E. E. David, Jr.

Artificial Auditory Recognition in Telephony

Abstract: Machines which automatically recognize patterns from a stream of acoustic events, for example a spoken command, would have great utility in both communications and data processing. This paper reviews two applications of an elementary recognizer to the problem of actuating certain logical functions, and indicates how more ambitious recognizers might be utilized. In this regard, the automatic measurement of a talker's voice pitch and voicing dynamics appears fundamental to speech analysis, and hence to many recognition schemes. Visual inspection of spectral data taken from different speakers supports this contention.

Segmentation of speech into discrete units suitable for recognition, including the possibility of overlapping elements, is discussed. There is reason to expect that such segments will span several elementary speech sounds (phonemes). To illustrate this approach, a set of rules is presented for associating visual spectral displays (sound spectrograms) with the perception evoked by the corresponding utterances. These rules are specifically tailored for a limited vocabulary consisting of ten spoken numbers, and were validated by naive subjects who used them to identify the utterances of 33 people. In a further experiment, spectrograms of the same material from 14 talkers were simplified by reducing them to binary elements. It was found that master patterns for each number, compiled from the ensemble of talkers, could identify the utterances with over 99% success. These results emphasize a "diversity" approach to speech recognition which operates on relations between gross spectral features and does not depend exclusively on any one property.

The difficulty of achieving efficient communication between man and his machines has become almost legendary among communication scientists. The root of the problem lies, in all probability, in the diverse nature of the sensory mechanisms and logical organizations involved. Human beings are thought to perceive information at a maximum rate of about 40 bits/second, while machines can take in data thousands of times faster. On the other hand, the human memory is apparently many times larger than present machine memories. And it appears to be organized in associative units, thereby relieving the access and indexing problems. Even more important is the ability of the human observer to make meaningful interpretations of patterns of events. Men are adept at reading handwriting, understanding speech, identifying musical instruments by their sound, associating two-dimensional pictures with their three-dimensional counterparts, tasks at which machines are notoriously inept. Such abilities are commonly described by the term *pattern recognition* or *sensory Gestalt*. It seems quite clear that if machines could perform like functions, man-machine communication would be considerably aided. A corollary is that efficient man-machine coupling can lead to methods for efficient vocal communication between people.

I will not attempt a definition of *recognition*. As a poet once said when asked the "meaning" of one of his poems, "A poem shouldn't mean, it should be." Nonetheless, recognition tasks have several characteristics in common. Probably the most obvious is our colossal ignorance of how recognition is performed in the nervous system. Another common feature is that somehow both human beings and many animals are able to classify many diverse physical stimuli into the same category, each category probably being characterized not by fixed physical properties of stimuli but rather by certain relations between their parts. This property is certainly manifest in the case of auditory recognition, which is my principal topic in this paper. The same words spoken by a man and a woman differ drastically in their acoustic content, but the listener has little difficulty in establishing that they are the same words. Clearly, there is some sort of perceptual transformation taking place. At the periphery of the nervous system, namely at the auditory nerve endings

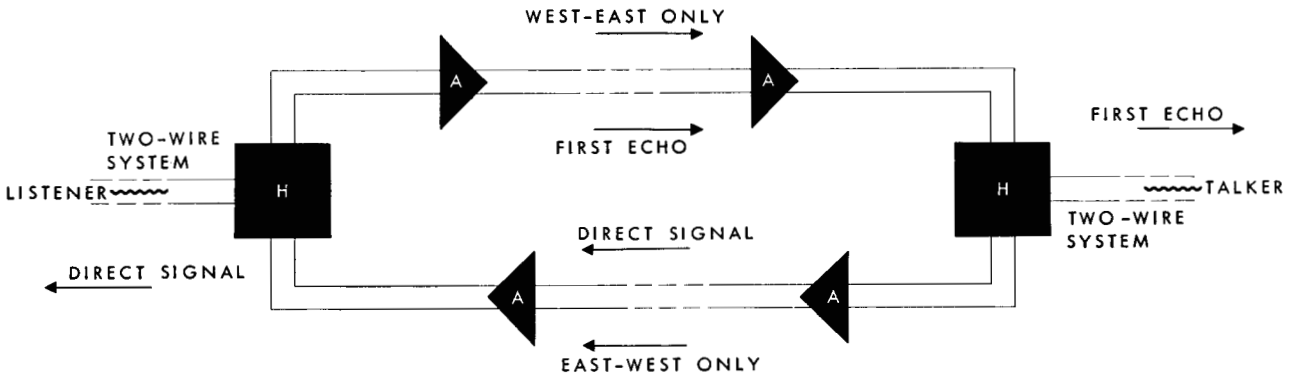


Figure 1 **Echo production on a telephone circuit.**

in the ear, these two disparate acoustic disturbances must excite different patterns of nervous activity. Somehow, farther along in the chain of events, in the brain, these patterns assume some form of equivalence. In fact, once the two sets of words are recognized as being the same, there must be a similar close neural correspondence. In any case, the essence of recognition lies in the perceptual realm.

We might expect problems of auditory recognition to crop up often in telephony. Speech is the major traffic through the telephone system — a system which has a logical as well as a transmission function to perform. In the early days of telephony, people had to deal only with other people. Switching and accounting functions were accomplished manually after voice or written instructions. When automatic switching was introduced, the customer was fortunately already supplied with a finger which was pressed into use for signalling to the deaf central office. Automatic message accounting, as it exists in today's telephone plant, takes advantage of the anatomy of the telephone network as well as the dialed information. But to handle the diversity of requirements inherent in today's communications business, people are still distributed throughout the network. Whenever one of these people must interact with the system, inefficiency and errors often result. These points of friction would be relieved by devices which could read characters or recognize the sounds of speech.

In this context, it is not surprising that some of the first modern attempts to automatize speech recognition were made in conjunction with telephone research. I would like to recount for you two applications of a very simple artificial recognizer and review the rather startling results it has spawned. The implication is that machine recognition can contribute importantly to efficient vocal communication between people. Indeed the so-called vocoder principle, as I will relate, forms a ready vehicle for the application of recognizers, be they elementary or sophisticated, in voice communication. On the elementary side, there are functions such as the detection of vocal pitch and inflection. I will try to point out the relevance of the pitch detection problem for communication. Automatic word recognition typifies a more ambitious class of problems. I will present the results of the two studies in word recognition which, though they were done some years ago, have never been reported in the literature. The philosophy guiding these studies is illustrative of what we feel is a fruitful approach to speech recognition. Finally, I will return to the pitch detection problem and indicate why I believe it is fundamental to speech analysis and speech recognition.

The telephone subscriber's interaction with the telephone system is not limited to dialing his connection, speaking occasionally to an operator, and paying his bill. When he talks over a long two-way circuit, often his voice is called upon to perform a switching function. This requirement arises because of an unfavorable interaction between the talker and the circuit, Fig. 1. A long telephone line, comprising a pair of wires in each direction and terminated by a hybrid or bridge, tends to produce an echo when a person speaks over it. The hybrid is intended to couple the directional pairs to a bilateral single pair, at the same time preventing any interaction between the directional pairs. The hybrid balance is not perfect, so for instance, when the talker at the right speaks over the East-West pair, the hybrid on the left, in addition to passing the direct signal to the listener, produces an echo on the West-East pair. This echo appears at the talker's own earphone. The hybrid on the right produces a similar effect and eventually a second echo results. This effect might appear not to be a major one from the standpoint of communication. However, laboratory experiments have shown that if a person's voice is reproduced in his own ears at a normal hearing level and with a delay of a few tens of milliseconds, his vocal mechanics can be seriously interrupted. He tends to stutter and becomes hesitant in speaking; some speakers are unable to produce any connected utterance at all. The effect increases, up to a point, with both time lag and echo strength. While in commercial service the echo is always greatly attenuated in level, it can still be annoying if its delay time is long. One remedy for this situation is merely to disconnect the return line when the speaker is talking. This task can be carried out by voice-operated switches known as *echo-suppressors*.[1] These switches are arranged as shown in Fig. 2. When the party at the left talks, his receiving line is disconnected, and similarly for the other party. If both parties talk, one of them "cap-

