has a recovery time T_a, which is composed of both a radiative time constant T_r and a nonradiative time constant τ. The absorber in Fig. 5(B) requires higher signal intensities to reach steady-state saturation than does the absorber in Fig. 5(A) if both have the same interaction constant B_a and the same radiative lifetime T_r. The three-energy-level absorber in Fig. 5(C), on the other hand, can reach steady-state saturation at much lower signal intensity.

In these three types of absorbers, steady-state saturation is reached—that is, $\alpha_A = \frac{1}{2}\alpha_A(0)$—when $P_B' = 1/BT_a$ or, using Eq. (1), when

$$P_B' \approx \frac{4\eta^2 \Delta\nu}{\lambda^3} \left(\frac{T_r}{T_a} \right) \tag{10}$$

In general, steady-state saturation begins when the rate due to stimulated transitions becomes comparable to the recovery rate associated with spontaneous transitions. In the two-level absorber, T_a/T_r is the fluorescent quantum efficiency. In other words, in the case of the two-level absorber with a purely radiative recovery ($T_a = T_r$), the signal-flux-power spectral density (in watts/cm^2/cycle) must exceed the flux-power density associated with the so-called zero-point fluctuations of all possible radiation modes. According to (10), the signal-flux-power spectral density must be

$$\frac{P'}{\Delta\nu} = \frac{4\eta^2}{\lambda^3} \text{ watts/cm}^2/\text{cm}^{-1} \qquad (11)$$

or

$$\frac{F'}{\Delta\nu} = \frac{2 \times 10^{19}\eta^2}{\lambda^2} \text{ photons/s/cm}^2/\text{cm}^{-1} \qquad (12)$$

These relationships show that it would be relatively easy to reach saturation in resonant systems at both microwave and lower frequencies. It is also very easy to reach saturation with gas lasers, because of their very narrow line widths. In solid-state lasers, however, the line widths are usually broadened as the result of coupling with the phonon spectrum. For a line width $\Delta\nu$ of the order of 10 cm^{-1}, for instance, a signal-flux-power density of the order of one kilowatt/cm^2 would be required to produce a pronounced saturation effect.

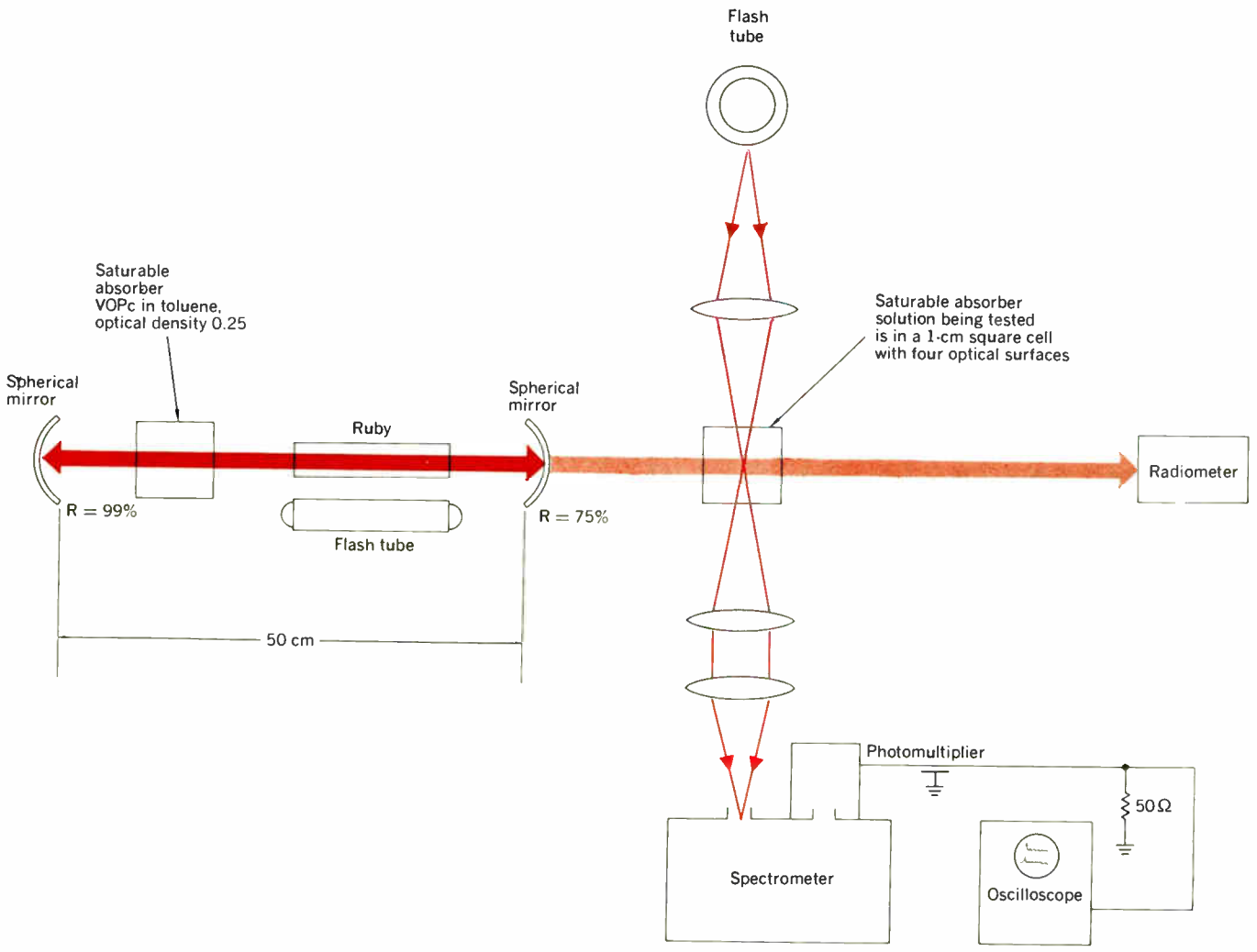
The energy density $W_B' = 1/B_a$ required for transient saturation—that is, for $\alpha_A = \alpha_A(0)/\epsilon$ when $T \ll T_a$—depends only on the interaction constant B_a. According to Eq. (1), W_B' should depend only on the homogeneously broadened line width $\Delta\nu$ and radiative lifetime T_r; that is,

$$W_B' = \frac{4\eta^2}{\lambda^3}\Delta\nu T_r \text{ joules/cm} \qquad (13)$$

The developed concepts for the saturable absorption were verified by spectroscopic measurements during illuminations by outputs of ruby lasers[13] made on an unpumped ruby crystal and various solutions of phthalocyanine.

The results of the tests of saturation of absorption and of spontaneous emission made on a ¼-inch ruby cube placed inside the resonant cavity of a ruby laser showed that the R_1 and R_2 lines of ruby behave at room temperature as a single, homogeneously broadened band. Both lines can be saturated by a laser pulse (at 6941.5 Å) having a duration of 0.2 ms according to a single expo-

Fig. 6. Equipment used for study of saturation of emission and absorption spectrum of solutions of phthalocyanine.



nential function. A ruby laser beam with energy density of 4 joules/cm² reduces the absorption in a ruby crystal to 37 per cent of the original, small signal value. The saturation can be described by the solution of a simple rate equation using an interaction constant determined from the value of the absorption coefficient. It was shown that at room temperature the model is valid down to nanosecond time durations.

As a test of the mechanism of saturable absorption, spectroscopic measurements were made with solutions of phthalocyanine[9] illuminated by giant ruby laser pulses. Q-switched ruby lasers with saturable absorbers such as VOPc (vanadyl phthalocyanine) in toluene in their cavities were used as the saturating sources. The equipment used for this experiment is shown in Fig. 6. Typical waveforms obtained in the dual-beam absorption tests are illustrated in Fig. 7. An example of a curve for saturation of absorption of VOPc in toluene is shown in Fig. 8.

The foregoing tests point out that the response to very intense optical signals can be explained in terms of a three-level-absorber model. This three-level-absorber response has been clearly evidenced for the metal-free solutions, which have a recovery time (toward the ground state) of about one microsecond, as shown in Fig. 7(B). In the metal solutions, on the other hand, the stronger spin-orbit coupling, as compared to the metal-free solutions, should give shorter, nonradiative transitions. The result is a more efficient three-level system in addition to a shorter recovery time, as shown in Fig. 7(C).

Although the results of the tests of saturation of absorption of the phthalocyanine solutions are not as straightforward as tests of an unpumped ruby, they are consistent with the theoretical analysis of saturable absorbers.

Semiconductor lasers for digital devices

The operation of laser digital devices, such as the laser neuristors[7] or the laser resonator switching devices to be described, is based on a large signal response of emissive and saturable absorber materials. A saturable absorber, to be useful as a laser switching device, must be easy to saturate by the optical signals developed by the laser. It follows that the minimum energy density W (in joules/cm²) that must be developed within the active region of the laser device to saturate the emitter material should be comparable to the energy density required to cause one switching operation of the device. It turns out that this energy density $W = PT_1$ is about the same whether the switching time is much shorter than or comparable to effective recovery time T_1. The value for this energy density is given by (13). It should be noted that it depends only on the homogeneously broadened line width $\Delta\nu$ and the radiative lifetime T_r, which is inversely proportional to the dipole moment or the oscillator strength of the transition. It is apparent, then, that for very-high-speed switching devices materials with very strong transitions or very short radiative lifetimes are desirable. Current-injection semiconductor lasers employ such materials. The energy density required to cause one switching operation in optically pumped lasers such as the ruby or the neodymium laser should be roughly six orders of magnitude higher than for GaAs lasers. The other favorable attribute of semiconductor lasers is that their pumping efficiency is at least one or two orders of magnitude higher than that of the optically pumped lasers.

