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.

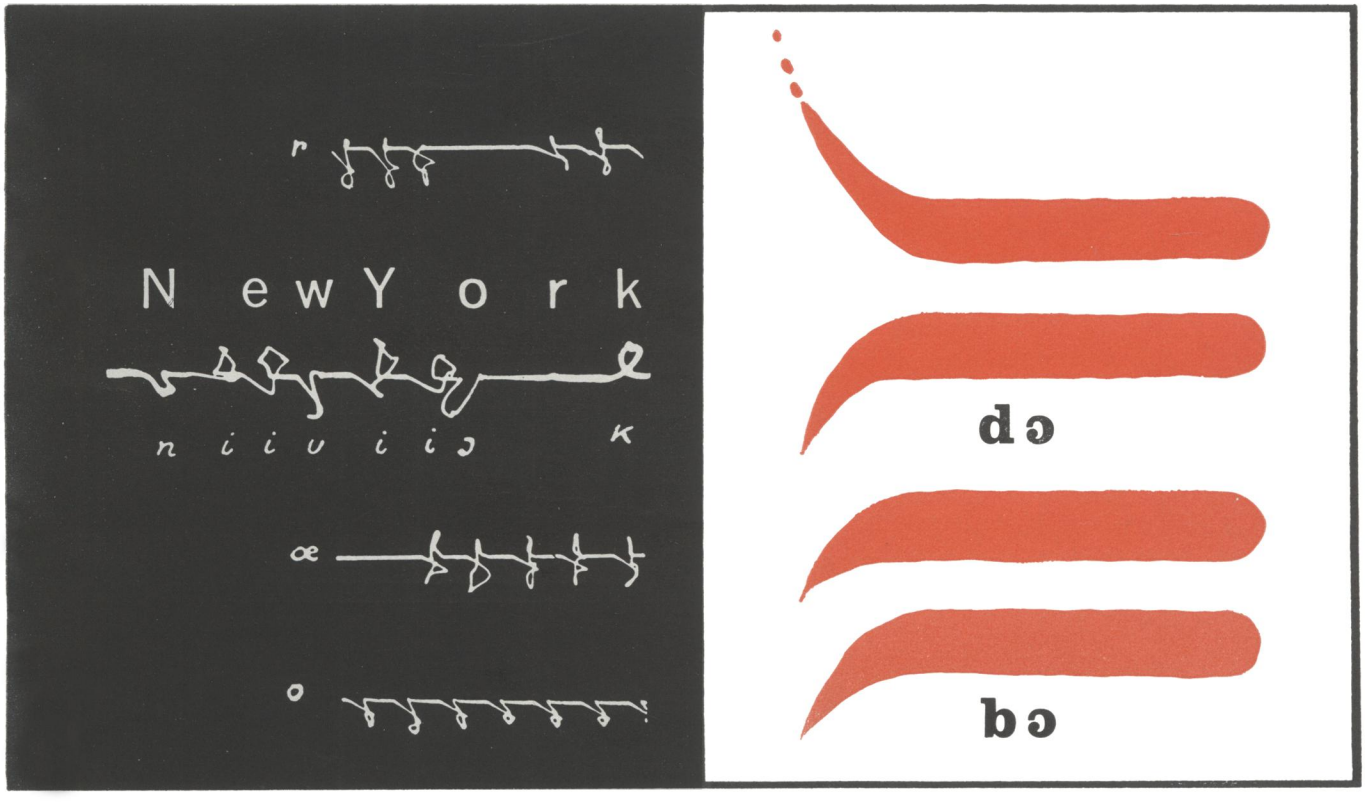


Fig. 1. Early Egyptian prealphabetic signs (left) were compared by Dreyfus-Graf to pictographs (center) produced by his speech recognizer. Today, strikingly simple hand-painted "cues" (right) can be used by machines to produce intelligible speech.

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]

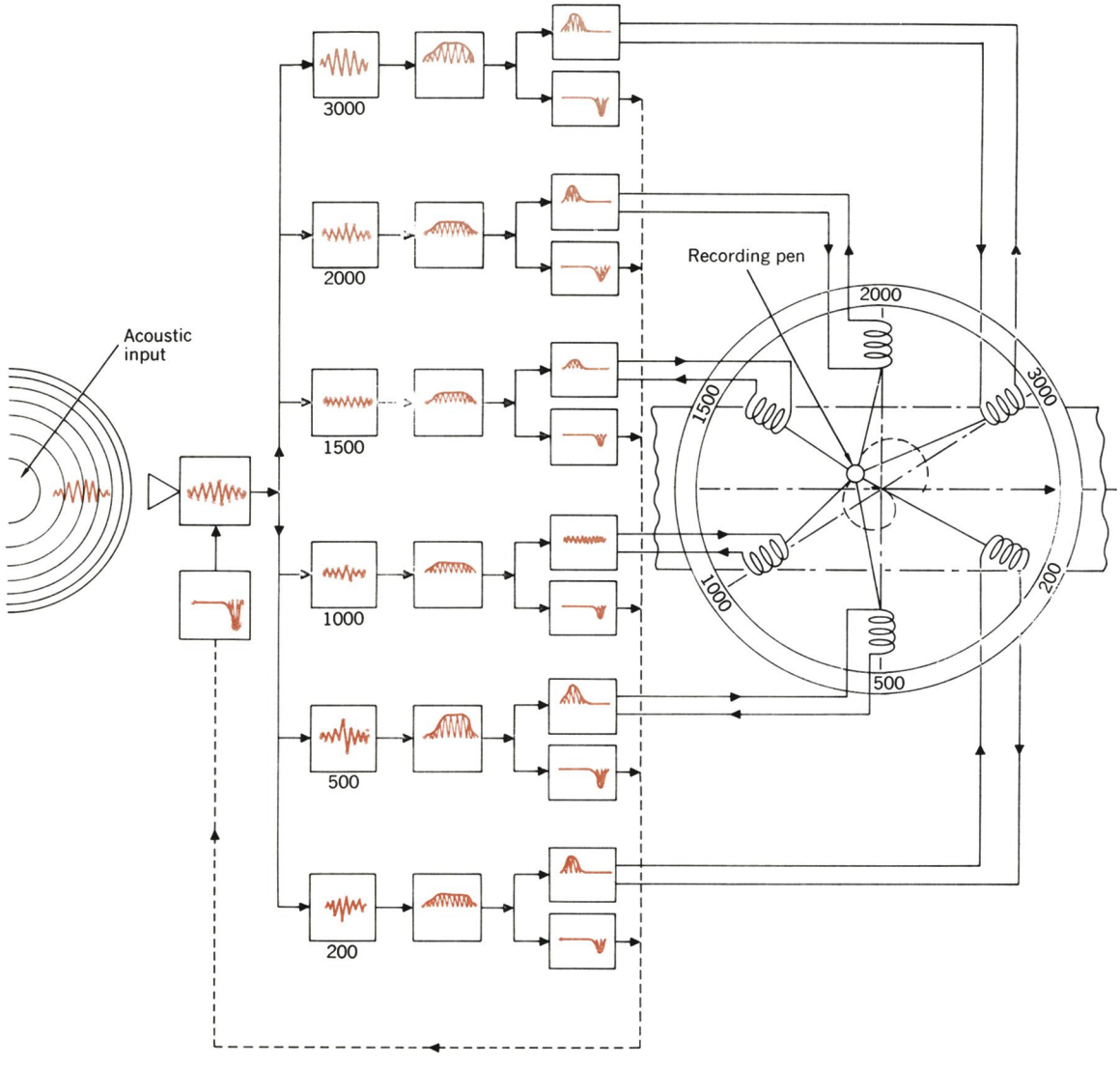
These remarks do not mean to imply, however, that speech-compression research has not contributed understanding to the automatic speech recognition problem; it certainly has. The point is only that these systems require the intervention of a *perceptive* human, which it is the aim of automatic speech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