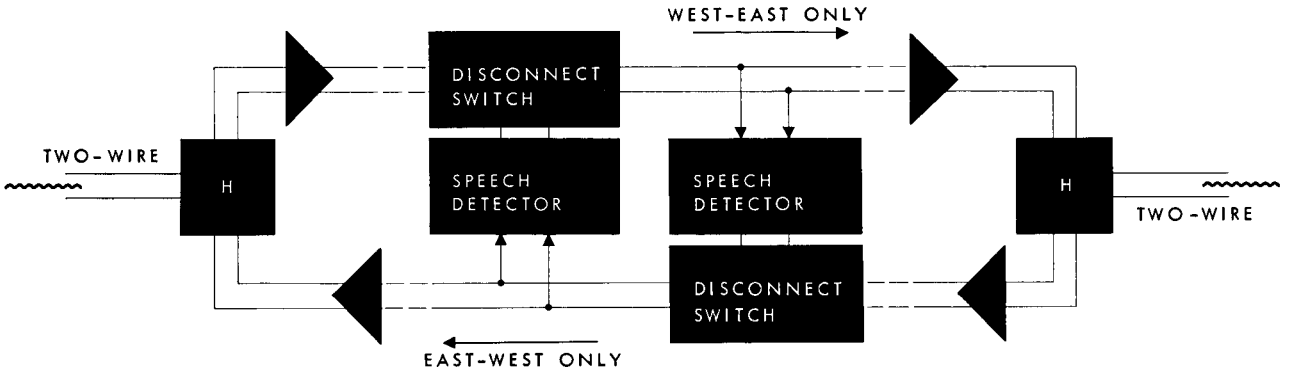


Figure 2 **Voice-operated echo suppressor on a telephone circuit.**

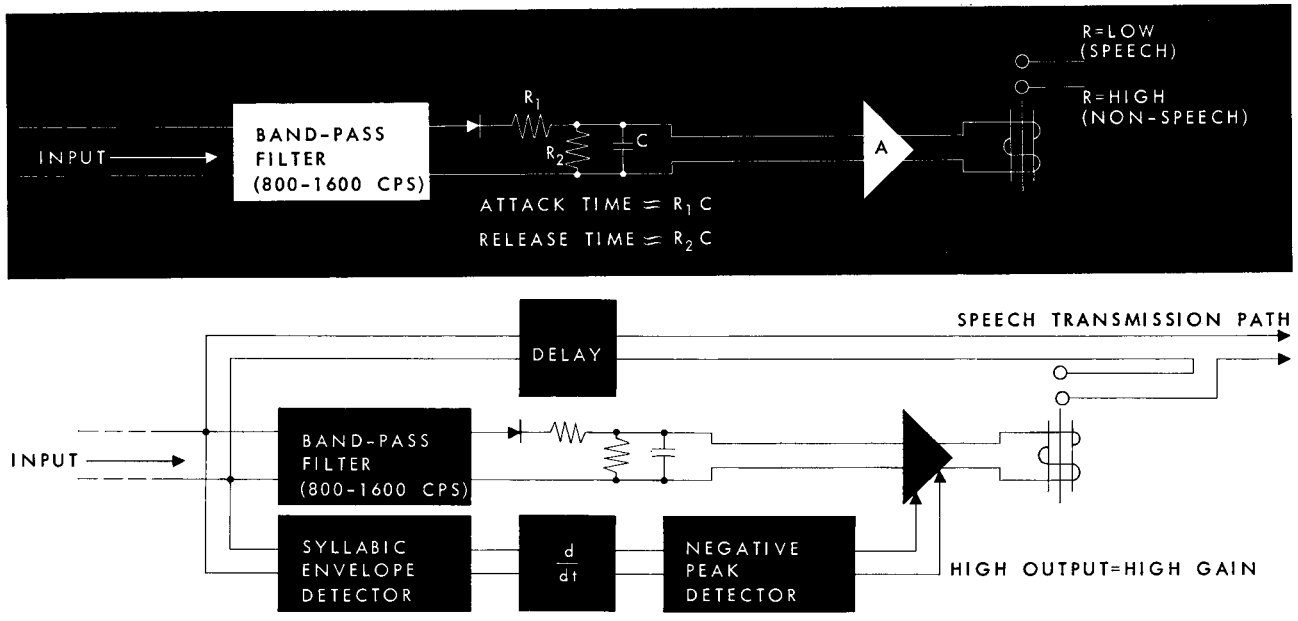
tures" the circuit, locking the other out.* Which one depends upon the time sequence and strength of the two talk spurts.

It is necessary for an echo-suppressor to recognize when speech is being transmitted. It must be able to discriminate between speech and the variety of other noises which might appear on the transmitting line. These include electrical disturbances picked up in the line or generated in the system, as well as environmental noise at the speaker's location. This rather primitive recognition function requires that all signals be classified into one of two categories, speech and nonspeech. Traditionally this discrimination has been accomplished on the basis of power spectrum alone. Fortunately, electrical interference and room noises tend to be either predominantly low frequency or impulsive, and can be suppressed to some degree, at least, by filtering and smoothing. The echo-suppressor recognizer therefore is merely a threshold-operated detector preceded by an appropriate filter, as shown in the upper part of Fig. 3. The relative frequency distribution of speech and noise dictates a pass-band of about 800-1600 cps for the filter.

Of course, if the noise level on the transmitting line is high, then the switch may err and disconnect the listener from an incoming message. Obviously, this is more likely to occur if he is listening from a boiler factory than from a sound-treated office. The threshold-adjusting mechanism shown in the lower part of Fig. 3 can help this problem. This circuit derives the speech volume envelope and examines its average decay rate. Typically the speech envelope shows sharp peaks and valleys as the various sounds succeed one another. By using the rapid decays as a cue, the speech-detector can discriminate against any noise not showing these "syllabic" envelope fluctuations.

*In modern practice, echo-suppressors are designed so that an absolute lockout is never possible. Rather than breaking the receiving line completely, the available gain is apportioned between the two lines so as to give one preference over the other.

Figure 3 **Automatic threshold-controlled speech detectors.**



Just this same characteristic has been used to realize a *commercial-suppressor* for use by classical-music fans.[2] This device discriminates between speech and music, allowing only the latter access to the listener's loudspeaker. Music has fast attacks but slow decays.

The second application of this rudimentary recognizer is in a switching system to take advantage of the one-way characteristic of conversations. Most people, especially men, are satisfied to listen while the other fellow is talking, and we expect him to listen when we talk. Indeed if there is a switch-type echo-suppressor on the line, voice signals can travel only one way at a time even in the case of female conversationalists, and the other circuit is idle during half the conversation. In modern communication systems there are often many circuits in both directions, while the system as a whole handles many conversations. Now clearly the probability of all the talkers at one end speaking simultaneously is quite small. Thus, there are on the average free circuits in both directions all the time. A sufficiently intelligent switching system at each terminal might make use of these idle pairs for further conversations, thereby increasing the traffic capacity of the system. This method of increasing circuit occupancy is known as Time Assignment Speech Interpolation or simply TASI. Just how many additional talkers can be accommodated depends upon several factors; (1) the number of independent circuits available in each direction, (2) the fraction of time that each person is speaking (the speech activity), and (3) the allowable "freeze-out" probability (the probability that a talker will find all circuits occupied when he starts to talk), which in turn is related to the fraction of all talkers' speech which will be frozen out. For instance, in the theoretical limit of a large number of conventional circuits between two terminals, the TASI gain is the reciprocal of the average talker activity while the freeze-out probability approaches zero.

Speech detectors, one on each input line, can provide the sensory input to TASI as indicated in Fig. 4. They are able to sense who among the potential talkers is "active" and needs a circuit, and also when a talker is inactive and can be disconnected. The output of the speech-detector can be an on-off signal to operate suitable logic in the switch control.

One possible logic provides that when a particular speech-detector indicates the presence of speech on its line, the programmed TASI switch connects the line to a transmitting circuit, continuing to operate on incoming voices until all circuits have been filled, at which time a freeze-out occurs if another voice comes in. When a talk spurt terminates, the corresponding line becomes available for another talker. At the receiving terminal, the automatic switching device must know who is talking over which circuit, so as to connect each talker to the correct listener. These data might be passed as a very short identification signal before each talk spurt, or alternatively over an auxiliary channel which would serve the entire bank of circuits. These signals can be generated automatically by the switch control.

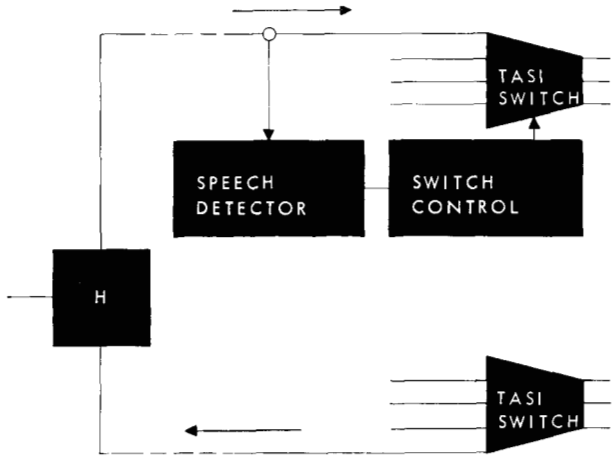


Figure 4 **Voice-operated logic.**

As you can see, the speech-detector is the heart of TASI; which, according to calculations, might multiply the number of available circuits by two or three times. This, indeed, represents a rather substantial dividend in efficient communication from a simple device, the threshold-operated speech-detector. Thus, in concept at least, the utility of devices to emulate human sensory functions seems firmly established. In fact, many of us feel that one link in an ultimate human-communication channel must be such a recognizer. For example, communication between people at a distance might ideally be performed by a man-recognizer link at the transmitter followed by a functionally inverse link at the receiver. The voice-coding philosophy originated by H. W. Dudley incorporates just such features.[3]

Dudley proposed several devices, known as vocoders, whose recognizer-analyzers might be likened crudely to the ear and brain of a listener, and whose receiver-synthesizers might be thought of as analog vocal tracts. This thought is illustrated in Fig. 5.

Ideally, such a system analyzes the talker's voice, selecting from it certain perceptually important features, transmitting a coded description of these features to the receiver, which then reproduces the talker's utterance. It is as though a man were separated from his vocal tract by some advanced medical technique, and the nerves controlling the tract elongated into a cable. The ear and brain reside at the receiving point while the vocal tract is transported to the message destination. To transmit a message over this system, we talk to this man and he repeats what he hears.