There is at present no verified theory available to describe the details of the radiative recombination processes in GaAs lasers.[14] In general, however, laser action in GaAs diodes is explained by radiative transitions between overlapping degenerate n and p regions at the junction. Emission takes place from filled states in the conduction band to empty states in the valence band at a frequency corresponding to the emission energy E_e. The material in the active region is relatively transparent, since its absorption edge E_a corresponds to a higher

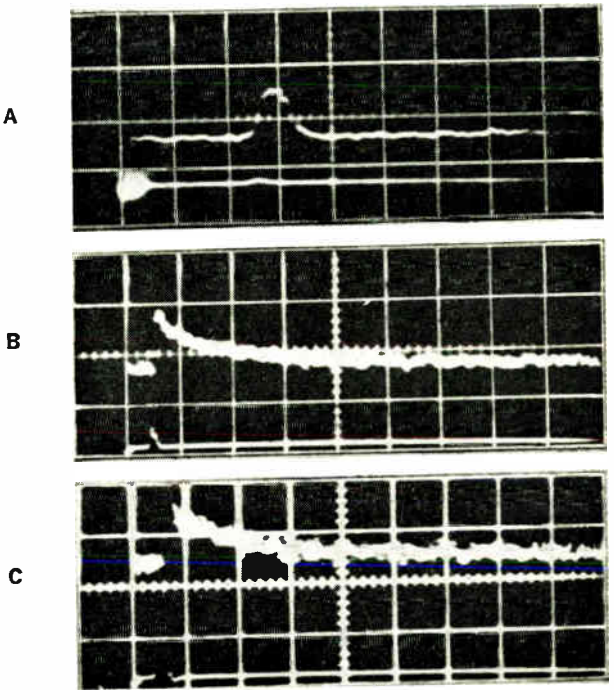
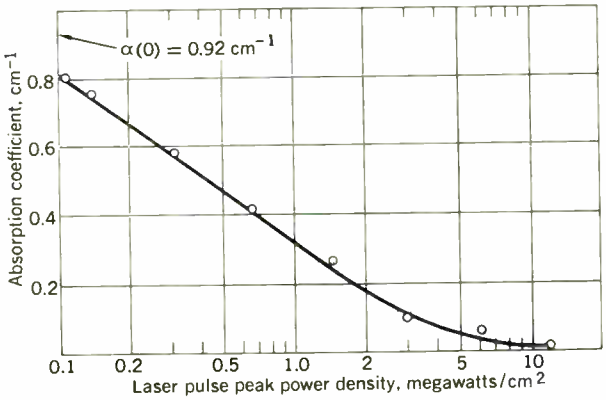


Fig. 7. Signal waveforms obtained in the test of saturation of absorption. A—VOPc in toluene at 6910 Å, low-signal optical density of 0.3. B—H_2Pc in chloronaphthalene at 6640 Å, low-signal optical density of 0.2. C—H_2Pc in chloronaphthalene at 6990 Å, low-signal optical density of 0.224. For upper traces, laser and flash tube were fired; for lower traces, only laser was fired. Vertical scale: 0.2 volt/div; horizontal scale: 0.1 μs/div for A, 0.5 μs/div for B and C.

Fig. 8. Saturation of absorption of VOPc in toluene at 6910 Å. Low-signal absorption coefficient $\alpha(0) = 0.92\ cm^{-1}$.



frequency; that is, $E_a > E_c$. The following simple phenomenological model is convenient for the prediction of large-optical-signal response of continuously pumped GaAs lasers.

Assuming that the degenerate valence band is essentially empty, the operation of GaAs can be interpreted in terms of a four-level laser, as shown in Fig. 9(A). The laser action takes place between energy states E_2 and E_1. The relaxation times τ_{32} and τ_{10} correspond to the diffusion time of the majority carriers, which can be considered equal to the carrier scattering time (of the order of 10^{-13} to 10^{-12} second). The fictitious energy states E_0 and E_3, in combination with the relaxation times τ_{32} and τ_{10}, represent the most conductive regions of the diode. Both E_1 and E_2 are impurity "tails," which are not really separate from E_3 and E_0, respectively. The external source supplies a constant current I for the population inversion for the active region. The ground energy state E_0 can be thought of as an infinite source of carriers from which the external current source transfers carriers to energy state E_3 at a rate I, in electrons per second, assuming that the process is 100 per cent efficient. These carriers will repopulate energy state E_2 with a carrier scattering time constant τ_{32}. The time constant T_1 represents the spontaneous emission time of n_2. It is further assumed that all of the recombination takes place in the idealized active region, which is situated at the junction and which contains a uniform distribution of the inverted population, $n_2 - n_1$. Assuming further that the laser state E_1 is normally empty, any population n_1 will be transferred to the ground state E_0 with a carrier scattering time constant τ_{10}. Thus, n_2 is the inverted population and n_1 is equal to zero.

If the radiative recombination is in the form of band-to-band recombination of free carriers, it should be proportional to a product of these two carriers.[15-17] There is strong evidence, however, that the emission in these diodes results from a recombination in the p regions involving injected electrons and acceptor (zinc) centers or an equivalent but large number of free holes whose density is not greatly affected by the bias voltage.[18-21] The rate of change of the inverted population n_2 can then be expressed by

$$\frac{dn_2}{dt} = -\frac{n_2}{T_1} - n_2 w_{21} + I \tag{14}$$

where w_{21} represents the stimulated transition probability per unit time resulting from a signal present in the active region. By letting $I = N_2/T_1$, Eq. (14) will assume the familiar form

$$\frac{dn_2}{dt} = \frac{N_2 - n_2}{T_1} - n_2 w_{21} \tag{15}$$

If we consider the relation between spontaneous and stimulated emission, w_{21} can be expressed using Eq. (1) by

$$w_{21} = \frac{\lambda^3 P}{8\eta^2 \Delta\nu T_1} \tag{16}$$

where T_1 is the radiative lifetime in seconds, P is the flux power intensity in watts/cm^2, $\Delta\nu$ is the homogeneously broadened line width in cm^{-1}, λ is the wavelength in microns, and η is the refractive index. Then

$$w_{21} = \frac{P}{P' T_1} \tag{17}$$

For $\eta = 3.6$ and $\lambda = 0.84$ micron,

$$P' = 155\ \Delta\nu \text{ watts/cm}^2$$

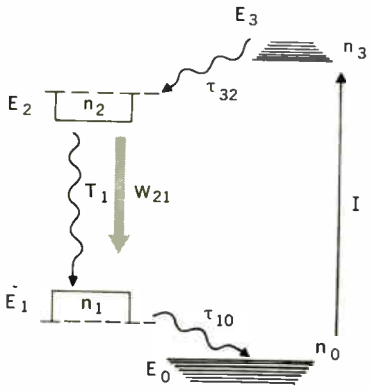
From (15), we have

$$\frac{dn_2}{dt} = \frac{N_2 - n_2}{T_1} - \frac{n_2 P}{P' T_1} \tag{18}$$

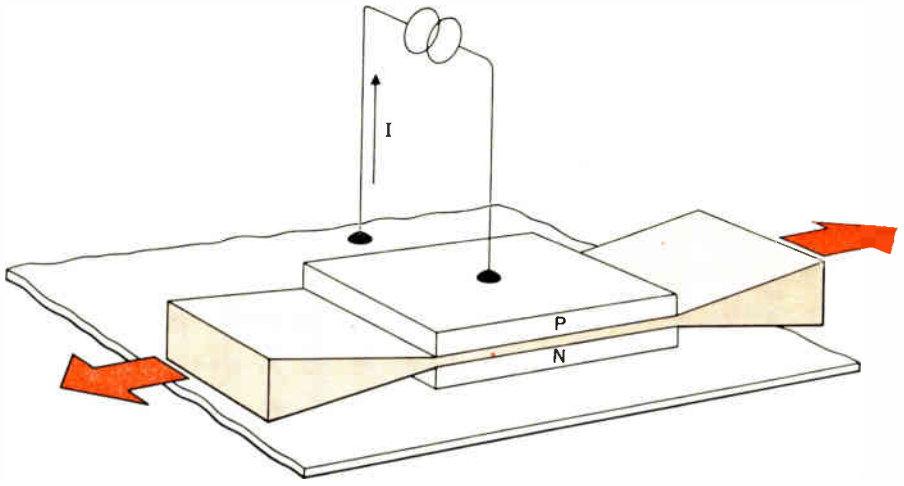
where $N_2 = IT_1$, with I in electrons/s/cm^3. P' is the flux power density at which the rate of stimulated emission equals the rate of spontaneous emission.