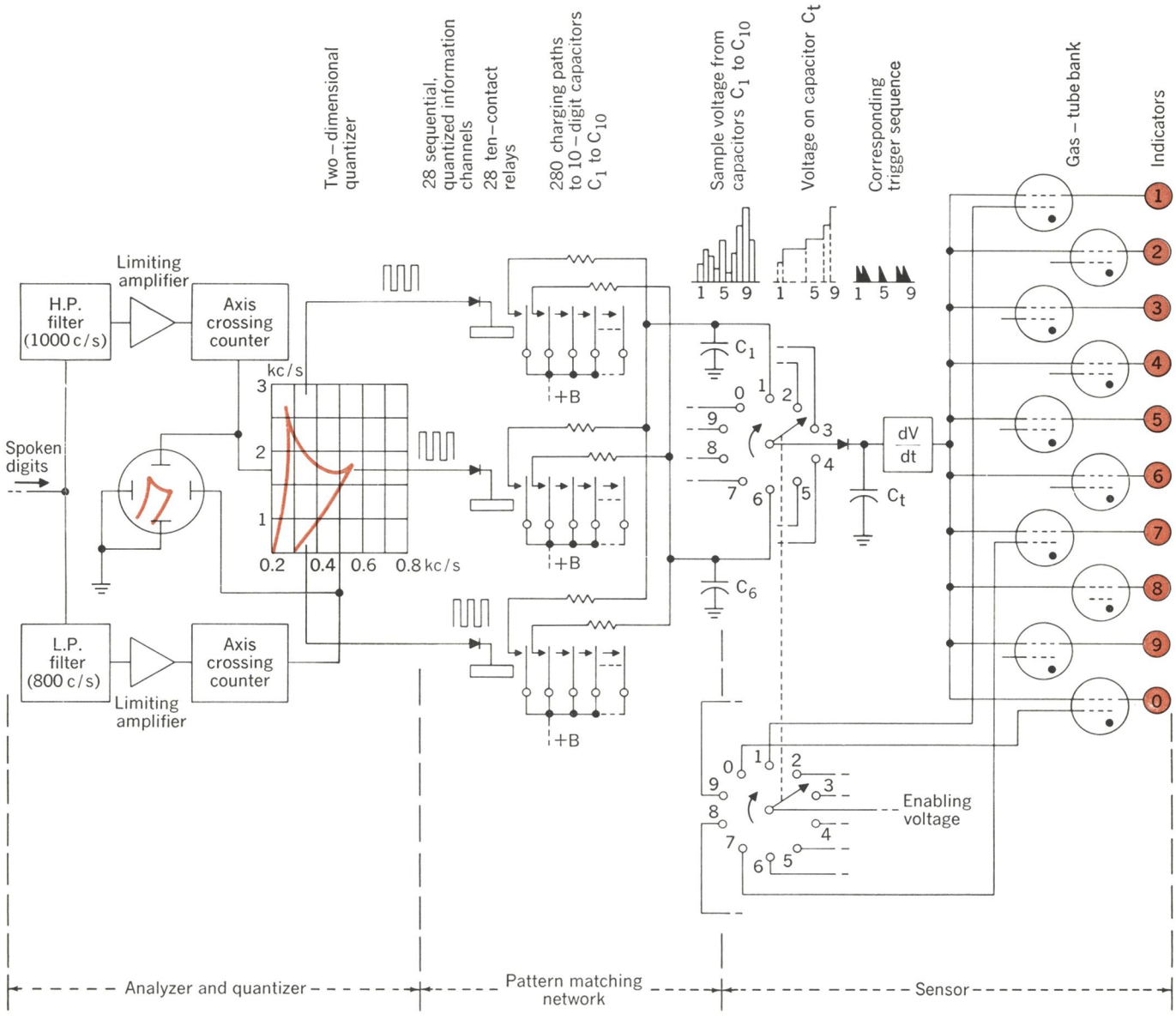
Five years ago, in an interesting survey on machine recognition of spoken words, Richard Fatehchand stated that only limited success had been achieved with speech recognition machines. No machine existed, he said, that would deal with continuous speech.[7] Today, there is still no such machine. Nor is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

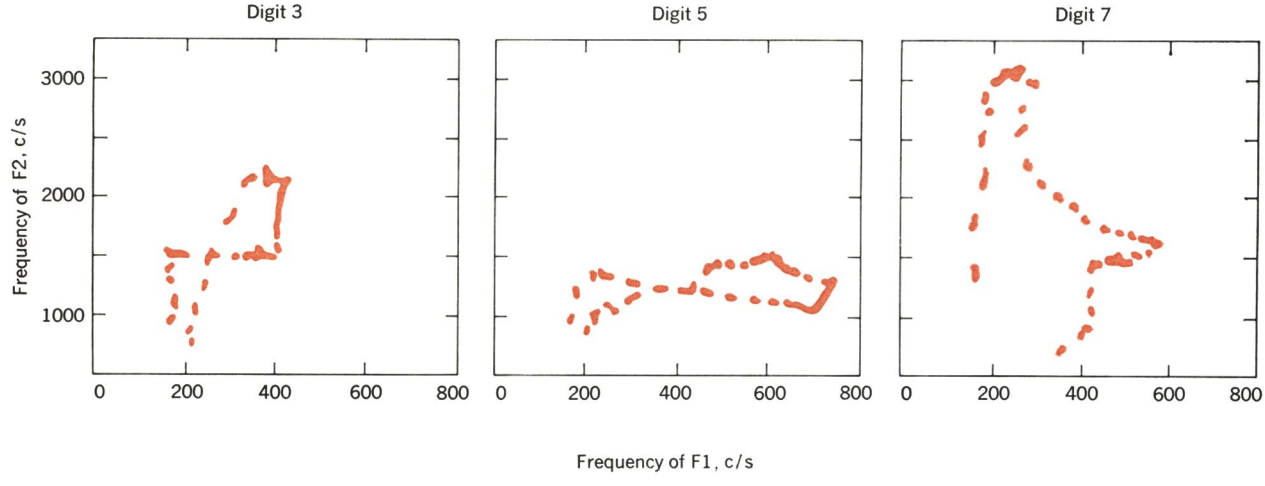
In 1956 an automatic speech recognizer based on an entirely different operating principle was designed at Northeastern University by J. Wiren and H. L. Stubbs.[11] This electronic machine was designed to sort out elementary sounds of speech (phonemes) by a process of successive binary decisions about 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.