In an elementary vocoder, the features to be described by the transmitted signals might be perceptually important components of the speech frequency spectrum. In a more ambitious vocoder, the signals might denote the sounds of speech or the words from a specified vocabulary, these having been recognized at the analyzer or coder. In this case, of course, the signals could be made to actuate a voice typewriter. We might picture this function as being accomplished by a pair of hands controlled by the "motor" output of the coder. A less

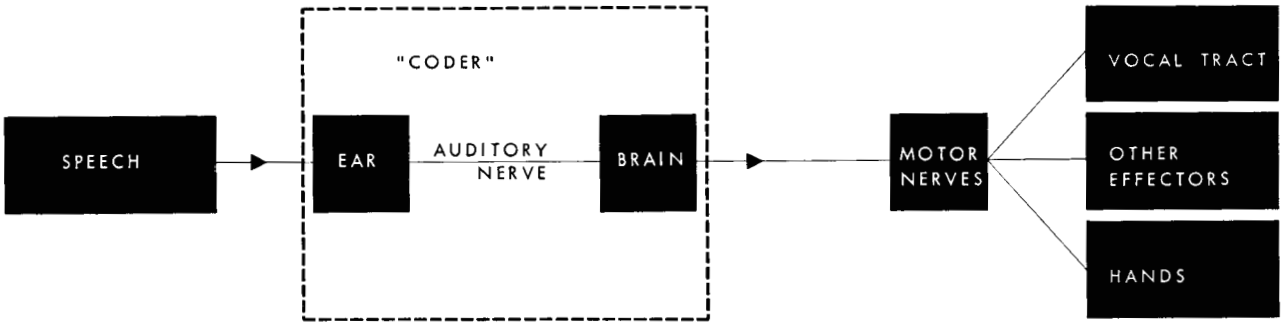


Figure 5 **Representation of efficient speech transmission.**

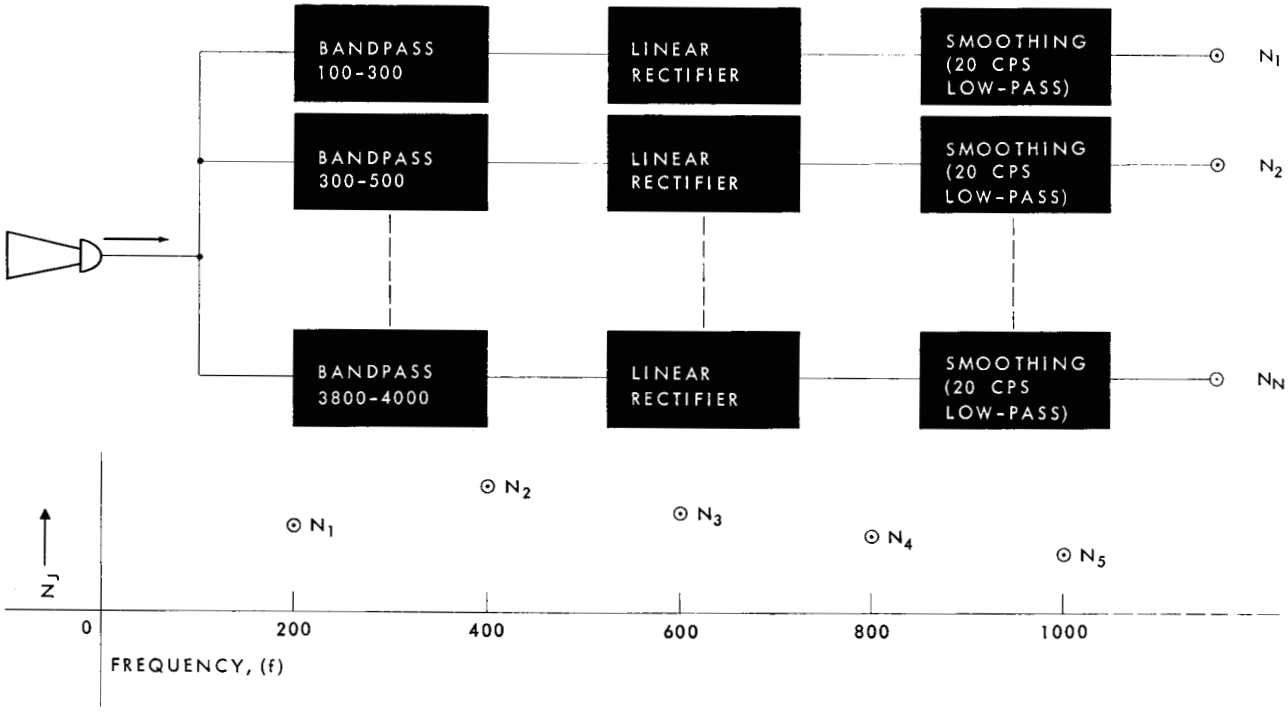
demanding task might require certain actions to be taken on voice command. Again, these actions could be carried out by appropriate effectors.

The utility of vocoder arrangement lies in the fact that an analog vocal tract, or a typewriter, or the other effectors require relatively little direction. The necessary control information for speech production, for instance, can in theory be packaged into something less than 50 bits/second. Speech production with such a low rate input has been demonstrated many times, and indeed is probably demonstrated each time a human being speaks. Thus, such a coding of speech information leads to a highly efficient use of transmission channel space in a communication system.

The simplest vocoder is the so-called *channel vocoder*. Its analyzer is shown in Fig. 6. The vocal tract control data are derived by passing the speech through a bank of band-pass filters, and by measuring at each filter output the energy of a succession of suitable time intervals. At any one time, then, the spectrum is represented by N numbers, if there are N filters, each approximately representing a sample point on the spectrum as indicated in the lower part of the figure. The succession in time of such number sets reveals the dynamic progression of the short-time power spectrum.

These data can be displayed for visual inspection by the so-called *visible speech* technique.[4] Here, in effect, a large number of overlapping filters provide the spectral numbers which are projected into a density display on a frequency-time plane. Such a display is shown in Fig. 7. At any frequency-time intercept, the density of the display depicts a spectrum sample number. A phonetic transcription is shown along the time axis. The vocoder, in effect, transmits an approximate description of such a

Figure 6 **Spectral analysis in a vocoder.**



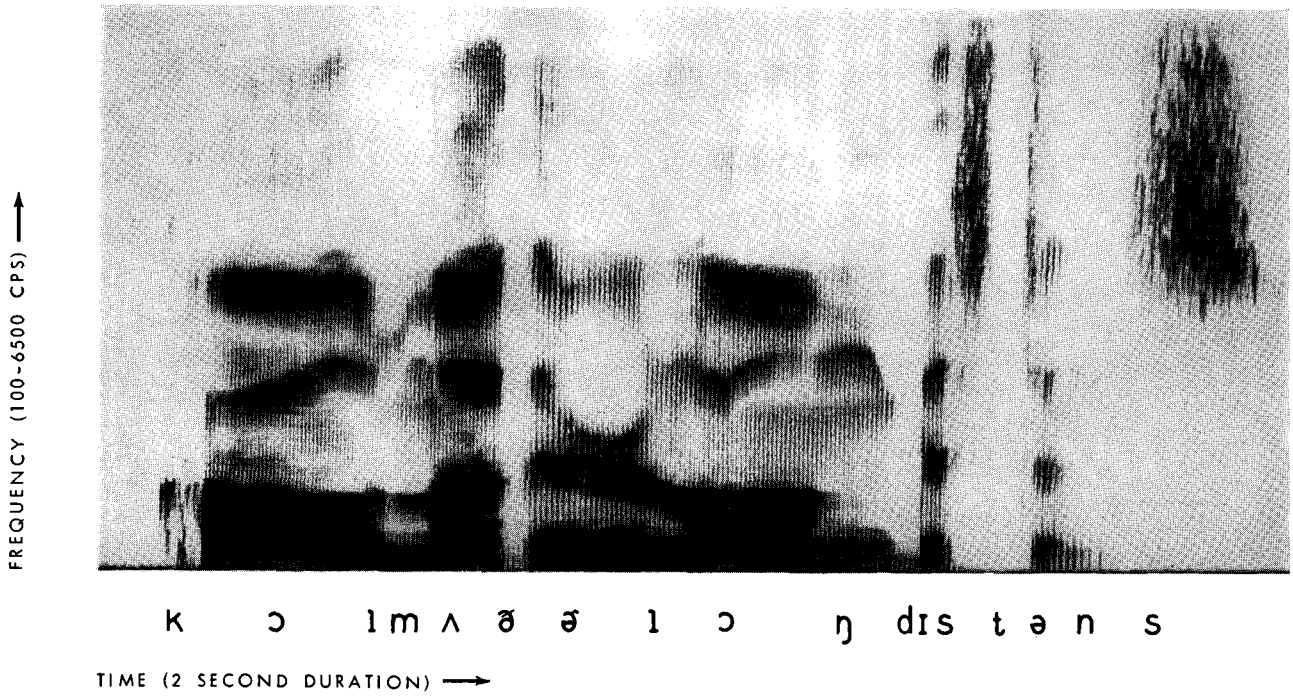


Figure 7 **Sound spectrogram of male speech.**

sound spectrogram. Thus you are looking at the information available to the vocoder speech synthesizer. In reproducing the input speech, the vocoder must regenerate a signal with a spectrum to match the succession of numbers measured by the analyzer. The method commonly used for this is merely to superimpose this characteristic on a flat spectrum generated at the vocoder synthesizer. Figure 8 shows this process. The energy from the generator is divided into narrow frequency bands, and the amount of energy in each is adjusted dynamically by a modulator in accordance with the analyzer measurements. The sum of all these channels has approximately the same spectral density as the input to the analyzer.

As I have described it, the vocoder performs no recognition function—the spectral data have no unique perceptual correspondence. Many different distributions can give rise to the same percept. Yet recognition does play a vital role in the vocoder, as can be seen in the following discussion.