In terms of the band-filling model,[19,21] n_2 represents only the inverted population that corresponds to the homogeneously broadened part of the emission band. In other words, n_2 is the part of the inverted population that for particular doping levels, forward bias current, and temperature contributes directly to the stimulated emission. The homogeneously broadened line width $\Delta\nu$ then

Fig. 9. A—Four-energy-level representation of GaAs laser. B—Simplified sketch of GaAs laser



A



B

could be estimated to be equivalent to kT, where k is the Boltzmann constant and T is absolute temperature. At liquid nitrogen temperature, 78°K, $\Delta\nu$ would be equal to about 30 cm^{-1} and the flux-power density P' would then be of the order of 5×10^3 watts/cm^2. This estimate is consistent with experimental results showing that as the current of a GaAs diode is increased above the laser threshold the stimulated emission dominates the recombination process by maintaining the inverted population n_2 at a fixed level. The accompanying "lifetime-shortening effect" can be expressed conveniently by defining $T_s = 1/w_{21}$ as the lifetime of the injected carriers in the active region due to stimulated emission recombination. According to (14),

$$\frac{dn_2}{dt} = -\frac{n_2}{T_1} - \frac{n_2}{T_s} + I \qquad (19)$$

or

$$\frac{dn_2}{dt} = \frac{N_2^* - n_2}{T_1^*} \qquad (20)$$

where

$$T_1^* = \frac{T_1 T_s}{T_1 + T_s} \quad \text{and} \quad N_2^* = IT_1^*$$

This lifetime-shortening effect could be caused by superradiance, the amplified stimulated emission, or by externally applied signals. In the operation of continuously pumped laser digital devices, the time constant T_1^* of (20) determines the upper limit on the switching times. In GaAs lasers this time can be well under 1 ns.[22] Although continuously pumped room-temperature operation would be desirable, at the present state of development of GaAs lasers only liquid nitrogen CW operation is easy to achieve.[23]

Laser inverter circuit

The quenching of oscillations of one laser oscillator by the output of another laser oscillator[24,25,5] suggests the possibility of a laser inverter NOR circuit that could be used as a basic building block for optical computers. Let us consider a continuously pumped laser oscillator, as shown in Fig. 10. The laser material is described by an effective emissive population N_e, a recovery time T_e representing the effective pumping rate, and an interaction constant B_e representing the stimulated transition probability w_e, where $w_e = B_e P$. The total flux power P consists of two parts: (1) a flux power P_r, developed within the laser oscillator as the result of resonant modes that produce an output signal P_o, where $P_o = \alpha_o P_r$; and (2) a flux power P_i that passes through the emissive population N_e as a result of an applied input signal $P_{in} = \alpha_{in} P_i$, which is not directly coupled into the resonant modes producing the output P_o. The linear loss produced within the resonator due to P_r will be designated as $\alpha_r P_r$, and the total loss will be designated as αP_r, where $\alpha = \alpha_r + \alpha_o$ and $P_o = \alpha_o P_r$.

The simplest way to describe the input–output relation for the circuit in Fig. 10 is to use Eq. (7) for the variation of the effective gain coefficient α_e with the total flux power density $P = P_r + P_i$. The steady-state solution for the resonator requires that

$$\alpha = \frac{\alpha_e(0)}{1 + BT_1 P} \qquad (21)$$

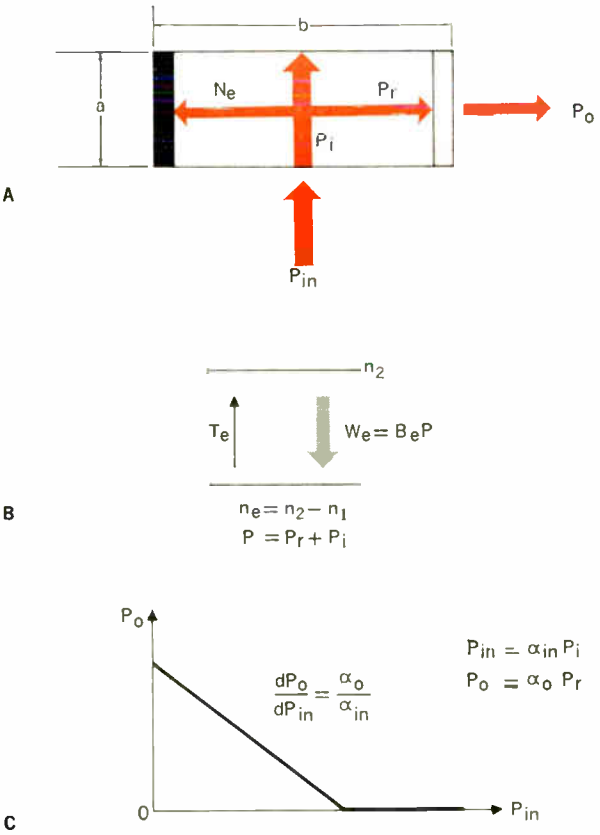


Fig. 10. NOR laser circuit. A—Schematic diagram of circuit. B—Model for laser material. C—P_o vs. P_{in}.

or

$$P_r = \frac{\alpha_e(0) - \alpha}{T_1 B \alpha} - P_i \qquad (22)$$

Equation (22) suggests a linear variation between the input P_{in} and the output P_o, with incremental gain of $-\alpha_o/\alpha_{in}$; see Fig. 10(C). It might appear that by making $\alpha_o/\alpha_{in} > 1$ one could obtain a linear gain of more than unity in the circuit of Fig. 10. However, according to this analysis, the inequality $\alpha_o/\alpha_{in} < b/a$ is required to assure that signal P_i will be forced by P_{in} and will not be self-sustained. Since P_o and P_{in} are power densities, the maximum overall gain of the output to input signals is still unity. Thus, this circuit gives inverting operation and an isolation between input and output; it is capable of a fan-in, but it is not capable of fan-out or a digital gain. At first glance, the laser inverter circuit does not seem to have a threshold for gain, as is required for stable digital circuits. However, if a gain is provided, as by a linear laser amplifier, the input–output transfer curve for two such inverters in series will give a threshold for gain for small signals as well as saturation of output for large signals.

When implemented with semiconductor lasers, such as GaAs, the inverter circuit should have a very fast response. The turnoff time for a large input signal, the fall time, is the decay time of the resonator. The rise time would be determined by the effective recovery time of the emissive population T_e, and would be somewhat longer

than the fall time. The rise time could be shorter than the lifetime of the minority carriers in the semiconductor lasers, which for GaAs lasers is of the order of 10^{-10} to 10^{-9} second.[23]

NOR laser logic circuits, then, could be visualized in the form of continuously pumped laser oscillators whose outputs can be quenched by signals coupled from similar oscillators by means of laser amplifiers. The amplifier–oscillator–amplifier combination must be stable, free of self-oscillations, and must have sufficient gain to allow the output of one circuit to inhibit at least two other oscillators; in other words, there should be a fan-out of two. There is a good possibility that a single, bilateral laser amplifier might be sufficient for this purpose, since in the NOR circuit the output signal is isolated from the input signal. A true unidirectional amplifier using some nonreciprocal effect, such as Faraday rotation, though desirable, may not be necessary.

A program is now under way at RCA Laboratories to develop a GaAs laser circuit.[26] By performing logic operations such as OR-NOT or AND-NOT, this device can serve as a basic building block for optical logic circuits. The laser inverter circuit consists of an amplifier and an oscillator. The oscillator junction area is a part of the amplifier junction area. In the absence of an input, the oscillator section of the inverter produces an output signal; but when an input signal is applied to the amplifier section, this signal is amplified to an intensity high enough to

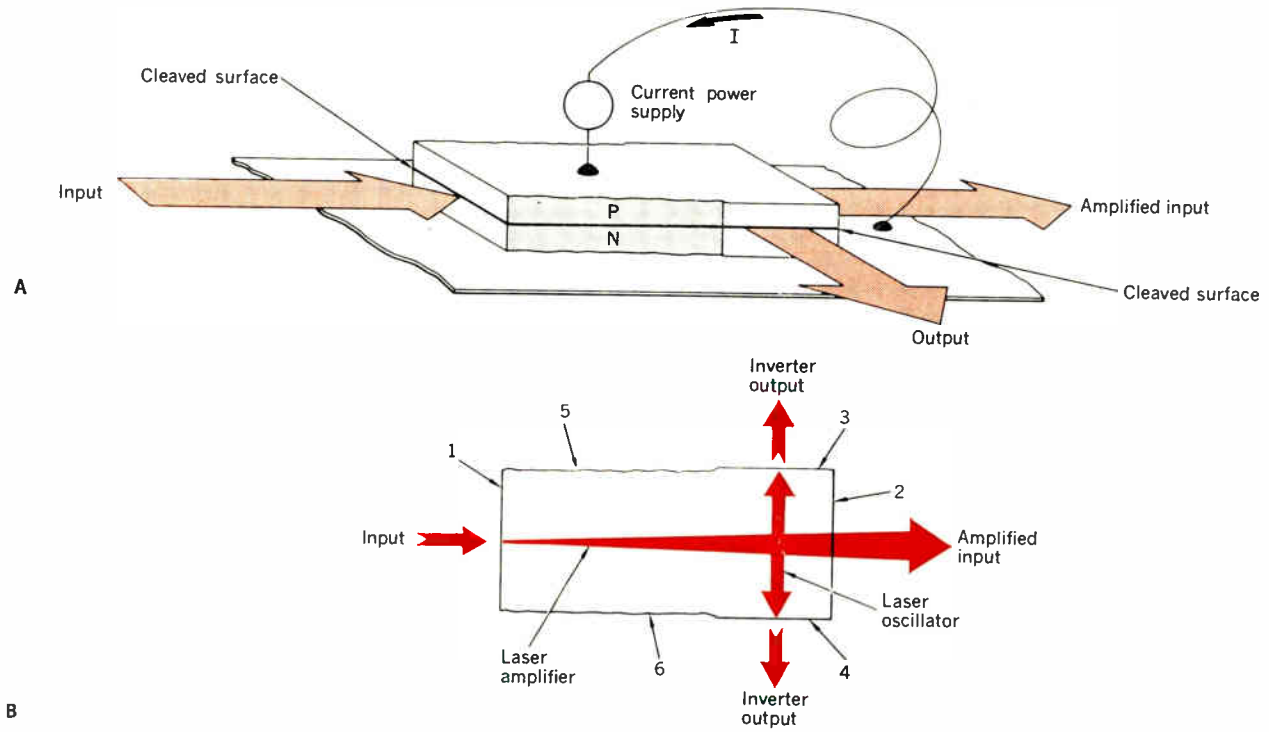
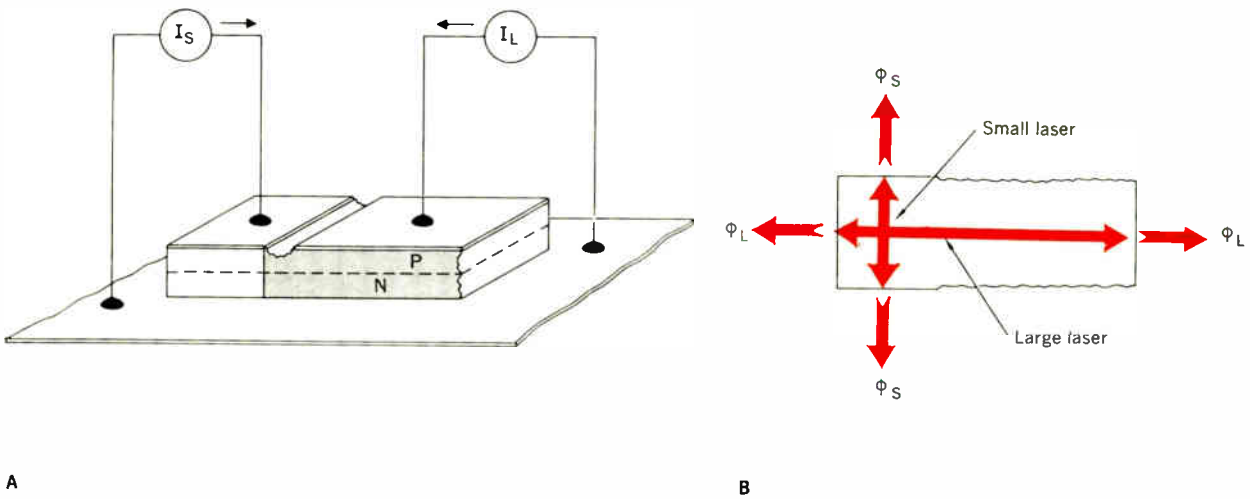