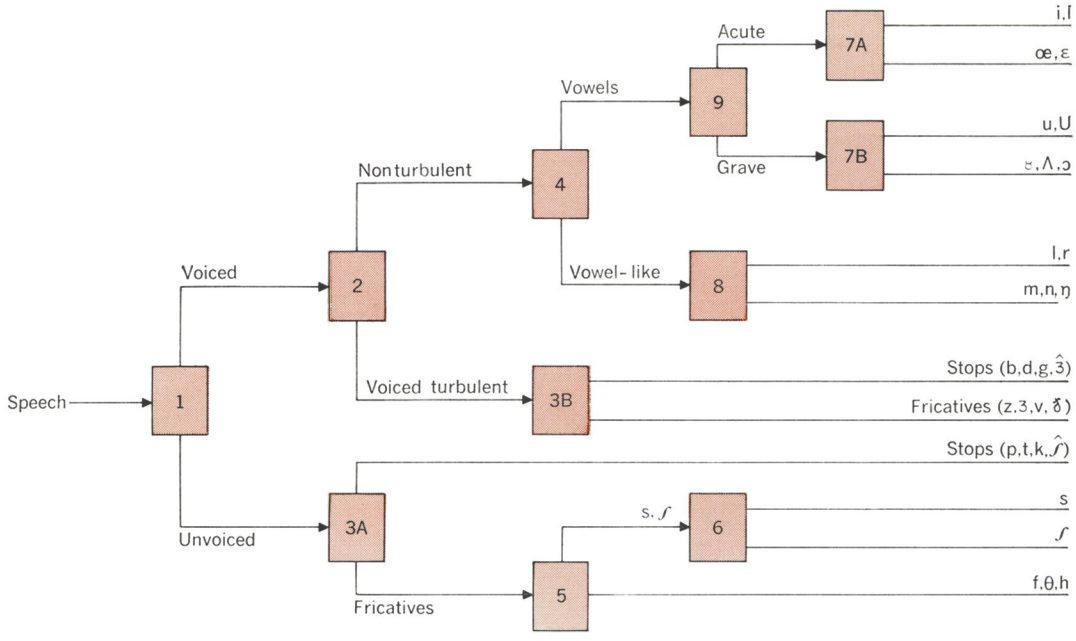


Fig. 5. Binary selection system for phoneme classification. Not all elements of this scheme were physically implemented.

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]

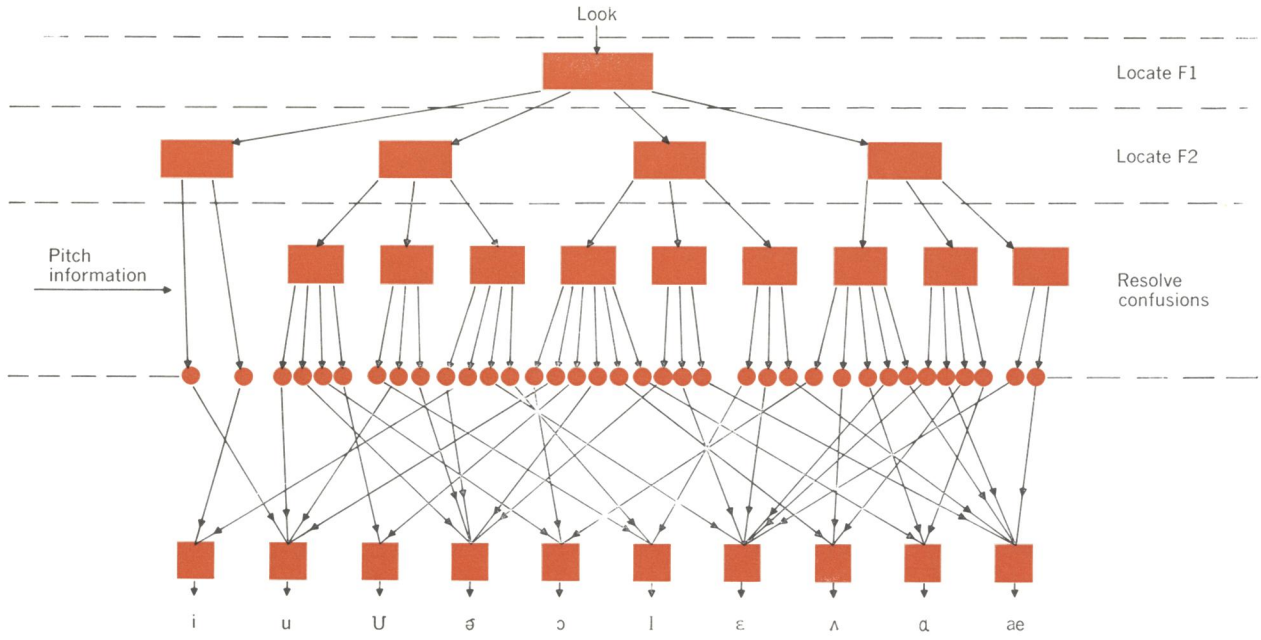


Fig. 6. General structure of the Forgies' vowel recognition program which classified each LOOK as belonging to one of the ten possible vowels.

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

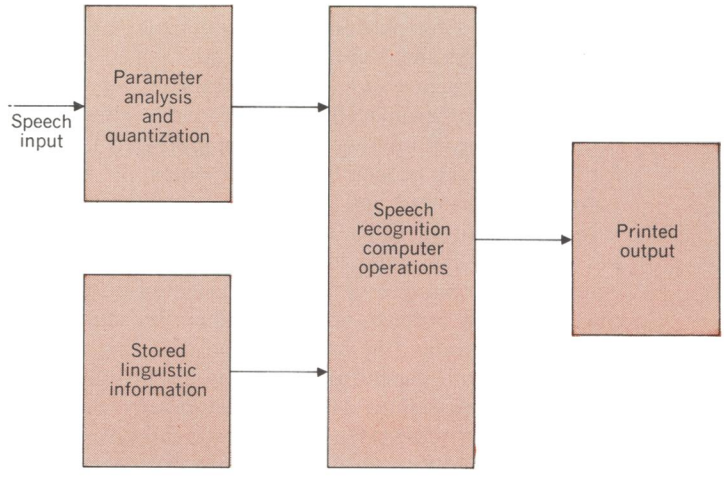


Fig. 7. Automatic speech recognition procedures must broadly take in the functions schematized here (after G. E. Peterson of the University of Michigan).