The energy produced by a talker in generating his speech can arise from two distinct sources, namely the vocal cords and turbulent airflow at a constriction in the vocal tract. The cords produce an approximately periodic excitation commonly found in *voiced* sounds such as the vowels. The fundamental frequency of the excitation corresponds to and determines the pitch of the voice. Turbulent airflow in the vocal tract produces a noise-like disturbance such as is found in the *s* and *sh* sounds. These *unvoiced,* or *voiceless* sounds have no pitch. Thus if the synthesized speech is to assume the talker's voice pitch and express his individuality, a flat spectrum of either noise or periodic pulses, whichever is appropriate at the time, is required. In addition, the repetition rate of the pulses must be the same as that of his vocal cord pulses. Thus, the microstructure of the flat spectrum onto which the general spectral features are superimposed must be predetermined at the analyzer, and this information transmitted to the synthesizer to control an artificial vocal source.

Figure 9 shows the vocoder analysis and synthesis in outline form while the voicing circuit is detailed. The vocal cord or voiced energy is provided by a harmonic tone generator of the appropriate fundamental frequency, sometimes called the buzz oscillator, and the unvoiced energy comes from a noise source which generates the so-called hiss sound. Switching between these sources is done by a relay. The analysis function to control this arrangement is shown on the left and requires recognition of the talker's voicing dynamics. Present methods for identifying voicing intervals in speech depend upon the over-all distribution of speech energy. During voiced periods, energy tends to be concentrated in the 200-800-cps band, while for unvoiced sounds the principal energy resides usually above 1500 cps. This differential is utilized by the threshold detector whose output causes the synthesizer relay to connect the buzz source. The hiss source is connected when the buzz source is not. To control the fundamental buzz frequency during voiced sounds, the lowest component of these sounds, the voice fundamental, is selected by a filter and its zero crossings counted. Such devices are quite vulnerable to predistortion of the spectrum or to added noise. Furthermore, these disturbances often occur in reverberant or noisy environments and in microphones and transmission circuits which precede the

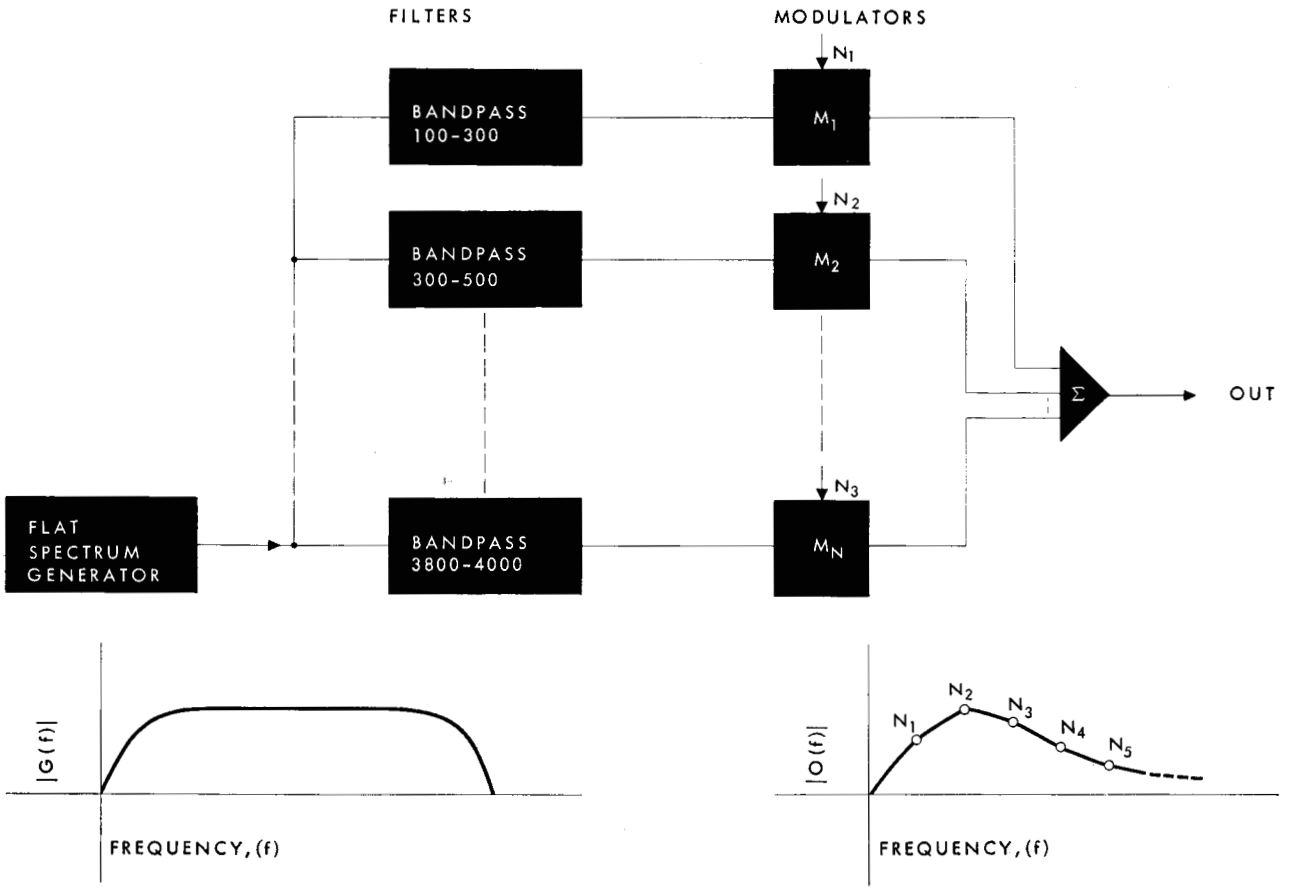
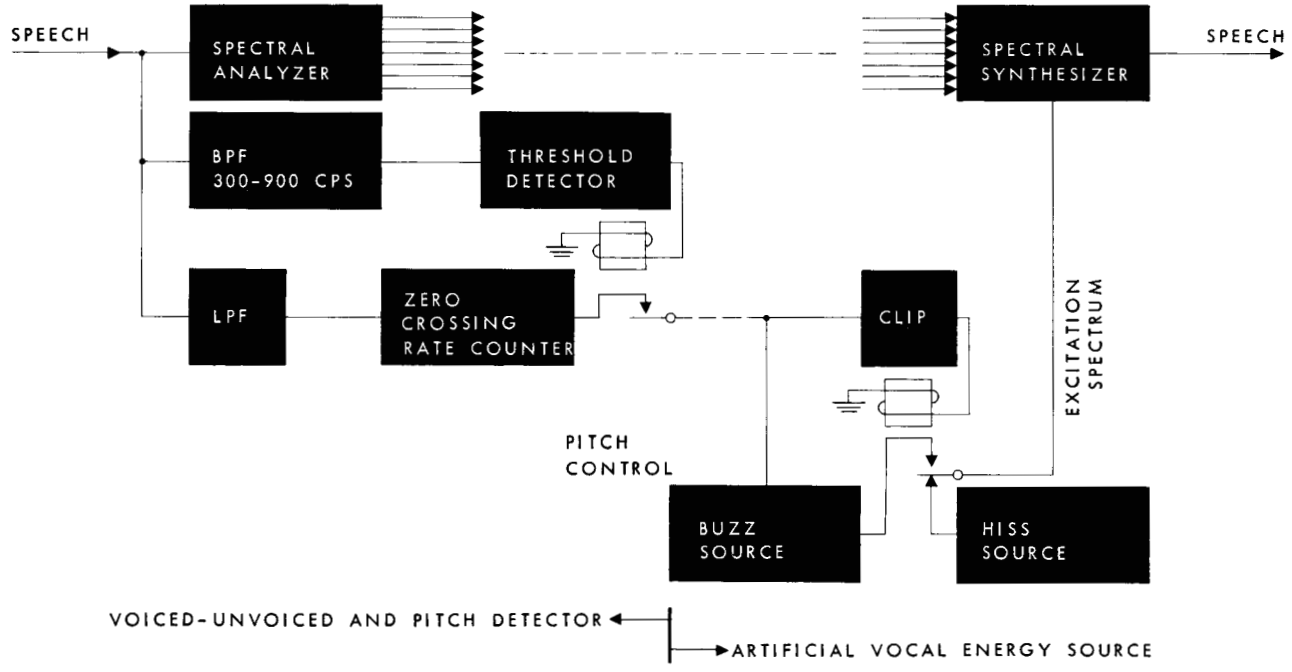