Fig. 11. Laser inverter. A—Pictorial representation. B—Active laser region.

Fig. 12. Dual laser oscillator.



lower the gain in the oscillator section to the point at which the output of the laser oscillator is quenched.

The construction and the operation of the laser inverter is sketched in Fig. 11. A laser amplifier is formed between sides 1 and 2 by decoupling the resonant cavity formed by these two sides, either by the use of nonreflective coatings on the sides[27] or by lapping side 2 so that it is no longer parallel to side 1.[4,28]

Efficient quenching of the output of a GaAs laser oscillator by means of a laser signal was demonstrated through use of a dual laser oscillator, as shown in Fig. 12. The dual laser oscillator, which contains two laser resonators within one laser structure, was made by carefully roughening part of the longer sides of a diode cleaved on all four sides, leaving two perpendicular resonators—one typically 10 by 4 mils and the other 30 by 10 mils. The resonators can be pumped separately with two current sources. Isolation was achieved by sawing a groove through the top metal contact into part of the semiconductor material while monitoring the leakage conductance. Typical isolation resistance between the two diodes is about one ohm.

The tested dual laser oscillators were produced from a solution-grown GaAs wafer, which yielded diodes that lased uniformly across the whole junction region. Typical operation is illustrated in Figs. 13 and 14. Figure 13(A) shows the waveforms of the currents applied to the large oscillator I_L and to the small oscillator I_S. Figure 13(B) shows the waveform of the output of the small laser. The quenching of the output ϕ_S of the small laser oscillator by the large laser oscillator output ϕ_L is demonstrated in Fig. 14. The family of curves shown is taken for fixed values of the small laser current I_S.

The operation of the dual laser oscillator demonstrated that the laser's quenching effect is a linear process; it can provide an inversion operation and a directionality for propagation of digital signals, but it lacks signal amplification, which is needed for digital circuits. The required amplification can be provided by a laser amplifier, as has been done in the case of the GaAs laser inverter. By decoupling the large laser oscillator cavity and converting it into a laser amplifier, the dual laser oscillator is converted into a laser inverter circuit.

Bistable circuit

A bistable laser circuit, such as a flip-flop, can be made from laser inverters or a single laser oscillator (see Fig. 15) that has a bistable operation if it has within its resonator a saturable absorber whose relation to the emitter is as shown in Fig. 16. In this figure the total loss coefficient α_T and the gain coefficient α_e are plotted as a function of

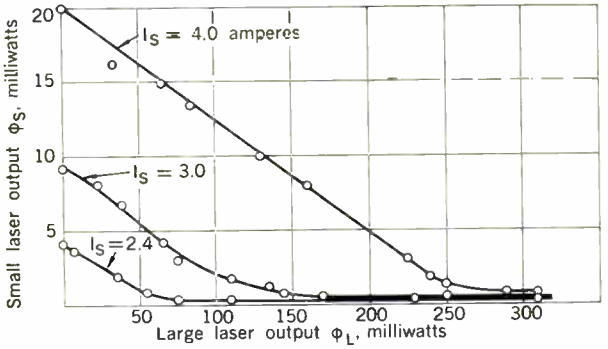


Fig. 14. Operation of dual laser oscillator, ϕ_S vs. ϕ_L.

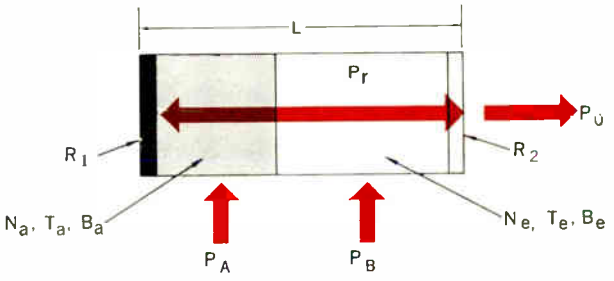
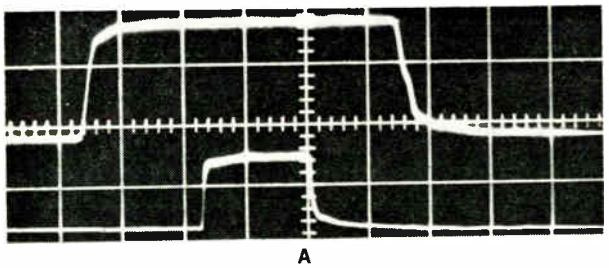
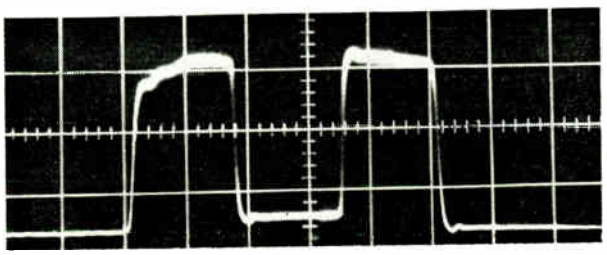


Fig. 15. Schematic diagram of laser resonator containing a saturable absorber.

Fig. 13. Signal waveforms for dual laser oscillator. A—Top trace: large-laser current I_L (2 amperes/div); bottom trace: small-laser current I_S (1.6 amperes/div). B—Output of small laser ϕ_S for $I_L = I_S = 4$ amperes and for $I_L = 4$ amperes, $I_S = 0$. Vertical scale: 0.1 volt/div into 50-ohm load; horizontal scale: 50 ns/div.

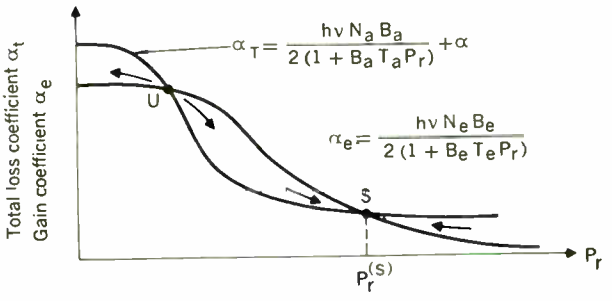


A



B

Fig. 16. Bistable operation of laser resonator in Fig. 15. Total loss coefficient $\alpha_t = \alpha_a$ and gain coefficient α_e are plotted against resonator flux-power density P_r.



the power density P_r in the resonator. The nonsaturable loss coefficient $\alpha = \alpha_r + \alpha_o$, where α_r represents the linear losses within the resonator and $\alpha_o = (2 - R_1 - R_2)/2L$, is the effective loss per centimeter length due to the reflectivities R_1 and R_2.

Note that point S is a stable equilibrium point. It gives the value of the steady-state signal $P_r^{(S)}$. Point U is an unstable equilibrium point. Inspection of Fig. 16 shows that $P_r = 0$ is another stable operating point. Operation of such a bistable, start–stop laser oscillator is sketched in Fig. 17. The saturable absorber recovery time T_a is chosen to be longer than the emitter recovery time T_e to assure a bistable rather than an astable operation. Initially, if the circuit is in the "0" stable state, it can be switched into the "1" state by input A. The oscillations of the circuit can be quenched by input B, thus returning the circuit to the "0" state. The outputs of such bistable circuits can control the threshold of the monostable circuits described in the next section. Inputs A and B can be provided by these monostable circuits.

If a three-level saturable absorber is used for the bistable circuit, a digital gain may be possible between input A and the output. A large digital gain cannot be expected from the bistable circuits, however, unless they are used in conjunction with laser amplifiers. The operation of a GaAs laser as a bistable circuit has been reported[29] using a GaAs laser device that is very similar to the large oscillator of the dual oscillator device described previously. More efficient bistable semiconductor laser circuits should be possible if special measures are taken to introduce an appropriate saturable absorber in the laser cavity. The repetition rate of a bistable circuit will be determined by the recovery time T_a of the saturable absorber.