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

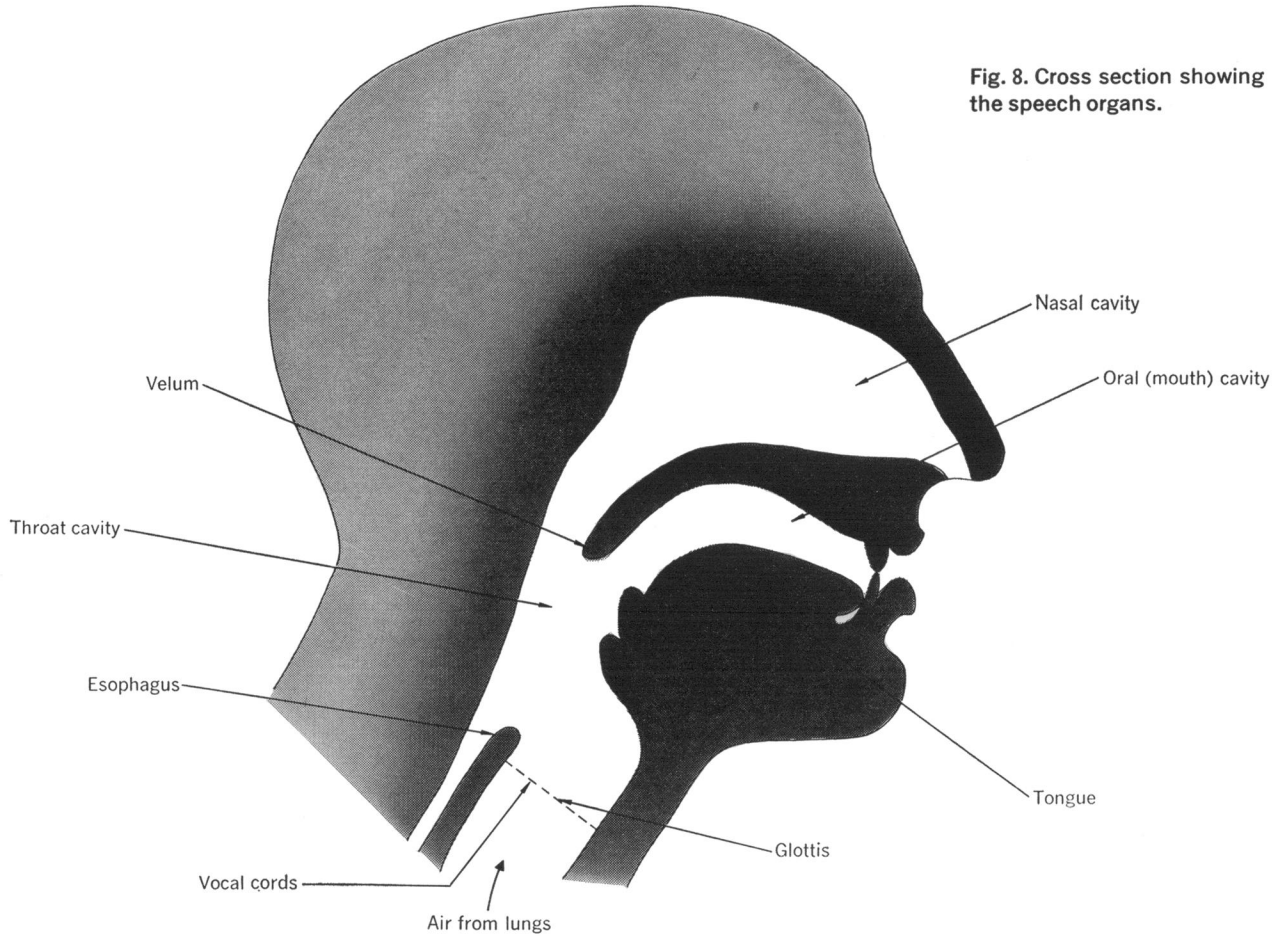


Fig. 8. Cross section showing the speech organs.

(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.

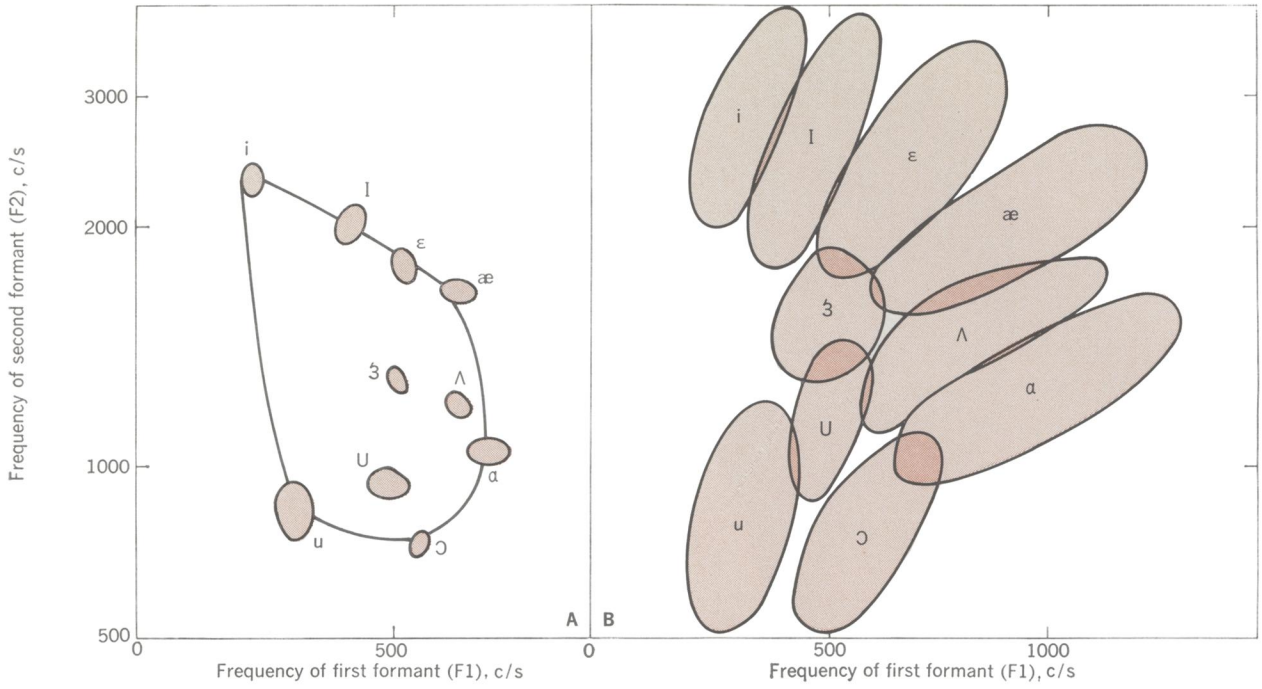


Fig. 9. A—Vowel sounds produced by an individual tend to form a fixed "vowelloop" in formant plots. B—These formant measurements made for a number of men, women, and children show a greater frequency spread.

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

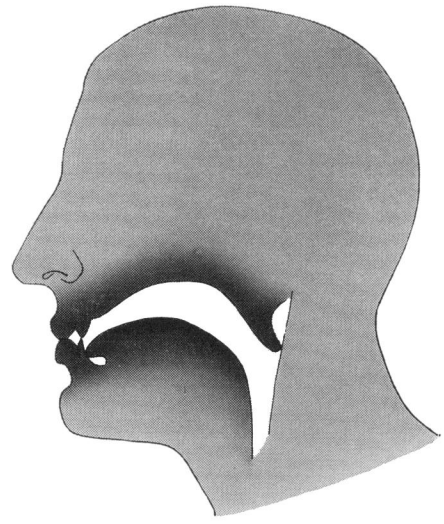
Vowel	Formant 1 Max	Formant 1 Min	Formant 2 Max	Formant 2 Min	Formant 3 Max	Formant 3 Min
u	480	210	1430	570	3300	1850
i	406	190	3100	2000	3900	2600
ɜ	652	360	2120	1130	2480	1400
ɑ	1040	592	1470	820	3180	2020
ɪ	534	206	2700	1710	3400	2340
ɛ	760	370	2570	1650	3300	2200
ʌ	910	550	1688	880	3250	1950

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.