Figure 8 **Speech synthesis in a vocoder.**

Figure 9 **Voicing and pitch detection in a vocoder.**



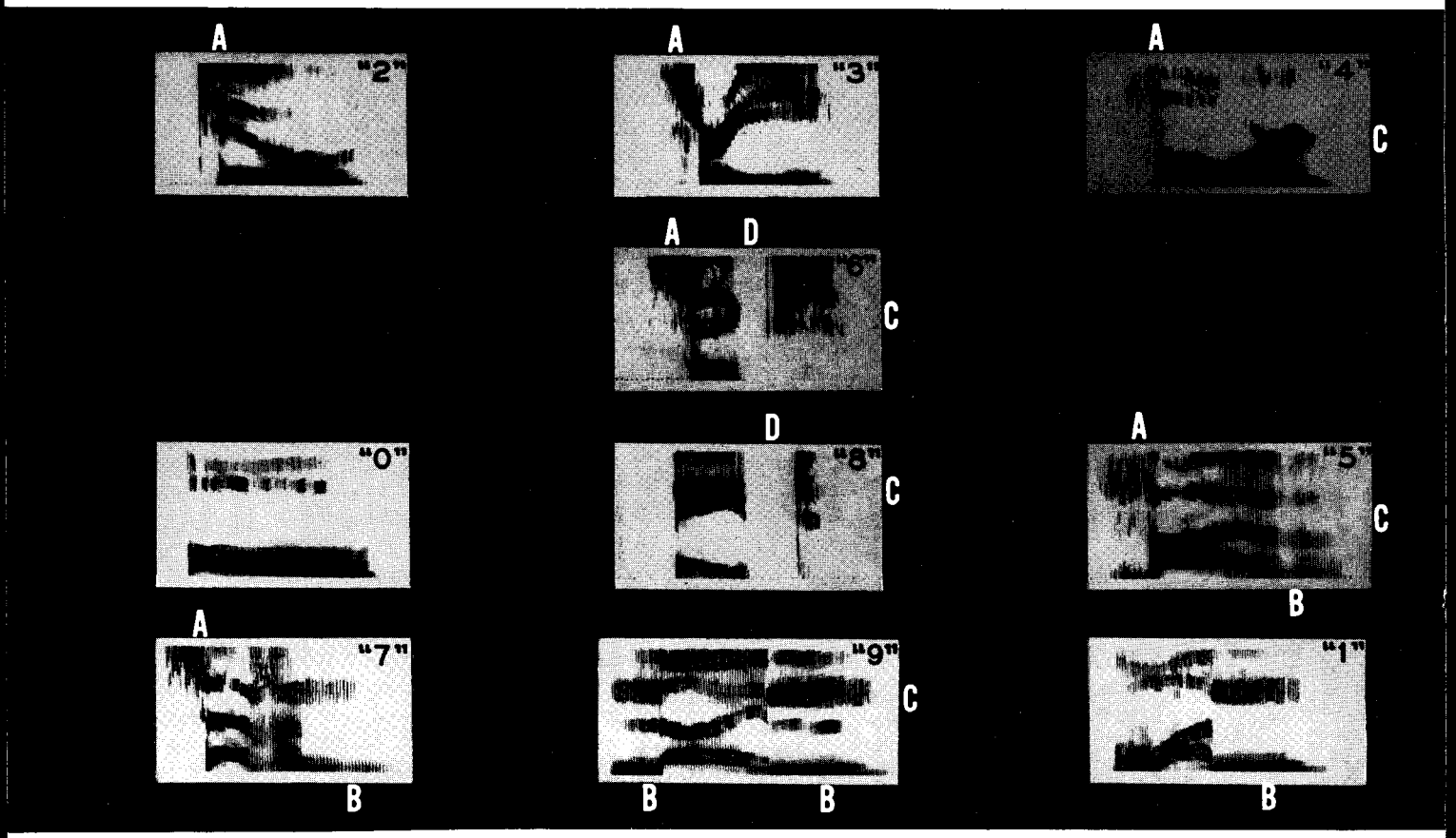
vocoder analyzer. The human analyzer on the other hand is more resistant to the adverse effects of these distortions, as has been shown by speech intelligibility tests. For instance, if white noise is added to speech only, say, 10 decibels below the speech level, listeners can still discriminate *vat* from *fat* and *bit* from *pit*. These discriminations require the voiced-voiceless distinction.[5] Needless to say, present voiced-unvoiced recognizers cannot emulate this feat.

A basic solution to the voicing recognition problem would be a real contribution to the speech-recognition field generally. For, as I will point out, it seems to be a prerequisite to speech analysis. From the practical point of view, it is this problem, in large measure, which has retarded the application of vocoders in the communication field, an application which would save a factor between 10 and 30 in the channel capacity necessary to transmit articulate speech. Quantitatively a reduction from 60,000 to about 2,000 bits/second would be well within reason.

Even greater economies could be achieved if recognition principles could be applied to the representation of the general spectral features of speech. Such features are displayed prominently on sound spectrograms of speech as shown in Fig. 7. The sentence pictured here says "Call mother long distance," as is indicated beneath it in phonetic script. Notice that during the voiced sections, for instance just after the *k*, the speech energy tends to be concentrated into three or four bands, usually called *formants*. Other sections can be characterized by a time sequence of events such as around the *s-t* in *distance* and still others by specifying crudely a spectral envelope such as the final *s* in *distance*. If the vocoder analyzer could recognize such features, it could derive an extremely compact spectral representation, one that could be passed over a channel of the order of 150 bits/second capacity.[6] Just how each of these features contributes to the perceptual value of an utterance is not completely understood. However, people can learn to interpret spectrograms in terms of sounds and words. The learning process is long and laborious, for, after all, the subject is learning to "hear" through another sense, namely vision.

A set of instructions for spectrogram reading were formulated during the early and middle 1940's by a group of phoneticians and engineers. The rules were published in the book *Visible Speech*.[7] The phoneticians' point of view prevailed in this coalition, and the rules were based primarily upon what was known at that time about the physics of speech production. The cues that they specify might be thought of as sufficient to define a sound, but whose presence is not necessary to the perception of that sound. Indeed there are many psychologically based factors which affect speech perception in a major way. For instance, a person often hears what he expects to hear. Furthermore, the *Visible Speech* rules concern the individual sounds of speech or, as linguists and phoneticians would call them, the phonemes. The concept of a few building blocks from which speech is put together is an extremely attractive one. It reduces the vocabulary size drastically—all the words of English can be synthesized from just 40 basic phonemes. It is quite another thing, however, to think of perceiving speech in terms of these elements. One might make analogy here between speech and literal text. The letters correspond to the phonemes in this analogy. I have often had the experience of not being able to spell a word which I have read many times. Too, when I misspell a word, I often recognize it as being wrong just because it doesn't "look right." What I am saying is that in reading, words are often perceived as a unit, not as a collection of individual letters. There are some interesting psychological data available on this point in a study by Postman and Adis-Castro.[8] They found that when familiar words were flashed briefly on a screen the minimum recognition time for an observer was nearly independent of word length. With less familiar words the perceived units are evidently shorter than words, and recognition time does not depend upon length. Although I know of no experimental data, the perceptual units of speech are under most circumstances longer than a single phoneme. Indeed interactions between adjacent phonemes was a recognized fact, stated in the *Visible Speech* book and confirmed many times since. In one recent demonstration, a phoneme-length section of connected speech was selected and embedded in several different speech samples. It was found that this same section in various contexts could give rise to several different perceptions. What people heard depended not only on the acoustical properties of the section, but on its surroundings as well. Prof. G. E. Peterson of the University of Michigan has recently begun speech synthesis experiments using elements consisting of two connected speech sounds.[9] He calls these elements *dyads*. The transition between the two sounds is in the middle of each dyad. The beginning of the first sound and the end of the second sound lie at the phonetically most stable position of each sound. It may be that the normal perceptual units of speech are more nearly akin to dyads than to the individual speech sounds themselves. Peterson and one of his students found that about 8,000 dyads are needed to synthesize the midwestern dialect of American Speech.[10] This number is certainly not too large for a human listener to cope with. Indeed it is entirely conceivable that the normal listener's perceptual vocabulary is many times larger than this. Thus, it may be that the making of a phonetic transcription, which involves resolution of a continuous utterance into a small number of discrete characters, must be preceded by an identification of longer speech segments.

R. H. Galt[11] in 1951 decided to draw up a set of rules for recognition of ten spoken numbers from spectrograms. Such rules might more nearly correspond to the perception of spectrographic patterns and might prove more easily applied to a variety of speakers. Galt obtained the services of two "experts" who had long experience in relating the visual spectrographic features to the perceptual features of the corresponding utterances. After a few trial sessions, Galt had his experts try to write down the rules they had used. After further refining, these rules



CHIEF SLOPE OF BAR 2 DOWN.

CHIEF SLOPE OF BAR 2 UP (OR HORIZONTAL). VERTICAL LOCATION OF LEFT END OF BAR 2 MIDDLE OR HIGH AND SEPARATED FROM BAR 1.

CHIEF SLOPE OF BAR 2 UP (OR HORIZONTAL). VERTICAL LOCATION OF LEFT END OF BAR 2 LOW AND USUALLY TOUCHING BAR 1.

A A PATTERN WHICH HAS IRREGULARLY SPACED VERTICAL MARKINGS IN THE LEFT-HAND REGION. TO THE RIGHT OF THIS IRREGULAR REGION WILL BE FOUND A REGION HAVING REGULARLY SPACED VERTICAL "STRIATIONS" IN WHICH REGION OCCURS THE CHIEF SLOPE OF BAR 2.

B A REGULARLY STRIATED REGION OTHER THAN THE REGION CONTAINING THE CHIEF SLOPE OF BAR 2.

C A PATTERN IN WHICH, GOING FROM LEFT TO RIGHT, BAR 2 AND BAR 3 TEND TO COME TOGETHER. OFTEN THESE BARS MEET AT OR NEAR THE RIGHT-HAND END OF THE REGULARLY STRIATED REGION CONTAINING THE CHIEF SLOPE OF BAR 2.

D A GAP OR BLANK SPACE EXTENDING FROM TOP TO BOTTOM ACROSS THE PATTERN.