Monostable circuit

If a saturable absorber material placed inside the laser resonator has $N_a B_a > N_e B_e$, $N_a < N_r$, and $T_a < T_e$, a monostable laser circuit is obtained. Such a circuit can be triggered by an input pulse signal to produce an output pulse. Figure 18(A) shows a monostable laser resonator with an input A, an input flux-power density P_A, and an output flux-power density P_o. The input signal, shown in Fig. 18(B), starts at time $t = t_{A1}$. The total loss coefficient $\alpha_t = \alpha_a + \alpha$ and the gain coefficient α_e of the circuit are sketched as a function of time in Fig. 18(C). Note that at time $t = t_{A2}$ the circuit reaches the condition for laser threshold; that is, $\alpha_e = \alpha_t$. The input energy W_A' required to reach laser threshold, and designated by the shaded area in Fig. 18(B), represents the lower limit of the input energy needed for triggering the circuit. The input pulse triggers the laser oscillator into a regenerative state, which in turn produces the output pulse shown in Fig. 18(D). The output has a delay time T_d, a rise time T_r, a fall time T_f, and a duration T_o. The pulse output can be of very short duration. The duration will be largely a function of the fall time T_f, which is determined by the total loss of the resonator; i.e., $T_f = 1/v\alpha$. The repetition rate will depend on the recovery time of the emissive population T_e. For the semiconductor lasers, the recovery time corresponds to the minority carrier lifetime, which in the case of GaAs lasers is in the range of 10^{-10} to 10^{-9} second.

A rough estimate of the energy gain G_W produced by this circuit is

$$G_W = \frac{W_o}{W_A} = \frac{N_e}{N_a} = \frac{B_a}{B_e}$$

A higher gain could be obtained if the strong regenerative

Fig. 17. Bistable circuit. A—Schematic diagram. B—Time variations of input A, input B, and output P. (Input A turns on laser oscillator; input B turns it off.)

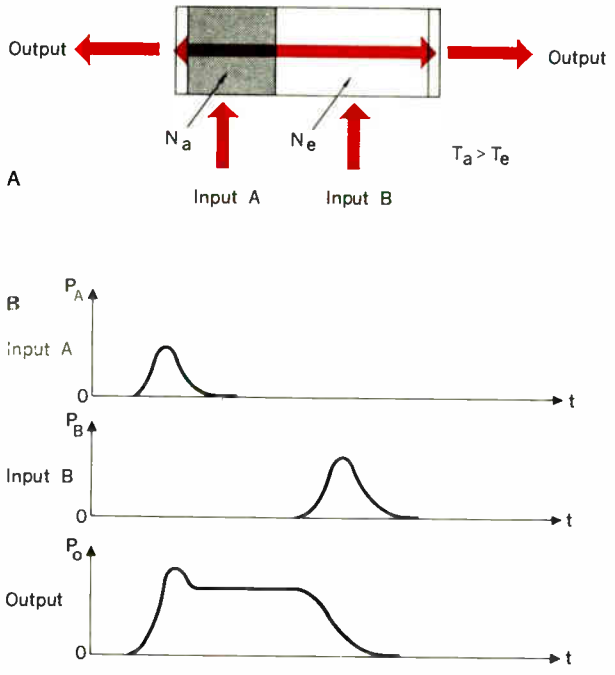
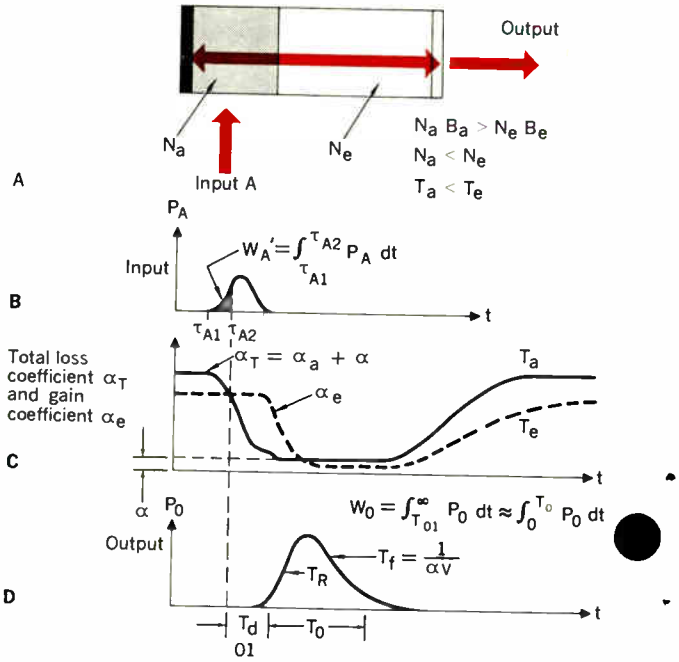


Fig. 18. Monostable circuit. A—Schematic diagram. B—Input signal as a function of time. C—Time variation of total loss coefficient α_t and gain coefficient α_e. D—Output signal as a function of time.



action occurred before the saturable absorber is saturated by the input.

Since more than one input can be applied, the circuit of Fig. 18 can be used as a monostable OR gate. Any one of a number of inputs could trigger the circuit to produce a pulse output, the shape of the output pulse being essentially independent of the input. Since the input cannot be coupled directly into the output, this circuit also provides isolation between the input and output.

A monostable laser resonator is shown in Fig. 19. Input *A* can lower the threshold for regenerative action, and input *B* can increase this threshold. If both inputs are in the form of pulses, as in Fig. 19(B), a large input pulse *B* will make the circuit inactive for the duration of the recovery time of the circuit. Since input *A* can be inhibited by input *B*, the latter controls the transmission of the former. If input *A* is an unconditional clock pulse, the circuit acts as an inverter for signal *B*. As indicated in this figure, the circuit is not expected to provide "digital gain" as an inverter. Therefore, input *B* must be obtained from a monostable amplifier circuit. If both inputs *A* and *B* are continuous bias signals, the operation of the circuit can be most conveniently described by the idealized transfer curves shown in Fig. 19(C). The solid-state curve represents the response of the circuit in the absence of bias signals. Bias signal *A* will reduce the threshold *T*, whereas bias signal *B* will increase the threshold.

At the present time it is not clear whether a good saturable absorber can be found to give semiconductor laser monostable circuits. The operation of a monostable laser circuit, however, was demonstrated using ruby lasers with solutions of phthalocyanine as saturable absorbers.[9-11] Equipment used for the demonstration of a monostable laser circuit is shown in Fig. 20. Laser 1, operating as an astable self-triggered oscillator, is used as a trigger source for laser 2, which is operating as a monostable laser circuit.

Typical operation of laser 1 as an astable circuit, or a relaxation oscillator, is described by Figs. 21 through 23. In Fig. 21 a comparison is made of the output energy of the ruby laser against the flash-tube power supply voltage. The operation of this relaxation oscillator as a function of time is shown in Fig. 22, in which the integrated laser output is recorded by an oscilloscope for the three condi-

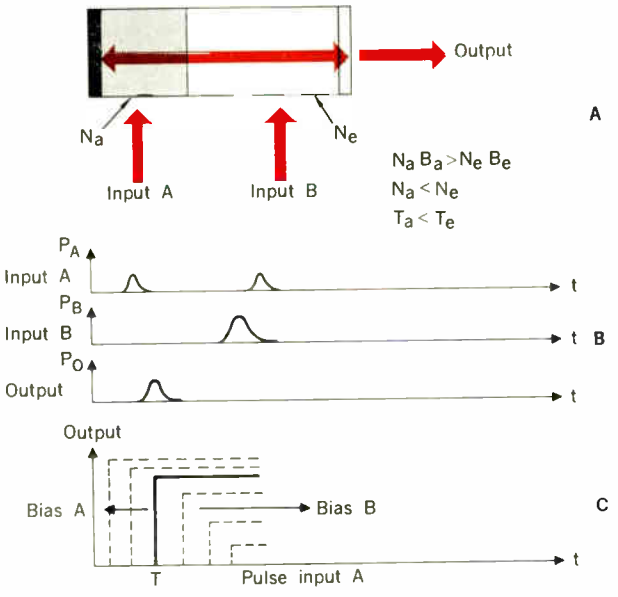
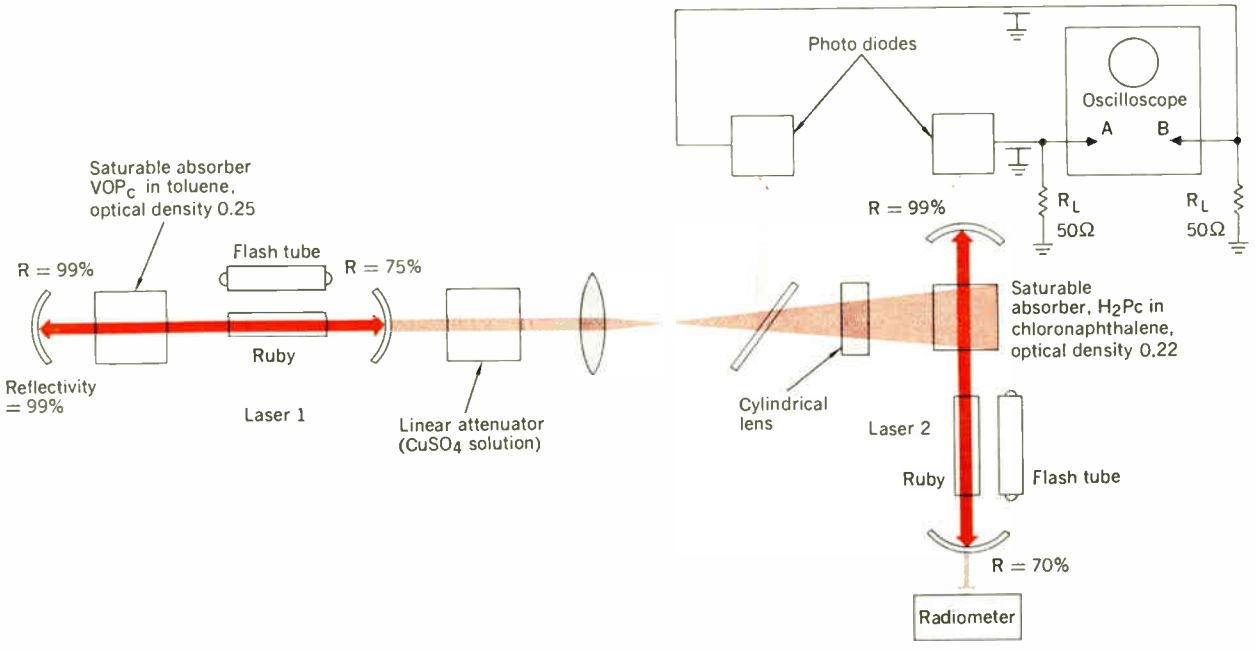


Fig. 19. Monostable circuit with two types of input. Input A enables the circuit to operate; input B inhibits operation. A—Schematic diagram. B—Time variation of input A, input B, and output. C—Idealized transfer curve representing operation of a variable-threshold monostable gate for pulse input A and bias inputs A and B.