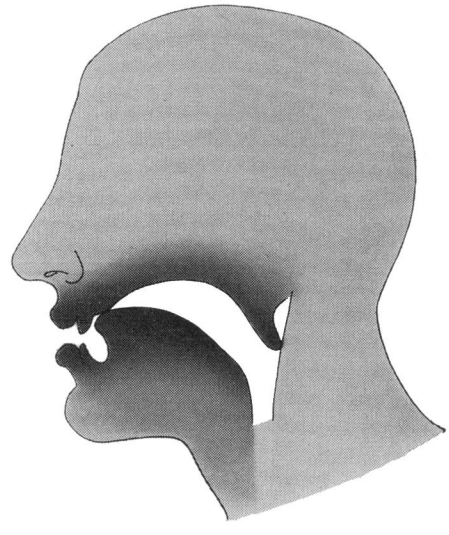
For the vowels, the vocal cords are usually vibrating (i.e., voiced), and the vocal tract is left relatively unimpeded. Different tongue hump positions, and rounding of the lips, produce the different vowels. The vowels usually have higher acoustic power than the consonants. Linguistic identification 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.

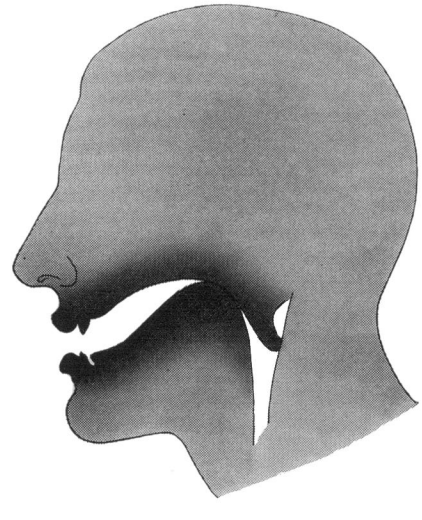
A predominantly turbulent air flow characterizes the fricatives, which can also be sustained. The air passes through a narrow opening at the front of the mouth and over the edge of the teeth. Vocal cords may or may not vibrate. For example, /s/ in *see* is an unvoiced fricative, while /z/ is *zoo* is 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.

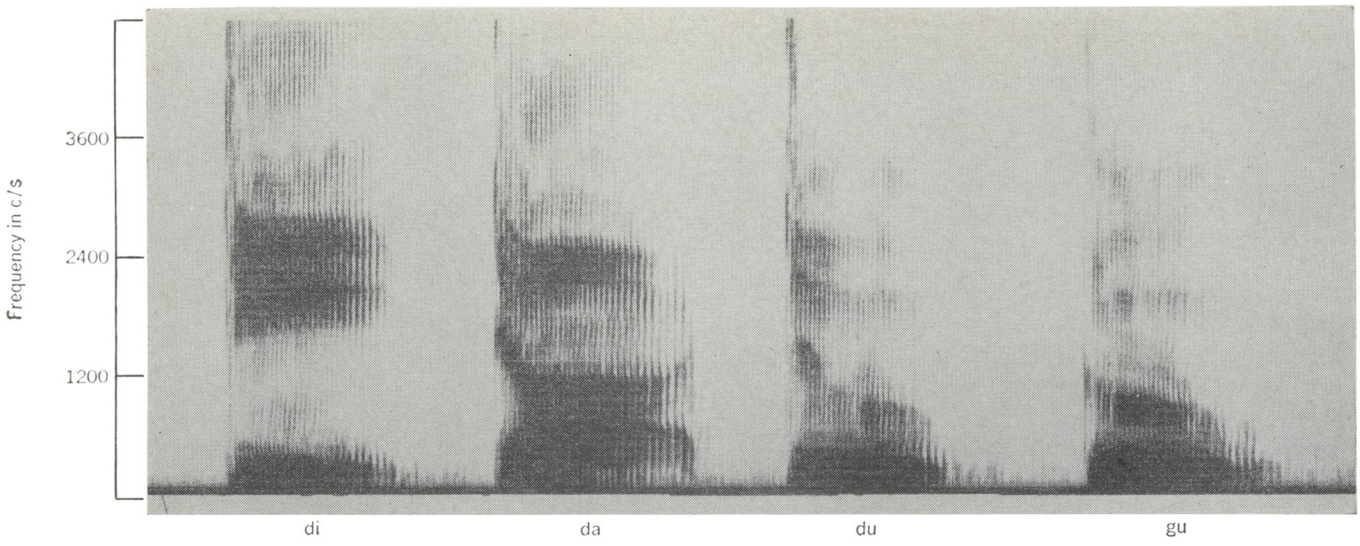


Fig. 11. Spectrograms of the sounds di, da, du, and gu, in which certain of the consonant transitions can be seen. These particular spectrograms, made under less than perfect recording conditions, give some indication of the problems a recognition machine might have in making identifications.