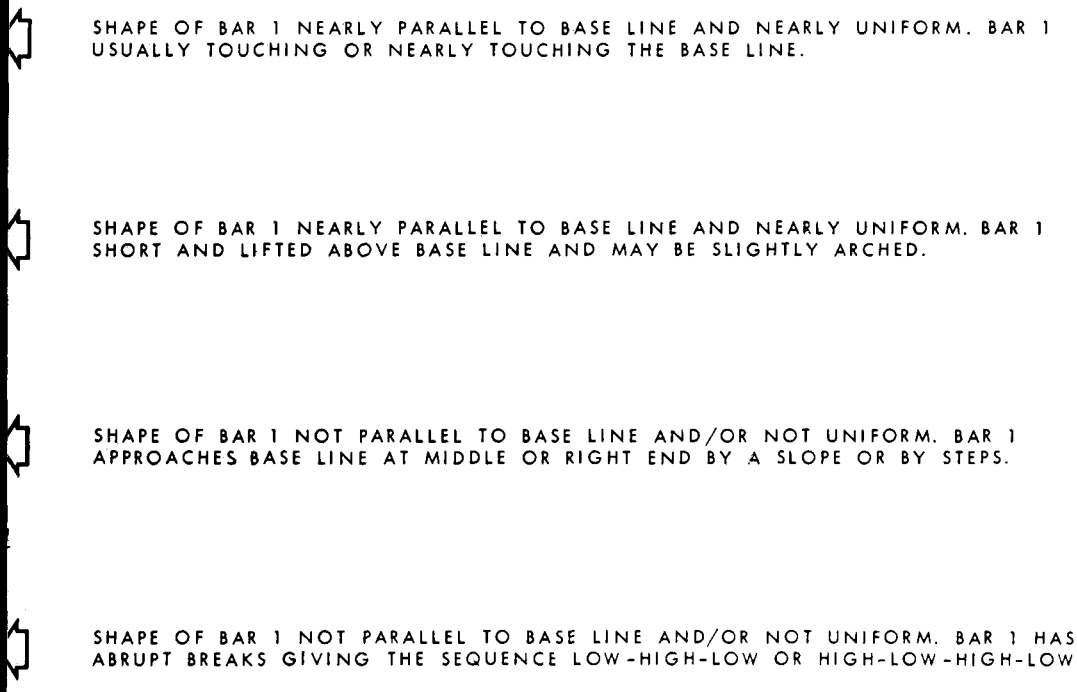


Figure 10 **Scheme for classifying patterns of digits.**

were then tested on 330 utterances of the spoken digits from thirteen men, seven women, and ten children. Galt found that naive subjects could identify the corresponding spectrograms with well over 90% success. Some of the rules are illustrated in Fig. 10, adapted from one of Galt's original figures. Six of them concern the frequency locations, slopes, and continuity of the first three formants. For instance, notice the up-ended descriptions; they sort the digits into categories according to the principal slope of the second formant. Digits "2," "7" and "Oh" show a down-slope, the others an up-slope. Another rule concerns the presence of an initial unvoiced section in the digits marked with *A*'s, namely "2," "7," "3," "6," "4," and "5." Still another rule concerns the presence of voiced sections with formants fixed in frequency—the digit "9" has two such sections, one at each end. Finally, one rule notes the presence of a silent gap, marked *D*, bisecting the digits "6" and "8."

For convenience, I have arranged Galt's rules in matrix form, as shown in Fig. 11. Along the top are listed the digits, along the side the rules, stating the features to be noted. A plus or minus in the matrix indicates the presence or absence of a particular feature in that digit. A zero indicates that a particular rule is not applicable to the digit. In identifying a spectrogram, the rules are applied, and a corresponding list of $+$'s and $-$'s prepared for that spectrogram. This list is then compared to each column in the matrix and a "score" derived for each column. A coincidence of either $+$'s or $-$'s counts $+1$ in scoring the column, a mismatch counts -1, and a zero nothing. The maximum score identifies the utterance.

Note that these rules concern rather gross properties of the patterns, and require no very precise measurements. Rather it is the relations between the features that are important. Further, the identity of each word does not depend exclusively on any one of its properties. To confuse one word with another requires the simultaneous confusion of several features. This "diversity" property is fundamental to the success of the rules since a particular utterance may exhibit some but not others of the features. The number of confusions separating each of the digits can be calculated simply by noting the number of differences between the columns. Counting each $+-$ pair as 1 and each $+0$ or -0 as 1/2, I have prepared such a table as shown in Fig. 12. Here each entry gives the number of distinctions separating the digits labeling the corresponding row and column. If Galt's rules accurately reflected the perceptual components of the spoken digits, then these numbers should specify how different each digit would sound. Of these, according to the table, "one" and "nine" are most alike, being only 2 distinctions apart.

RULE	DIGIT									
	1	2	3	4	5	6	7	8	9	OH
FORMANT ONE										
NEARLY PARALLEL TO BASELINE AND CONTINUOUS	−	+	+	+	−	+	−	−	−	−
TOUCHING BASELINE IN MOST PLACES	0	+	+	+	0	−	0	0	0	0
SLOPES DOWNWARD AND HAS NO ABRUPT BREAKS	−	0	0	0	+	0	−	+	−	+
FORMANT TWO										
PRINCIPAL SLOPE DOWNWARD FROM LEFT TO RIGHT	−	+	−	−	−	−	+	−	−	+
LEFT END LOW, TOUCHING FORMANT ONE	+	0	−	+	+	−	0	−	−	0
FORMANTS TWO AND THREE										
FORMANTS TEND TO COME TOGETHER FROM LEFT TO RIGHT	−	−	−	+	+	+	−	+	+	−
FORMANTS PROCEED UPWARD TOGETHER FROM LEFT	−	−	+	−	−	−	−	−	−	−
OTHER										
WORD BEGINS WITH UNVOICED SECTION	−	+	+	+	+	+	+	−	−	−
PRESENCE OF VOICED REGIONS WITH FIXED FORMANTS	+	−	−	−	+	−	+	−	+	−
PRESENCE OF GAP	−	−	−	−	−	+	−	+	−	−

Figure 11 **Tabulation of recognition rules.**

There are two significant facets of speech recognition which can be discussed in the context of Galt's rules. The first involves what is commonly called the vocabulary, the second, population. Vocabulary size refers to the number of categories in the classification matrix. In general, the larger the vocabulary, the more rules needed, but there is a more important aspect. Galt's rules as they were used in his experiment concerned only ten spoken digits. Any utterance presented had to fall in one of the ten categories. Furthermore, the utterances presented for classification were guaranteed by the experimenter to fall in one of the ten categories. There were no "elevens" or "ninetys" presented to the subjects for classification. This is what I would like to call a *selected* vocabulary. To make Galt's vocabulary what might be called *complete*, would require the addition of only one other category, namely one called *not one of the digits*. All of the other words of English would fall into this latter category. This addition would undoubtedly multiply the number, subtlety, and complexity of the necessary rules many-fold, if the same level of performance were maintained. This is another way of saying that recognition success on a limited, *but complete*, vocabulary stands a better chance of uncovering pattern properties which can be generalized efficiently to higher orders of discriminability than similar success on a limited, *selected* vocabulary. Thus a

Figure 12 **Tabulation of digit separations.**

	1	2	3	4	5	6	7	8	9	OH
1		5½	6	5	3	7	2½	5	2	3½
2			2½	2½	5½	4½	3	6½	6½	3
3				3	6	4	5½	6	6	5½
4					3	3	5½	5	5	5½
5						5	3½	4	3	4½
6							6½	3	3	6½
7								6½	3½	3
8									3½	3½
9										4½
OH										

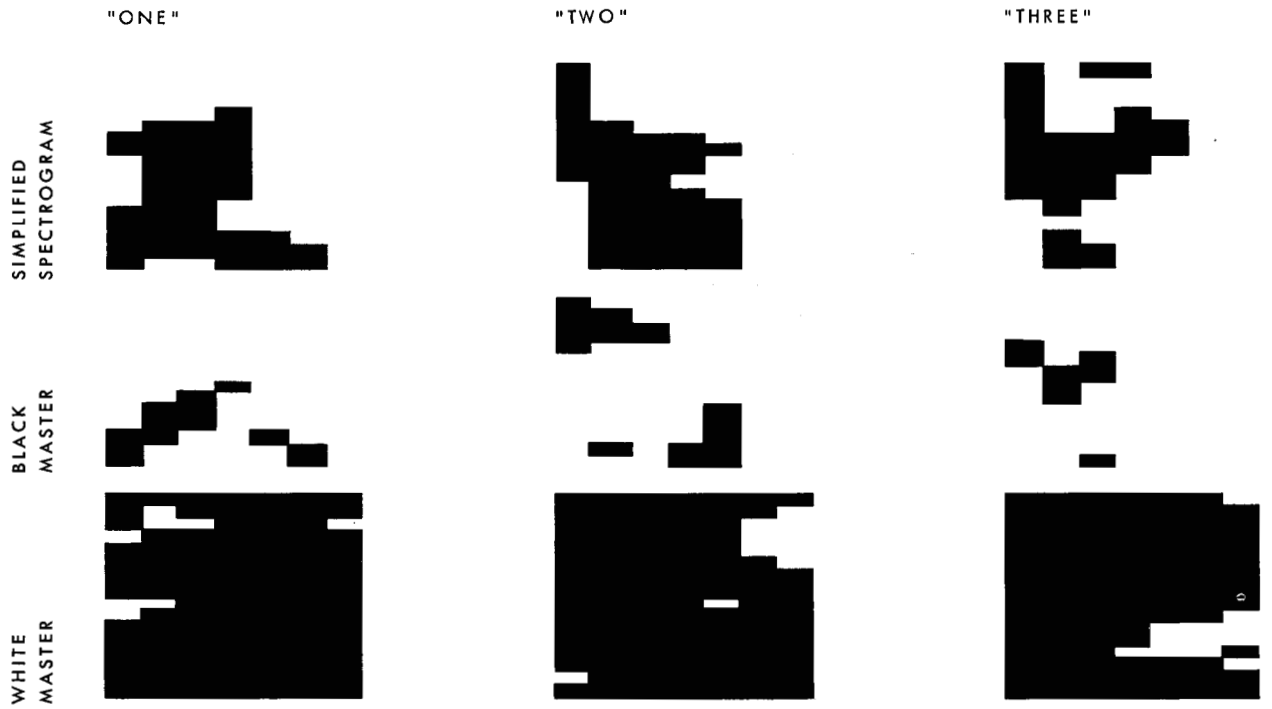


Figure 13 **Simplified spectrograms and master patterns.**

limitation on the size of a complete vocabulary, while it represents a simplification, might be thought of as a rather realistic constraint. After all, the vocabulary size in everyday communication is restricted more than one might think by environment, subject matter, and grammatical and linguistic structure.