Fig. 20. Experiment for demonstration of ruby laser with phthalocyanine saturable absorbers operating as relaxation oscillator (laser 1) and monostable circuit (laser 2).



tions of Fig. 21 and for a flash-tube power supply voltage of 1.7 kV. Figure 23 shows a traveling-wave oscilloscope picture of an output pulse wave shape detected by an ITT W-128 diode, which has a subnanosecond rise time.

The spectroscopic tests of the saturable absorbers used for lasers 1 and 2 were shown in Figs. 6 through 8. For the test shown in Fig. 20, the output waveforms for lasers 1 and 2 are shown in Fig. 24. Both lasers were fired at a flash-tube power supply voltage of 1.8 kV. The delay of the output of laser 2 is a function of the inverted population; it could be increased by increasing the distance between the spherical mirrors of laser 2. Delays were obtained from 0.1 to 0.5 μs. Operation of a monostable circuit with such a long delay is possible only with metal-free solutions of phthalocyanine, which when saturated have recovery times of about 1 μs.

The transfer curve for the monostable laser 2 is shown in Fig. 25. A gain of two was demonstrated even though laser 2 appeared to be less efficient than laser 1. Operating as a relaxation oscillator with VOPc in toluene as a saturable absorber, laser 2 again produced only 30-millijoule output pulses.

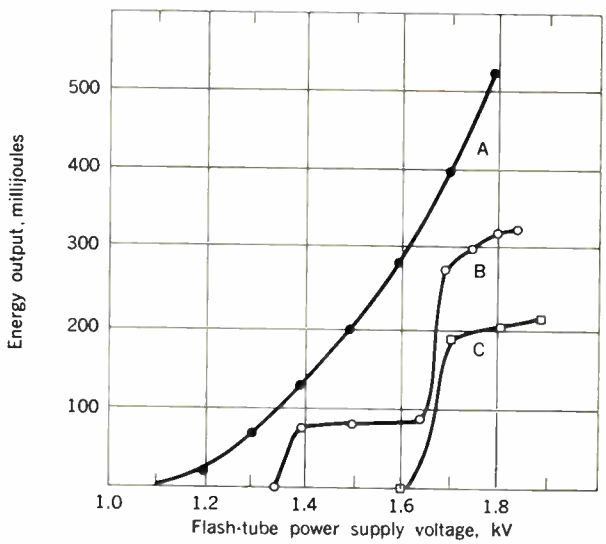


Fig. 21. Output energy of a ruby laser as a function of flash-tube supply voltage (A) without a saturable absorber, (B) with H_2Pc in toluene, optical density of 0.2, and (C) with H_2Pc in toluene, optical density of 0.3.

Fig. 22. Integrated output of a ruby laser. Horizontal scale: 50 μs/div; vertical scale: 0.5 volt/div; power-supply voltage: 1.7 kV. A—No saturable absorber. B—H_2Pc in toluene, optical density of 0.2. C—H_2Pc in toluene, optical density of 0.3.

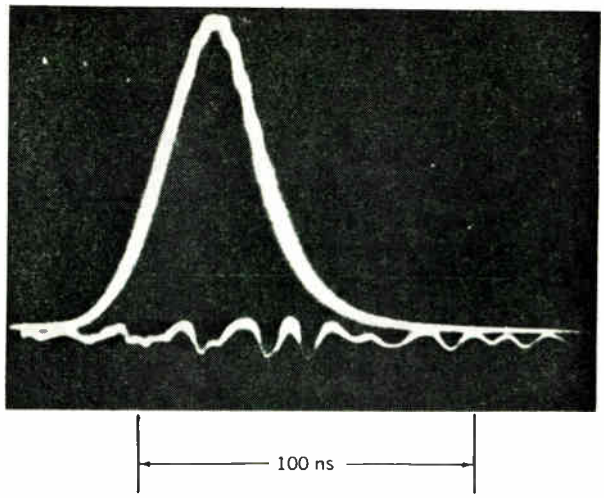


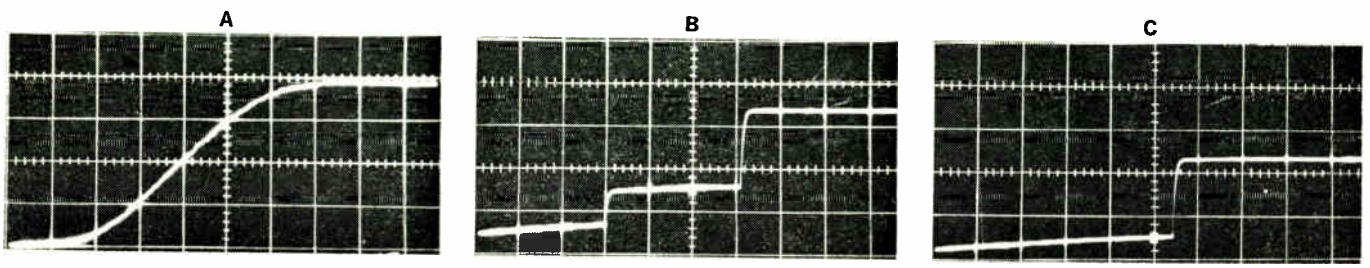
Fig. 23. Output pulse of a Q-switched laser.

A triggerable operation was obtained without a delay time between the input and the output using a similar ruby laser system but aligned so that the output of laser 1 couples into the resonant modes of laser 2. In this case the c axes of both ruby crystals were made to be parallel and the two laser pump assemblies while positioned within two confocal spherical mirrors were separated by a flat mirror with 80 per cent reflectivity.

Initially, an attempt was also made to use an unpumped ruby as a saturable absorber for a ruby laser. The desired relation between the matrix elements of the unpumped ruby absorber and the pumped ruby emitter was to be obtained by having a resonant laser cavity only for one polarization and an appropriate orientation of the c axis of the crystals. At room temperature, the performance of this triggerable laser oscillator, using a ruby as the saturable absorber, was only marginal. In liquid nitrogen environment, however, this approach is expected to be more successful.

Conclusions

It has been shown how lasers can be used as basic components for digital computer circuits. The experiments using ruby lasers were conducted chiefly to demonstrate effects that could lead to the development of laser digital devices. Semiconductor current-injection lasers are most attractive for digital devices because of their small size, high pumping efficiency, and high speed of operation. Two approaches are suggested: (1) the development of NOR laser circuits in the form of stable laser amplifier–



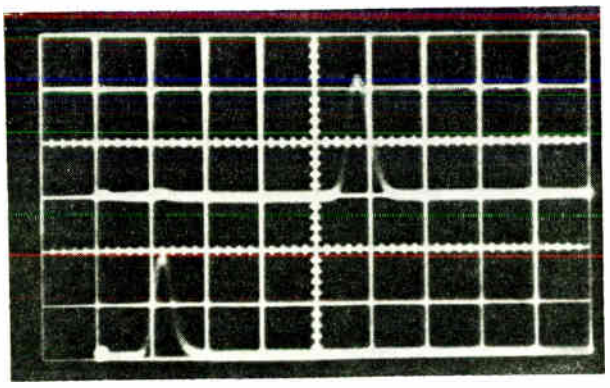
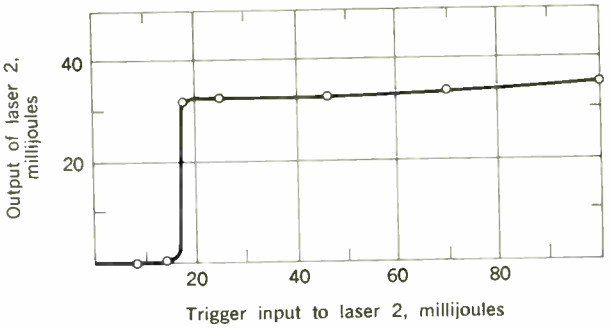


Fig. 24. Output waveforms of laser 1 and laser 2 in Fig. 20. Horizontal scale: 0.1 μs/div; upper trace: laser 1, 10 volts/div; lower trace: laser 1, 5 volts/div. Both traces have the same sweep internally triggered by the lower trace signal (laser 1).

Fig. 25. Input-output curve for monostable laser 2 triggered by output of laser 1.



laser oscillator combinations and (2) the study of semiconductor laser devices that make use of the saturation of optical absorption in addition to laser amplification and laser quenching effects.

Semiconductor laser materials technology has arrived at a stage at which more and more emphasis is being placed on research in the area of useful devices and their applications. Laser digital circuits will be one of these applications. In the next year or so, we should be in a better position to ascertain the role lasers will play in the digital computer field. We are optimistic that in the near future some forms of working laser digital circuits will be developed. In view of the present state of the art in GaAs lasers, it is anticipated that these devices will be operable at liquid nitrogen temperatures.

This article is based on two papers presented at the Symposium on Optical and Electro-Optical Information Processing Technology, Boston, Mass., Nov. 9–10, 1964. The original papers will appear in the proceedings of the symposium, *Optical and Electro-Optical Information Processing*, edited by J. T. Tippett, L. C. Clapp, D. Berkowitz, C. J. Koester, and A. Vanderburgh, to be published in 1965 by M.I.T. Press, Cambridge, Mass.