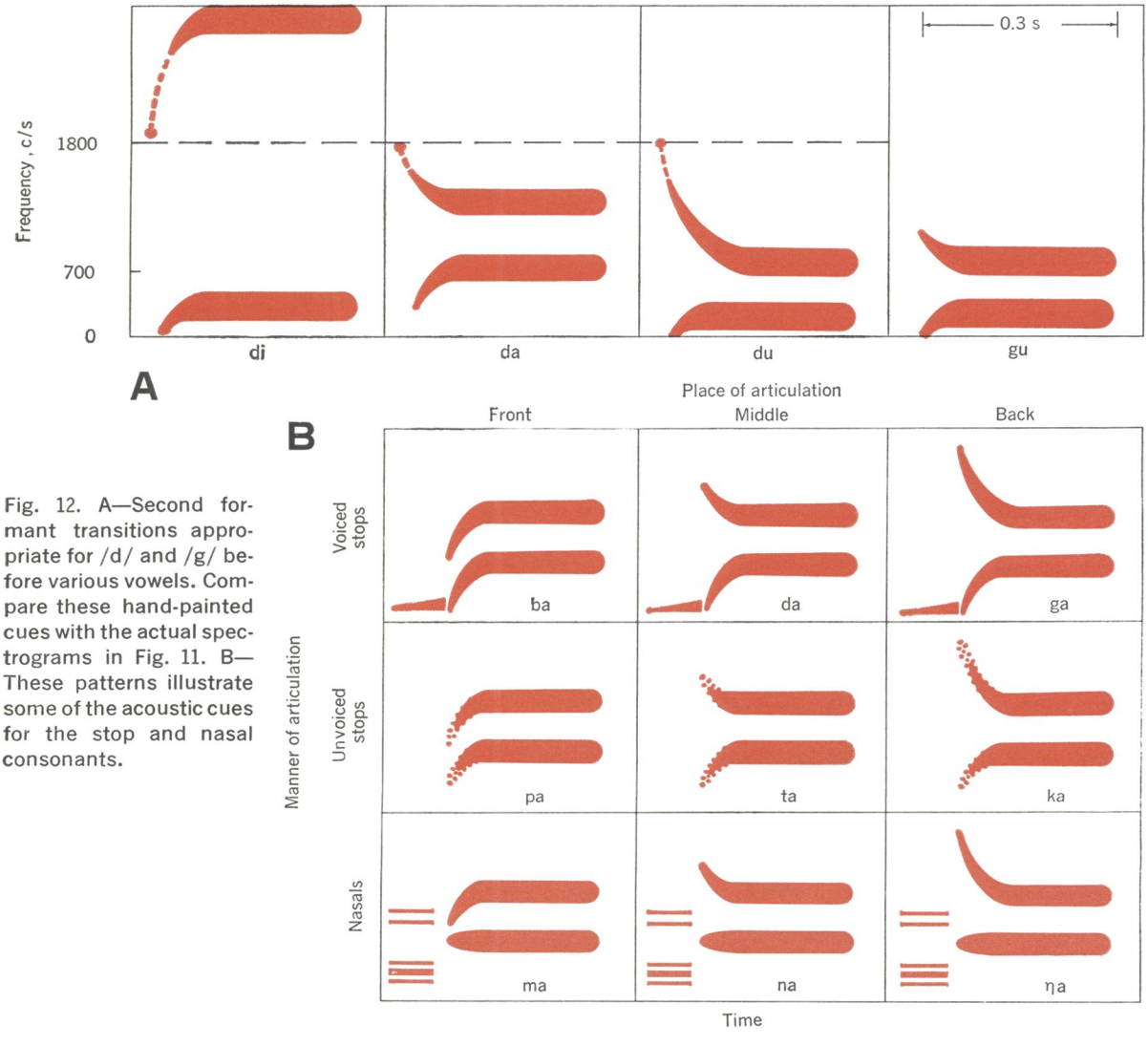


Fig. 12. A—Second formant transitions appropriate for /d/ and /g/ before various vowels. Compare these hand-painted cues with the actual spectrograms in Fig. 11. B—These patterns illustrate some of the acoustic cues for the stop and nasal consonants.

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

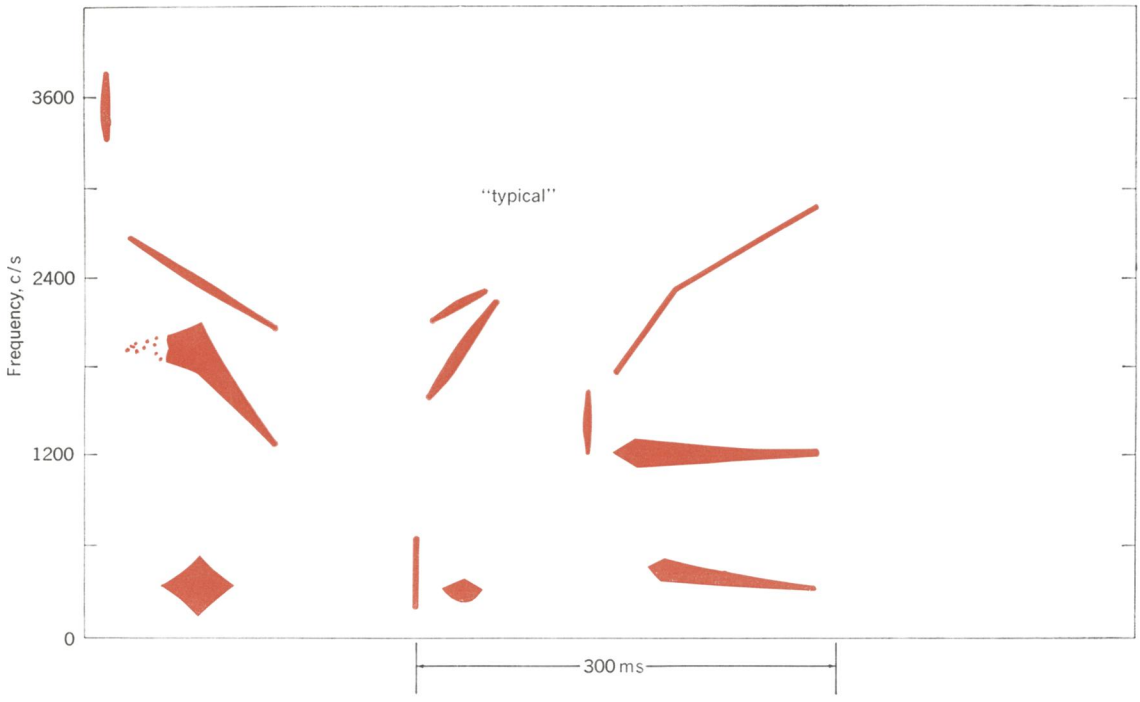


Fig. 13. Hand-painted spectrographic pattern for the word "typical."

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.

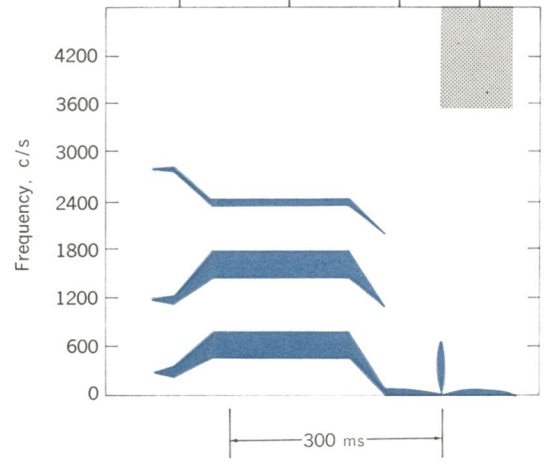
The other summation of acoustic research, for the ten years until 1957, is that of Pierre Delattre.[22] His paper provides an excellent view of the historical development of the work that led to the isolation of the many acoustic cues, which he breaks down for all the classes of sounds (fricatives, nasal 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.

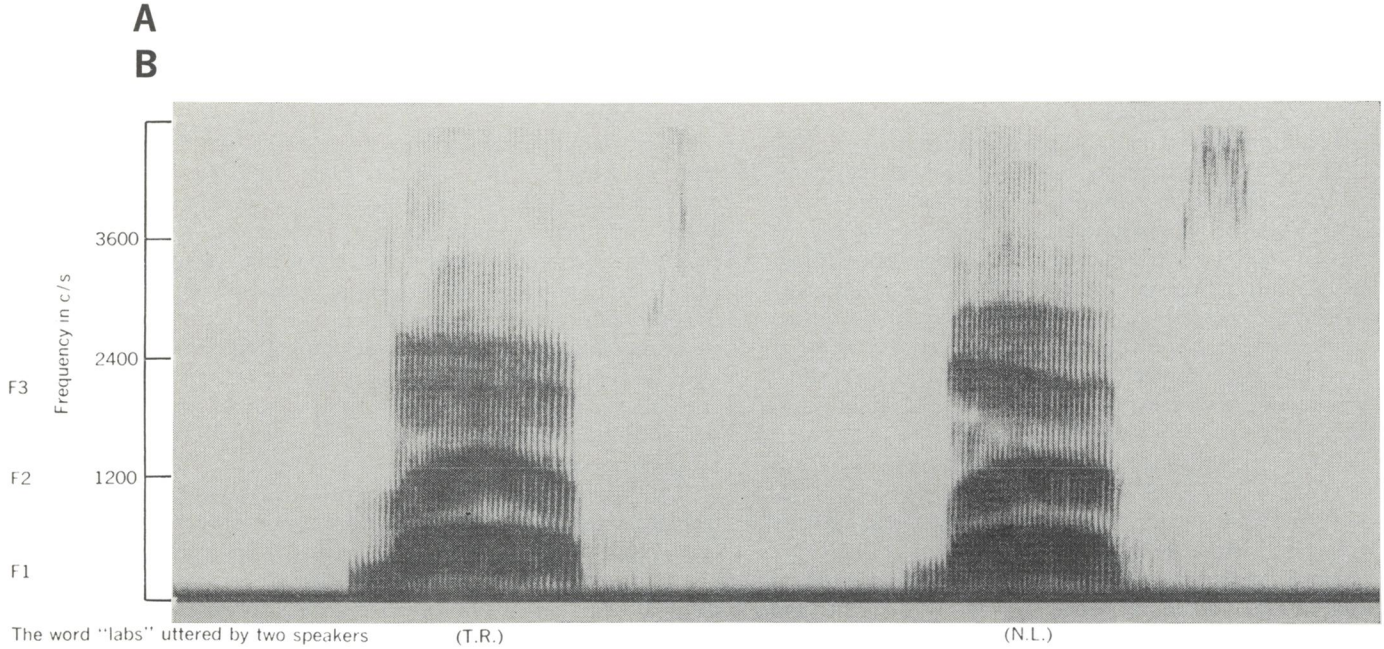
More recent papers in the important Haskins "opus" include a study of the effect of learning on speech perception (in 1961), which showed that there is an increased discrimination across phoneme boundaries,[48] an elaboration of their method of speech synthesis by rules[49] (in 1962), and a description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