Population size refers to the number of subjects from whom speech samples will be accepted for recognition. Of course, the complexity of a recognition task increases, within both selected and complete vocabularies, with population size. In addition, there are probably some members of society whose pronunciation is so unorthodox that their utterances out of context would not be recognized correctly by most listeners. These people are probably not desirable members to include in the recognition population. The performance criterion for a recognition scheme should be based on perceptual experiments.

Other facets of Galt's rules are significant. The rules, as they are written, presuppose a single word whose beginning and end are closely defined. What additional problems does a connected utterance pose? Certainly, the perceptual speech units to be recognized will not in general be words, but will be some other perhaps as yet unnamed entity. One of the entities will undoubtedly be silence. Others may be similar to Peterson's dyads. In any case there will be many more than the 40 English phonemes. Furthermore, they will not be of uniform length. Thus, a set of recognition rules will have to include a criterion for selecting the proper units for identification.

Note also that Galt's rules are written in terms of spectrographic features which themselves must be recognized if the rules are to aid machine recognition. Finding a formant or a silent gap in an acoustic complex is a recognition problem just as is the identification of a spoken number. In this respect the rules remind one of the ancient explanation of the cosmos in which the earth is supported by four elephants which stand on the back of a tortoise—a legitimate question is, what does the tortoise stand on? The rules are not really so unsupportable, however. After all, they reduce the recognition problem by one level of complexity.

I know of no totally successful work on spectrographic feature recognition. However, L. G. Kersta[12] in 1947 did an interesting piece of work in which he got at this problem in a sense. In retrospect, it can be said that Kersta's approach took advantage of the same sort of relations incorporated in Galt's rules. His results demonstrate how great a simplification of spectrographic information can suffice for identification. He again assumed a selected vocabulary consisting of ten spoken numbers, and took spectrograms of them from each of nine men and five women. He divided each spectrogram into a mosaic of square elements each of which measured 200 cps by 67 milliseconds. If the integrated density in a particular element were 1/2 or greater than the integrated density in the darkest element, then it was represented as being entirely black. If the integrated density were less than 1/2 then it was represented as being entirely white. Thus, Kersta achieved a highly simplified spectrogram, as is indicated for three digits in the top row of Fig. 13. For each digit he then compiled two master patterns. One consisted of all black elements common to the utterances of the fourteen speakers. These are shown in the second row of Fig. 13 for the digits "one," "two," and "three." The second master pattern consisted of all common white elements as shown by the blank elements in the bottom row of Fig. 13. When the original patterns were compared

		1	2	3	4	5	6	7	8	9	OH
BLACK MASTER	X	0.8	0	6	4	6	33	10	19	0	0
	Y	4	15	14	3	6	4	8	6	15	4
WHITE MASTER	X	6	6	13	5	2	1	4	1	19	5
	Y	1	2	3	8	9	11	12	8	0.7	6
"AND" ERROR PROB. %		0.05	0.00	0.70	0.22	0.13	0.38	0.38	0.20	0.00	0.00

0.21% AVERAGE

Figure 14 **Tabulation of recognition performance.**

with each master, it was found that quite often a single original pattern contained elements in common with the masters from several numbers. This situation resulted in a single spectrogram being classified into several categories. Just how often this occurred for both the black and white masters can be seen in Fig. 14. The percent probability that a particular digit will match a master other than the correct one is marked x, and the probability that other digits will match a particular master is marked y. For instance, for the black masters, spoken "twos" were never identified as any other number while other numbers were called "two" with a probability of 15%. Now if the black and white masters are applied in tandem with a logical "and" requirement; that is, both black and white elements must match, the error probability falls startlingly. The maximum error probability for any digit is less than one percent, while the average over all digits is about 0.2%. In other words, the number utterances were identified with over 99.5% success. Clearly, Kersta has extracted some common elements from the utterances of 14 people.

To digress for a moment, I particularly wanted to mention Kersta's experiment because it was done by digital simulation. The simplified spectrographic data from the 140 utterances were punched onto IBM cards, as were the master patterns. An IBM card sorter then carried out the matching procedure and compiled the results. It might well be that this was the first use of digital equipment in speech-processing research. It presaged a technique which promises to catalyze research in efficient voice and visual communication. Needless to say, IBM research and IBM equipment are playing a major role in this area.

Kersta's master patterns might be thought of as templates, and the identification procedure as a matching process. In thinking of broader application of the scheme, it is natural to ask what would happen to the master patterns as the population were expanded from 14 people. Certainly as all personal variations, such as duration and dialect, are absorbed all the patterns will contain null elements exclusively; there will be no elements in common among the simplified spectrograms. The scheme might be stretched in this direction by using a probabilistic principle among the spectrograms, rather than requiring absolute coincidence among them all. This modification might be applied both for determination of the master patterns and for the matching procedure. Nonetheless, the template approach seems limited by the gross variability among talkers. However, it does illustrate the degree of success that can at present be attained with the limited-selected vocabulary, and a limited population. More than this, it again affirms the principle of using many gross features or decisions rather than a few precise ones in speech recognition.

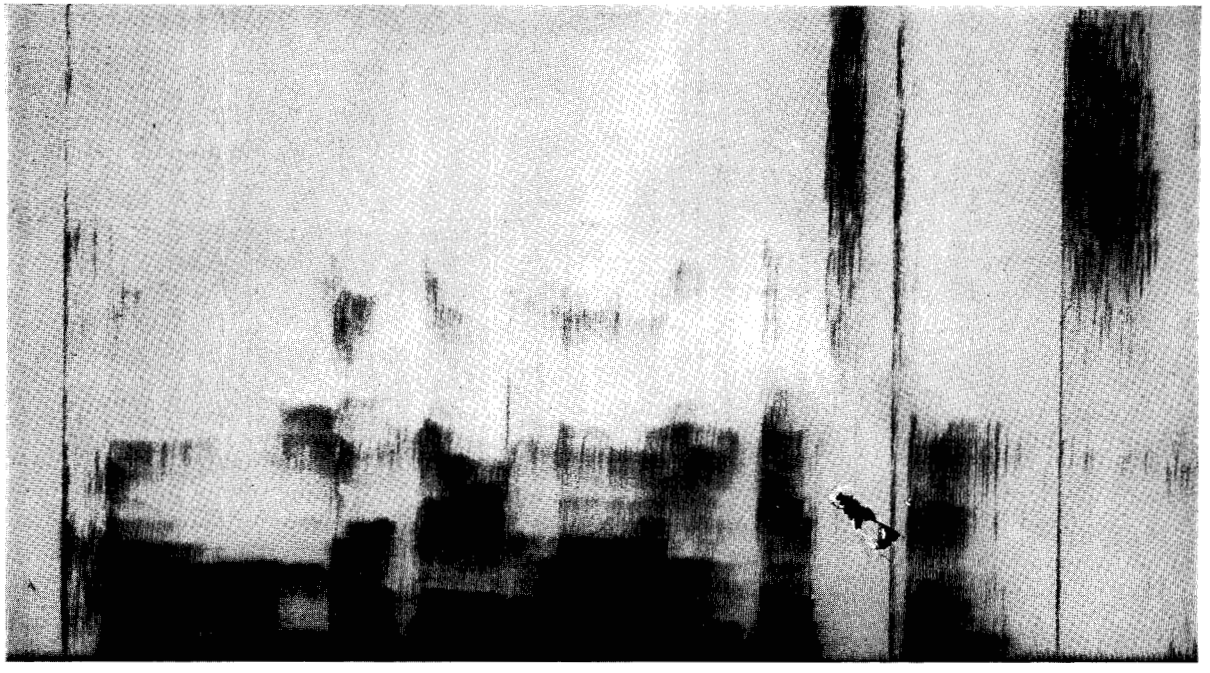
Thus, while there have been several more or less successful but limited demonstrations of speech recognition, many of the fundamental questions remain to be resolved. As I have pointed out earlier, one of the most central of these is the recognition of formants, silence, and the other important features. Importance here implies perceptual importance. There may well be features visible on spectrograms which have no auditory consequence. These should be ignored in the recognition process. In examining spectrograms of the same words spoken by several talkers, one cannot fail to be impressed by the gross differences in the spectral patterns. You will agree, I think, after comparing spectrograms of the same words spoken by a man in Fig. 7 and by a woman in Fig. 15. Some of these differences at least are not real in the sense that they reflect true differences in the acoustic character of the utterances. Consider, for instance, the vowel produced by the usual model, namely a linear, passive network excited by repetitive, periodic impulses as shown in the upper line of Fig. 16. The phonetic value of the vowel is determined by the network response, the voice pitch by the impulse repetition rate. Two voice pitches are illustrated, one high and one low. The vowel sound will then consist of repeated transients at the same repetition rate. As indicated in Fig. 16, its spectrum will be composed of harmonically related tones whose amplitudes depend upon the response of the network at the frequencies of the tones. One can look upon the harmonic magnitudes as "samples" of the network amplitude response, close together for a low pitch, further apart for a high pitch. If the vowel is then passed repeatedly through a slowly scanning filter, a reduced resolution "spectrogram" results. The degree of resolution depends upon the relation between the filter width and the harmonic spacing. If the voice pitch is low, and the filter is broad enough to encompass two or three harmonics, a reasonably good