The research reported in "Laser Digital Devices" was sponsored by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, N.Y., under Contract AF30(602)-3169. This work was done in partial fulfillment of the requirements for the Eng.Sc.D. degree at the Department of Engineering Columbia University. The author acknowledges the support of J. A. Rajchman, S. E. Harrison, and many other co-workers at RCA Laboratories.

REFERENCES

1. Amodei, J. J., "High-Speed Adders and Comparators Using Transistors and Tunnel Diodes," *IEEE Trans. on Electronic Computers*, vol. EC-13, Oct. 1964, pp. 563–575.

2. Biard, J. R., *et al.*, "Optoelectronics as Applied to Functional Electronic Blocks," *Proc. IEEE*, vol. 52, Dec. 1964, pp. 1529–1536.

3. Koester, C. J., "Study of Optical Fiber Techniques for Data Processing," RADC-TDR-62-478, AD 299 007, Final Rept. on Contract AF30(602)-2440, American Optical Co., Aug. 1962.

4. Koester, C. J., "Development of Glass Fiber Lasers," RADC-TDR-64-79, Final Rept. on Contract AF30(602)-2829, American Optical Co., May 1964.

5. Fowler, A. B., "Quenching of Gallium Arsenide Injection Lasers," *Appl. Phys. Letters*, vol. 3, no. 1, July 1, 1963, p. 1.

6. Koester, C. J., and Swoop, C. H., "Some Laser Effects Potentially Useful in Optical Logic," presented at the Symposium on Optical and Electro-Optical Information Processing Technology, Boston, Mass., Nov. 9–10, 1964.

7. Kosonocky, W. F., "Feasibility of Neuristor Laser Computers," *Proc. Symp. Optical Processing of Information*, Washington, D.C., Oct. 23–24, 1962. Baltimore: Spartan Books, Inc., 1963.

8. "Neuristor Logic Technology," RADC-TDR-64-123, Final Rept. on Contract AF30-(602)-2761, Applied Research Dept., Defense Electronic Products, RCA, Camden, N.J., June 1964.

9. Sorokin, P. P., *et al.*, "Ruby Laser Q Switching Elements Using Phthalocyanine Molecules in Solution," *IBM J.*, vol. 8, Apr. 1964, p. 182.

10. Stark, P. E., *et al.*, "Saturable Filter Investigation," Semiann. Tech. Rept. (July–Dec. 1963), NOnr-4125(00), NRO 15-702, Lear Seigler, Inc., Laser System Center, Feb. 18, 1964.

11. Bret, G., and Gires, F., "Giant-Pulse Laser and Light Amplifier, Using Variable Transmission Coefficient Glasses as Light Switches," *Appl. Phys. Letters*, vol. 4, no. 10, May 15, 1964. p. 175.

12. Wittke, J. P., and Warter, P. J., "Pulse Propagation in Laser Amplifier," *J. Appl. Phys.*, vol. 35, June 1964, pp. 1668–1672.

13. Kosonocky, W. F., "Laser Digital Devices," presented at the Symposium on Optical and Electro-Optical Information Processing Technology, Boston, Mass., Nov. 9–10, 1964.

14. Burns, G., and Nathan, M. I., "P-N Junction Lasers," *Proc. IEEE*, vol. 52, July 1964, pp. 770–794.

15. Hall, R. N., "Recombination Processes in Semiconductors," *Proc. IEE (London)*, Mar. 1960, p. 923.

16. Mayburg, S., and Black, J., "Dependence of Recombination Radiation on Current in GaAs Diodes," *J. Appl. Phys.*, vol. 34, May 1964.

17. Mayburg, S., "Direct Recombination in GaAs and Some Consequences in Transistor Design," *Solid-State Electron.*, vol. 2, 1961, p. 195.

18. Nathan, M. I., and Burns, G., "Recombination Radiation in GaAs by Optical and Electrical Injection," *Appl. Phys. Letters*, vol. 1, no. 4, Dec. 1, 1962.

19. Nelson, D. F., *et al.*, "Band-Filling Model for GaAs Injection Luminescence," *Ibid.*, vol. 2, no. 9, May 1, 1963.

20. Dousmanis, G. C., *et al.*, "Effect of Doping on Frequency of Stimulated and Incoherent Emission in GaAs Diodes," *Ibid.*, vol. 3, no. 8, Oct. 15, 1963.

21. Dousmanis, G. C., *et al.*, "Temperature Dependence of Threshold Current in GaAs Lasers," *Ibid.*, vol. 5, no. 9, Nov. 1, 1964.

22. Konnerth, K., and Lanza, C., "Delay Between Current Pulse and Light Emission of a Gallium Arsenide Injection Laser," *Ibid.*, vol. 4, no. 7, Apr. 1, 1964, pp. 120–121.

23. Marinace, J. C., "High-Power CW Operation of GaAs Injection Lasers at 77°K," *IBM J.*, vol. 8, no. 5, Nov. 1964.

24. Koester, C. J., "Some Properties of Optical Fibers and Lasers (Part B)," *Proc. Symp. Optical Processing of Information*, Washington, D.C., Oct. 23–24, 1962. Baltimore: Spartan Books, Inc., 1963.

25. Koester, C. J., "Possible Use of Lasers in Optical Logic Functions," *Proc. 1963 IEEE Pacific Computer Conf.*, p. 54.

26. Kosonocky, W. F., *et al.*, "GaAs Laser Inverter," *Digest Internat'l Solid-State Circuits Conf.*, Philadelphia, Pa., Feb. 17–19, 1965.

27. Crowe, J. W., and Craig, R. M., "Small-Signal Amplification in GaAs Lasers," *Appl. Phys. Letters*, vol. 4, no. 3, Feb. 1964.

28. Koester, C. J., and Snitzer, E., "Amplification in a Fiber Laser," *Appl. Opt.*, vol. 3, Oct. 1964, p. 1182.

29. Nathan, M. I., *et al.*, "A GaAs Injection Laser with Novel Mode Control and Switching Properties," presented at the Solid-State Device Conference, Boulder, Colo., July 1964.

Authors

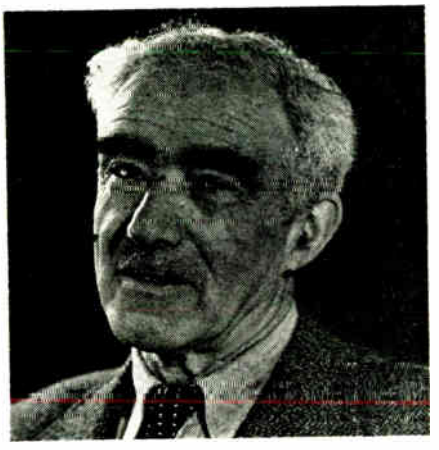
Nilo Lindgren received the B.S. degree in electrical engineering (communications) from the Massachusetts Institute of Technology in 1948, and thereafter studied art, literature, and psychology at various schools, including the Tyler School of Fine Arts, the University of Pennsylvania, and the Institute of Gestalt Therapy in New York City. He worked as a technical writer and editor in the Research Division of the Philco Corporation, Philadelphia, Pa., and in the Research Department of the Hughes Aircraft Company, Culver City, Calif., where he also was involved in writing for industrial films. He later served as technical editor for the Research Department of Grumman Aircraft Company, Bethpage, N.Y., for a number of years. More recently, he was an editor on *Electronics*, McGraw-Hill Publishing Company, for which he wrote many articles, including special surveys on microminiaturization and bionics. Mr. Lindgren is a member of Eta Kappa Nu.

Michael Petrick is at present on the staff of Reactor Engineering as a group leader of the Heat Engineering Section at the Argonne National Laboratory, Argonne, Ill. He did his graduate work at the University of Notre Dame, from which he received the master's degree, and at the Illinois Institute of Technology, from which he received the Ph.D. degree in chemical engineering in 1958. Since he started his association with Argonne National Laboratory his interest has been primarily in the field of two-phase flow, with emphasis on boiling water reactor problems and general two-phase heat transfer and fluid flow problems.

Dr. Petrick is the author and co-author of numerous articles on the subject of two-phase flow. More recently, he has been devoting his efforts to research work in liquid-metal magnetohydrodynamics. His current studies involve overall binary cycle efficiencies and experimental programs on nozzles and generators.

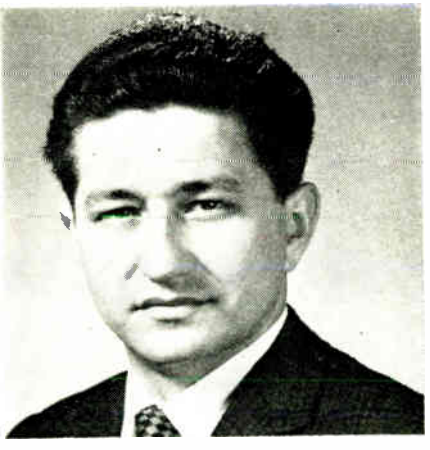
J. B. Atkins (SM) completed his undergraduate work at the University of Arkansas, from which he received the B.S.E.E. degree in 1960. At present he is enrolled in the Graduate School of Syracuse University and is writing a thesis, which will complete his requirements for the M.S.E.E. degree. He worked with the IBM Corporation in Poughkeepsie, N.Y., during the summer of 1959 and rejoined the company following his graduation in 1960. He worked in the Data Systems Division for 1½ years, chiefly in statistical analysis and the design of logic circuits. He has spent the past 3½ years in the Components Division, where his chief concern is the design and development of various logic circuits. Mr. Atkins is a member of Eta Kappa Nu and Tau Beta Pi.