The word "labs" uttered by two speakers (T.R.) (N.L.)

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.

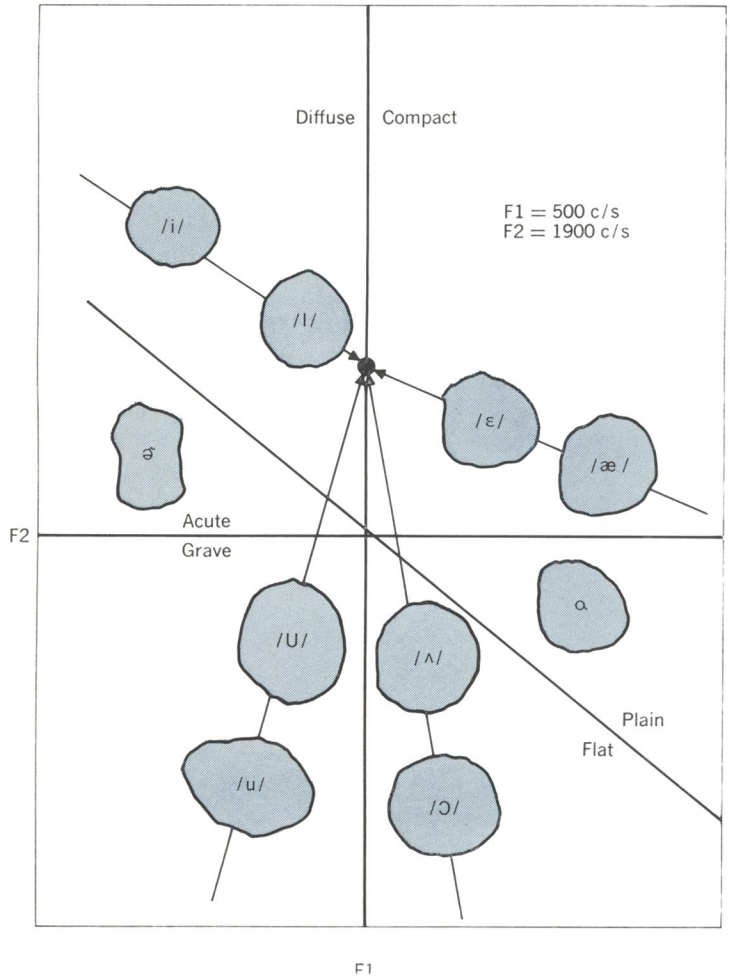
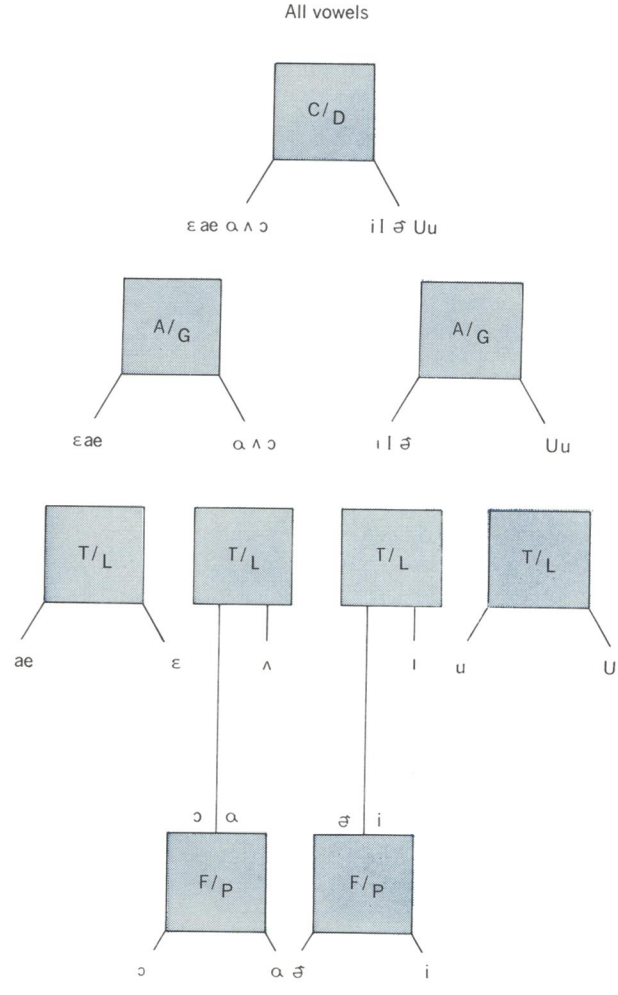


Fig. 16. Decision tree for the ten vowels of the Hemdal–Hughes program. Four binary feature-pairs were found sufficient to distinguish all these vowel sounds.