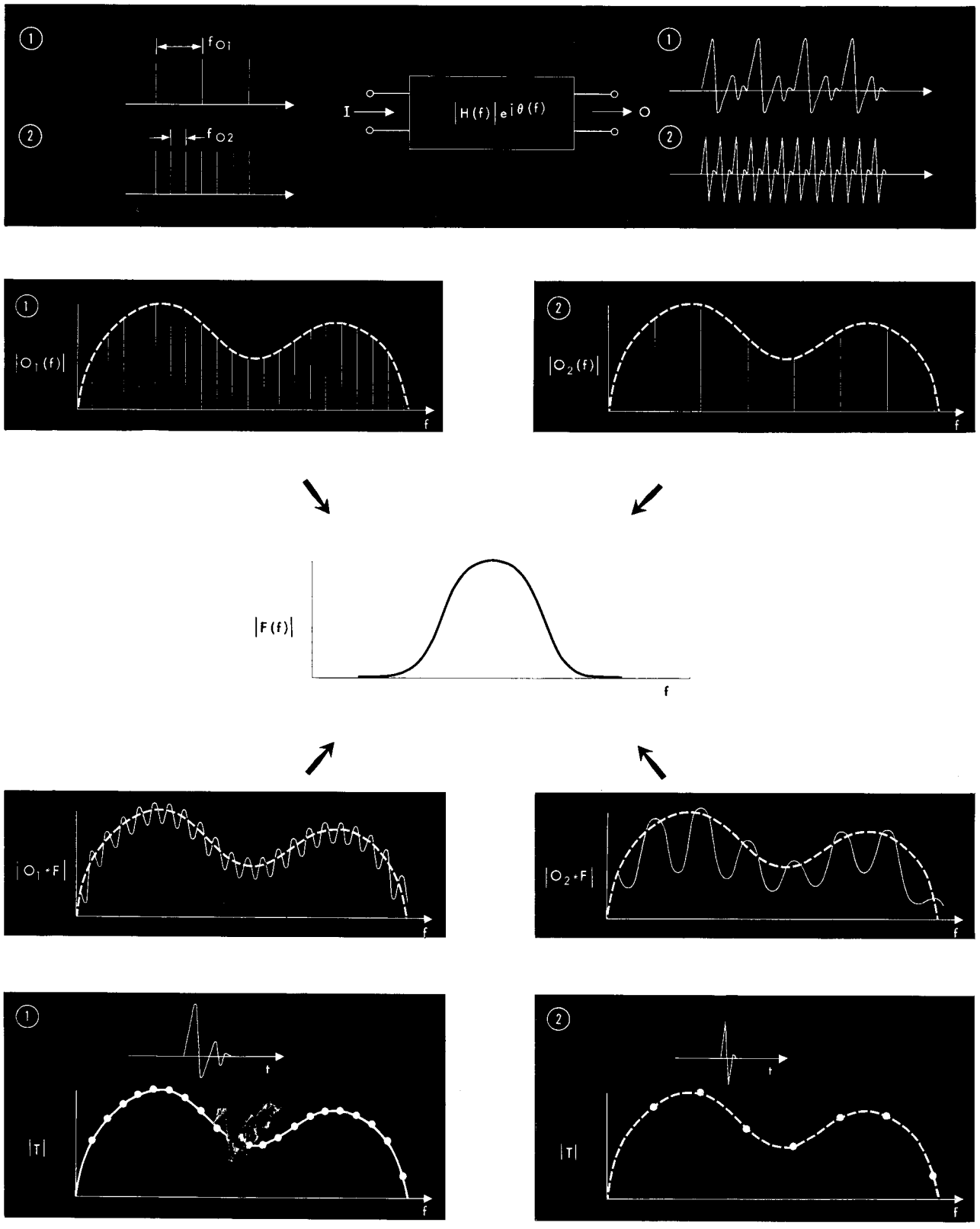
approximation to the network amplitude response will result, as shown on the left in Fig. 16. This is not true when the voice pitch is high, and the filter width is such that it includes only one harmonic at a time, as shown on the right. The spectrograms in these two instances will appear to the eye, and to any instrument observing the spectral analysis, as being quite different. In the former case, the formants would tend to be clearly evident; in the latter, harmonic maxima could be easily confused with formant maxima. Thus a confusing structure can be superimposed on a spectral analysis by the discrete and variable nature of voiced sounds. Of course, if the pitch frequency is known, then the effect can be removed subsequent to the analysis. A better method, shown at the bottom, is to segment the speech into pitch period length segments before analysis. The spectrum of each period is the same as the response of the network at a finite number of points; these points lying at the positions of the vowel's harmonic tones. An estimate of the network response can then be had from these points, which in the case of the period-by-period analysis are guaranteed to fall on the response curve. Thus, the parameters of speech analysis should be a function of the voice pitch itself. Seen from this viewpoint, pitch detection is fundamental to the whole problem of speech analysis, and quite likely to speech recognition as well, for a variety of talkers and pitches.

While a pitch-synchronous analysis will be helpful, it will not, of course, be an "open sesame" to the promised land of recognition. Rather, basic studies in at least three fields are called for. First, we should pursue further study of human anatomy and physiological processes. I don't want to imply that a speech recognizer must follow human methods exclusively. However, the close relation between speaking and hearing would argue for this view. The argument becomes more valid as recognition aims encompass a larger population of talkers and a larger vocabulary. Second, physiological research must be supplemented by measurements of human ability and behavior in the process of recognition. Such psychological work should not be merely a tabulation of the properties of human behavior, but should be guided by underlying models, perhaps incorporating a considerable amount of mathematical and logical formalism. One problem that deserves attention is that of establishing the relative perceptual importance of various acoustic features. The more important should be weighted more heavily in a recognition scheme. Thus, for instance, in scoring a spoken digit against Galt's rules, some coincidences and anti-coincidences should be counted more heavily than others. Which ones, we do not know at present, although we might guess. One way to investigate this matter is through psychoacoustic experiments in auditory perception.

Another possibility in finding perceptual correlates is to have a computer act as a naive subject in Galt's experiment. By observing its own errors, it might over a period of time compile a successful set of rules. Identifi-

Figure 15 **Sound spectrogram of female speech.**





308 *Figure 16* **Two methods of spectral analysis.**

cation of spectrographic features might be accomplished similarly.

Finally, the problem of recognition must be attacked by considering logically and mathematically how a human being or a machine could recognize speech patterns, and by testing experimentally the validity of proposed schemes. Here the use of digital computers to simulate the necessary experimental equipment will yield very considerable economies of time and effort, while retaining the necessary flexibility.

Work in each of these fields is progressing, and communication across their boundaries is providing a good deal of stimulation. As this research proceeds, the limitation on recognizers will come to rest upon how much human time, effort, and money we are willing to expend in simulating human functions.

References

1. A. B. Clark and H. S. Osborne, "Long Distance Telephone Circuits in Cable," *Bell Sys. Tech. J.*, **11,** 520-545, 1932.
2. R. L. Ives, "Music Pulse Analyzer Rejects Voice Signals," *Electronics,* **30,** 183-185, April 1957.
3. H. W. Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, **11,** 169-177, 1939.
   H. W. Dudley, "The Carrier Nature of Speech," *Bell Sys. Tech. J.*, **19,** 495-515, 1940.
   H. W. Dudley, "Fundamentals of Speech Synthesis," *J. Audio Eng'r. Soc.*, **3,** 170-185, 1955.
4. R. K. Potter et al, "Technical Aspects of Visible Speech," *J. Acoust. Soc. Am.*, **17,** 1-89, 1946.
5. G. A. Miller and P. E. Nicely, "Analysis of Perceptual Confusions Among Some English Consonants," *J. Acoust. Soc. Am.*, **27,** 338-345, 1955.
6. J. L. Flanagan, "Bandwidth and Channel Capacity Necessary to Transmit the Formant Information of Speech," *J. Acoust. Soc. Am.*, **28,** 592-596, 1956.
7. R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech,* D. van Nostrand Co., Inc., 1947.
8. L. Postman and G. Adis-Castro, "Psychophysical Methods in the Study of Word Recognition," *Science,* **125,** 193-194, 1957.
9. G. E. Peterson and W. S-Y Wang, "Segmentation Techniques in Speech Synthesis," *J. Acoust. Soc. Am.*, **30,** 739-742, 1958.
10. G. E. Peterson and W. S-Y Wang, "Segment Inventory for Speech Synthesis," *J. Acoust. Soc. Am.*, **30,** 743-746, 1958.
11. R. H. Galt, Bell Telephone Laboratories, unpublished notes dated October 2, 1951.
12. L. G. Kersta, Bell Telephone Laboratories, unpublished notes dated July 1, 1947.

Received May 27, 1958