Henry Ehrenreich received the A.B. degree in physics from Cornell University in 1950. After a year's graduate work at Columbia University he returned to Cornell, where he did his dissertation work in solid-state theory and received the Ph.D. degree in 1955. He then joined the General Electric Research Laboratory, Schenectady, N.Y., where he remained until 1963. During the fall term of 1960 he was a visiting lecturer at Harvard University, and in 1963 was appointed Gordon McKay Professor of Applied Physics at Harvard. His interests in solid-state theory have been concentrated in the transport properties and the band structure of semiconductors and in the optical properties of solids. He has also worked in the areas of ultrasonic attenuation in insulators, the many-body problem, and ferromagnetism in metals. Dr. Ehrenreich is a Fellow of the American Physical Society and a member of Sigma Xi and Phi Beta Kappa.

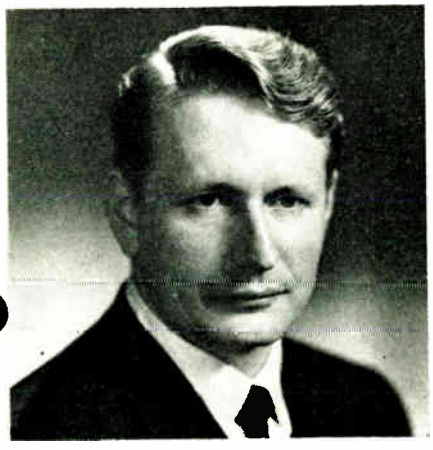


J. L. Salpeter received the Ph.D. degree in physics in 1909 from the University of Vienna, Austria, where he also had a number of papers published on atmospheric electricity. His early work included one of the first theories on the reflection coefficient of an ionized gas for electromagnetic waves. He is also the author of a textbook on calculus for natural scientists, first published in 1913. Upon completion of his university studies, he entered industry, working on incandescent lamps and radio tubes, first in Hungary and then in Austria. He invented a process for manufacturing small-grain coiled tungsten filaments, which was used in Europe for many years for vibrationproof lamps.

In 1939 Dr. Salpeter traveled to London and eventually to Australia, where he became associated with Philips Electrical Industries as a research physicist. He was engaged for a number of years in studies on fluorescent lighting, television optics, and magnetic materials. At present he lives in retirement in Adelaide, South Australia.



Oskar A. Reimann came to the United States from Germany in April 1952. He attended St. Michael's College and the University of Toronto, and received the B.S. degree in physics from St. John Fisher College, Rochester, N.Y., in 1962. He is currently taking graduate courses in physics at Syracuse University. Since 1962 he has been employed at Rome Air Development Center, where as a member of the Data Processing Section of the Information Processing Branch he is engaged in the development of computer elements and components. Since joining RADC he has been particularly concerned with "all-optical" digital computer elements. The initial effort under this task was the feasibility determination of the neuristor laser computer, as proposed by the Radio Corporation of America. The properties of glass-laser fibers were studied to determine their applicability to the neuristor computer as well as to digital data processing in general. He is now working on the development of semiconductor laser digital devices and related efforts.



Walter F. Kosonocky (SM) received the B.S. and M.S. degrees in electrical engineering in 1955 and 1957, respectively, from Newark College of Engineering. At present he is a candidate for the Sc.D. degree in engineering at Columbia University. Since 1955 he has been employed at RCA Laboratories, where he is now in the Computer Research Laboratory. His work has included development of the ferrite aperture plate memory system, transistor application in computer circuits, investigation of ferrite cores and micromagnetic techniques for high-speed computers, and a study of pattern-recognition systems. Since April 1962 he has been engaged in a feasibility study of the use of lasers as digital computer components. He received RCA Laboratories Achievement Awards in 1959 and in 1963 for his contributions to the application of parametric devices for digital computer logic and memory systems and of tunnel diodes and transistors for high-speed computer systems, respectively. He is a member of Tau Beta Pi and Eta Kappa Nu and is the author of six technical papers.

Classified Advertising

Positions Open

The following positions of interest to IEEE members have been reported as open. Apply in writing, addressing reply to address given or to Box Number, c/o IEEE Spectrum, Advertising Department, 72 West 45th St., New York, N.Y. 10036.

Classified Advertising Rates for this column: $6.00 per line. No advertising agency commission is granted. Copy must be received by the 2nd of the month preceding date of issue.

Electrical Engineer. Chicago international consulting engineering firm needs electrical engineer experienced in all phases of electrical design of hydroplants. Salary commensurate with experience. Some relocation costs and fringe benefits. Submit resume including salary requirements to the Personnel Manager. Harza Engineering Company, 400 West Madison Street, Chicago, Illinois 60606. An equal opportunity employer.

Faculty member required for undergraduate and graduate teaching and research. Control system specialist with Ph.D. in Electrical Engineering. Industrial experience desirable. Rank open. Competitive salary. Write: Chairman, Department of Electrical Engineering, McMaster University, Hamilton, Ontario.

Electrical Design Engineers. Synchronous, induction, or D.C. machine design. Experience in programming rotating machine design for computer desirable, but not required. Send resume to R. Green, Electric Machinery Mfg. Company, 800 Central Avenue, Minneapolis, Minnesota 55413. An equal opportunity employer.

Laser Systems Analysis for Optical Radar Applications. M.S. or Ph.D. in applied mathematics, electrical engineering or physics. Minimum of 3 years industrial experience in the following fields: missile systems analysis, threat analysis, operations research, signal to noise estimation, analog simulation, error analysis, etc. Send resume to Donald J. Kopczick, Korad Corporation, A Subsidiary of Union Carbide Corp., 2520 Colorado Ave., Santa Monica, California 90406. An equal opportunity employer.

Ph.D. faculty member needed for teaching and research. Good publication record essential. Write: Head, EE Department, University of Tennessee, Knoxville, Tennessee 37916.

Sales Engineers. Challenging openings at several offices in the U. S. Also an opening for application engineering work and handling of customer requirements between Home Office and Field Sales office. A step to more responsibility in either the Home Office or Field Sales office work. Send resume to R. Green, Electric Machinery Mfg. Company, 800 Central Avenue, Minneapolis, Minnesota 55413. An equal opportunity employer.

Electronics Engineer for Laser Development Programs. M.S. degree preferred, for design of electronic systems utilizing Lasers as output devices. Commercial Programs include semiconductor pulse circuitry for use with gallium arsenide Lasers as well as high power modulators with ruby devices. At least 3 years experience in radar systems or pulse circuitry required. Send resume to Donald J. Kopczick, Korad Corporation, A Subsidiary of Union Carbide Corp., 2520 Colorado Avenue, Santa Monica, California 90406. An equal opportunity employer.

Electrical Engineer. International consulting engineering firm needs Chief Distribution Engineer for overseas assignment in family status. Must be capable of training foreign nationals and eventually assuming responsibility for complete charge of distribution department. Salary commensurate with experience. Fringe benefits. Please send complete resume including salary requirements to the Personnel Manager, Harza Engineering Company, 400 West Madison Street, Chicago, Illinois 60606. An Equal Opportunity Employer.

Electronics engineer, for radio system engineering, including site surveys, map studies, performance and reliability calculations for VHF/UHF, 40% time travelling, 60% at Milan Hq's. Spanish mother tongue, English desirable, with more than 5 years Project Engineering. All applications handled confidential. Submit resume including salary requirements to PEO Automatic Electric, Via Bernina, 6 Milano—Italy.

Engineering Sales Trainees West Coast. B.S.E.E. about 25 years old, prefer some industry experience. Factory, warehouse and office training course leads to Field Engineer positions in Calif., Oregon or Washington with full benefits and company car. Large worldwide electrical manufacturing corporation. Apply District Manager, 3323 San Fernando Road, Los Angeles.

Professor—Alaska: To teach undergraduate electrical engineering, especially the design and operation of moderate-sized power systems, and to develop research or consulting. Ideal preparation would be Master's plus several years experience. Write J. G. Tryon, Department of Electrical Engineering, University of Alaska, College, Alaska 99735.

Faculty Member needed for teaching modern courses in an undergraduate program in Electronic Engineering. Ideal for person whose first interest is teaching. Prefer candidate with a recent M.S. degree and a strong interest in solid state electronics. Send resume to Dean of Engineering, California State Polytechnic College, San Luis Obispo, California.

Manager, Product Development. An Electrical Engineer to fill an excellent career position with a non-defense manufacturer. Position involves developing products for new or improved applications from already established basic concepts, for low-priced, mass-produced precision mechanisms. He will assume responsibility for development and coordination from initial concept to customer installation. Experience in design & test of switches, solenoids, or time-delay relays is essential, together with knowledge of and background in switching theory, circuitry transients, and contact metallurgy. Excellent working and living conditions. Location in metropolitan center in Mid-West. Salary is commensurate with qualifications. Real opportunity for advancement. Please send detailed resume of experience including products handled and salary history. Box 6026.

Two Ph.D. faculty members needed for growing graduate program. One in Power area and the other with interest in Networks, Information Theory or other. Write: Chairman, E.E. Dept., Vanderbilt University, Nashville, Tennessee 37203.

Positions Wanted

Classified advertising Rates for "Positions Wanted" column: $6.00 per line per insertion. 50% discount for members of IEEE (please supply membership number with copy). Copy must be received by the 2nd of the month preceding date of issue. Address replies to Box Number given, c/o IEEE Spectrum, Advertising Department, 72 West 45th St., New York, N.Y. 10036.

Sales—Electrical Designer, 38, resident of N. Delaware, seeks position with electrical equipment manufacturer. Box 8027.

Electrical Engineer, 29, MS. 6 years industry experience, desires college research or experiment station position or R&D work with small company. Southeast preferred. Major interests: digital computer technology, radar systems analysis. Box 8028.

Electronics Engineer, retired, seeks part time position, N.Y. City area. Box 8029.