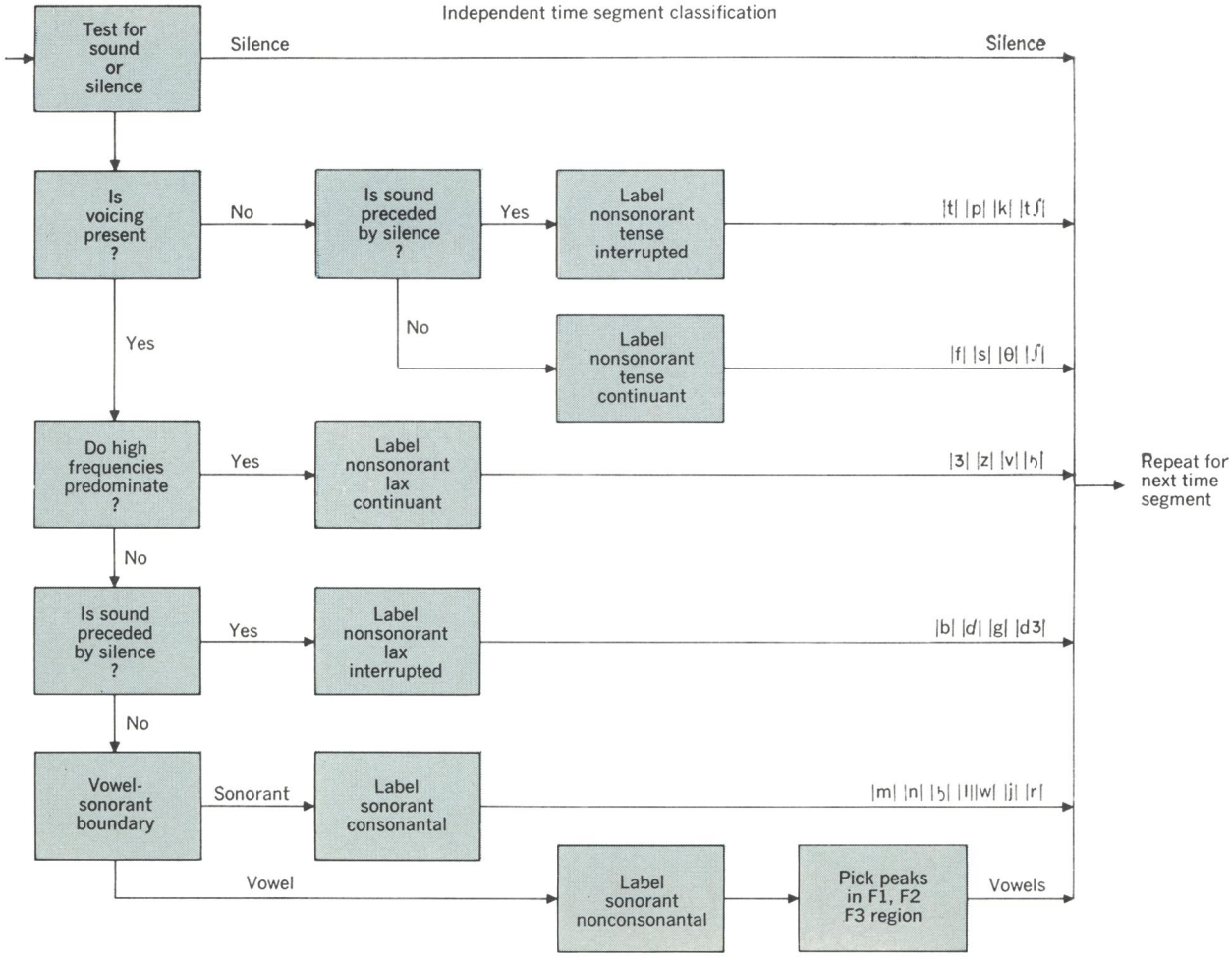
sufficient for identifying the compact/diffuse feature—F1 threshold was set at 500 c/s. For (3): a downward/-upward shift of a set of formants or all formants in the spectrum characterizes the flat/plain feature; thus, the physical correlate was determined by the sum of F1 and F2. For (4), the tense/lax 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.

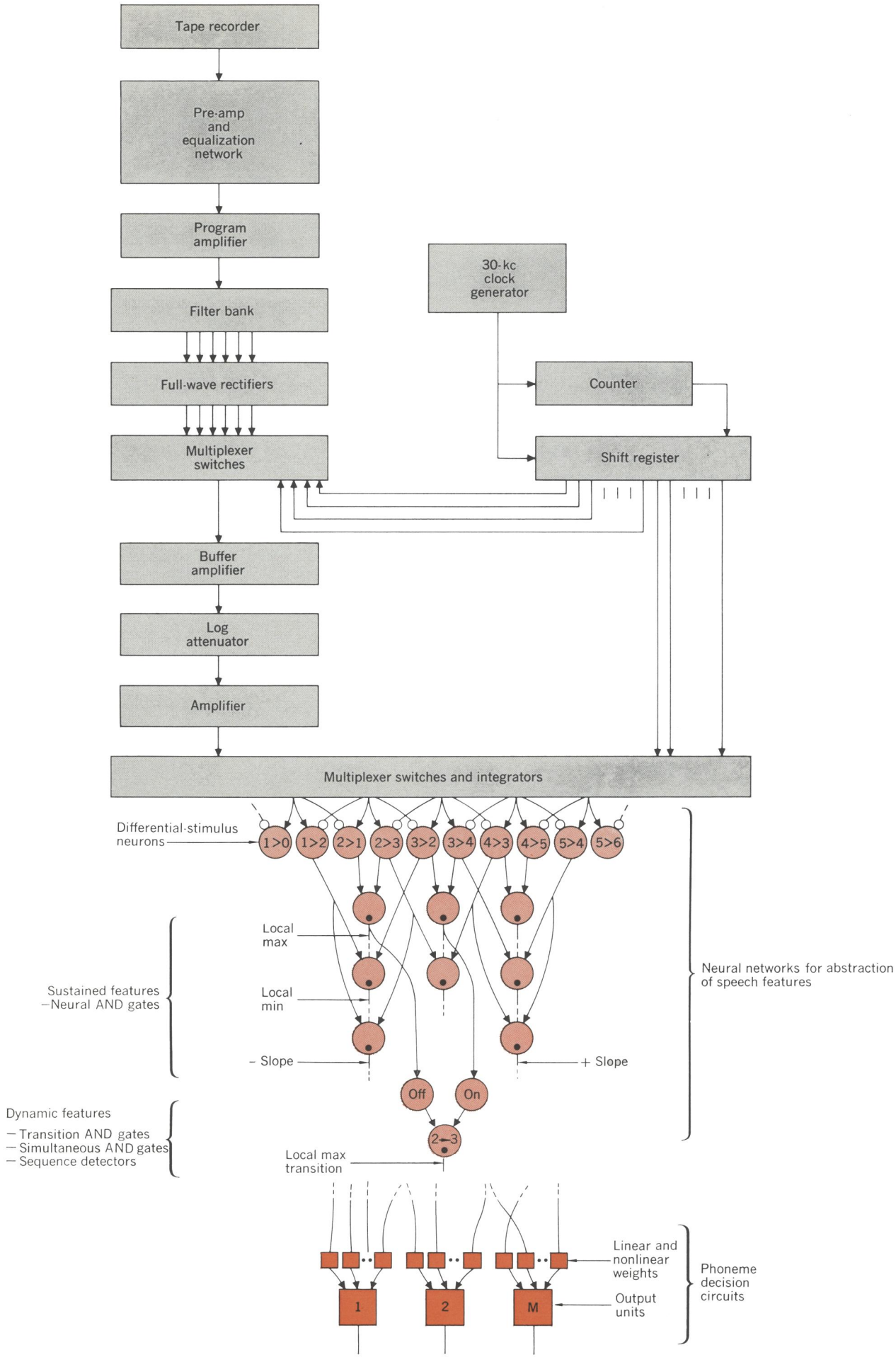
The Hemdal–Hughes decision tree for the ten vowels, in Fig. 16, shows how the four pairs sort out the vowels. An indication of how the computer program was set up appears in Fig. 17. Acoustic data are examined in either single time segments or in combinations of segments until the various consonants, 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, wed, 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]

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."

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.

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. Hemdal, 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.*