Speech Recognition by Computer

Designing a machine that listens is much more difficult than making one that speaks. Significant improvements in automatic recognition may come only with a better understanding of human speech patterns

by Stephen E. Levinson and Mark Y. Liberman

Modern computers have prodigious powers, but they would be still more useful if more natural ways of communicating with them were possible. The evolution of spoken language has made it well adapted to the needs of human communication. It is fast and nearly effortless. It requires neither visual nor physical contact and it places few restrictions on the mobility of the body or on the use of the hands. A machine capable of recognizing human speech could combine these advantages with the quite different powers of the computer. Such a machine could provide universal access to large data bases through the telephone network. It could provide for the control of complex machines by vocal command and make possible sophisticated prosthetic devices for the handicapped.

After more than 40 years of research, however, the automatic recognition of natural or conversational speech remains a utopian goal. Current speech-recognition devices have small vocabularies and little ability to deal with fluent sequences of words; usually they must be trained to recognize only one speaker's voice. Even so, the advantages of automatic speech recognition are so great that devices capable of recognizing isolated words or short phrases from a vocabulary of between 10 and 30 words are commercially available and are economically practical in some applications. In research laboratories there are speech recognizers with vocabularies of up to 1,000 words, systems that recognize limited-vocabulary sentences with brief pauses between the words and systems that recognize connected speech with fair accuracy if the vocabulary is small, the syntax is limited and the speaker is careful.

The interaction of technology and economics will undoubtedly lead to speech-recognition systems of greater capability. We cannot accurately predict the pace of such development. We are certain, however, that mere elaboration and extrapolation of current technology will not lead to the development of machines that match the human capacity for recognizing speech. Major progress depends on new discoveries.

Why is the problem of recognition so hard? The core of the difficulty is the complex and variable way linguistic messages are encoded in speech. Spoken language enables people to express their thoughts in sound and to recover messages from the sounds produced by others. This curious two-way mapping between concepts in the mind and vibrations in the air presupposes the participants have some common conceptual framework, so that the message received is at least approximately equivalent to the one that was sent. It is not enough, however, to share knowledge of the things one might want to say. Monolingual speakers of English and of Finnish may have many potential messages in common without being able to understand each other's utterances at all. In order to speak and to understand, people must share a system for encoding messages in sound and for decoding speech sounds to yield meanings. In other words, they have to know the same language.

Speech communication with a computer can be understood in an analogous way. The computer "knows" (in some extended sense of the word) about a domain its users also know about. It is useful for information in this domain to be exchanged, and speech happens to be the chosen medium of communication.

Consider a conversation between a computer and its users concerning the inventory of a warehouse. The computer "knows" how many of each item are on hand and where each article is stored. Its data base also lists costs and suppliers. People probably think about the warehouse and its contents in many ways, but the structure of the computer data base is sufficiently similar to one mode of human thought for certain kinds of communication to be possible. The users have questions that the computer can answer, at least in principle, such as, "Do we have any blue pencils in stock?" The users also have things to say that the computer can profitably "understand," such as, "There is no more room in bay 13." If such communication is to be accomplished through the medium of speech, the computer and its users must agree on how to encode such messages in sound and how to reverse the process. They have to "know" the same language.

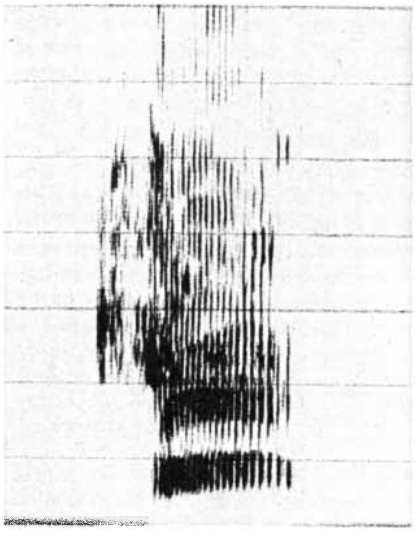
We are mainly interested in languages such as English, which are called natural because they are defined implicitly by the everyday usage of ordinary people. Computers now function with formal languages such as FORTRAN, which are defined by an explicit set of rules consciously established by specialists. At least for now, computers do only what they are specifically programmed to do. They do not live in the world of people and learn from everyday experience. Hence for a computer to "know" a natural language, it must be provided with an explicit and precise characterization of the language, or at least with a characterization of what the programmer takes the language to be. In all existing and currently conceived speech-recognition systems the formal description of a natural language covers only a fragment of the language, and the formalism reconstructs the fragment in ways that are probably quite different from the implicit knowledge of a native speaker. Even imperfect linguistic abilities on the part of a computer, however, are enough to make possible useful communication with people.

It will help in understanding various approaches to natural-language recognition if we begin by considering some aspects of language and speech in their own terms. We shall then discuss methods of recognizing isolated words and review some procedures for analyzing connected speech. Finally we shall describe a speech-recognition system developed at Bell Laboratories that attempts to combine the major elements of human speech communication into a single operating unit.

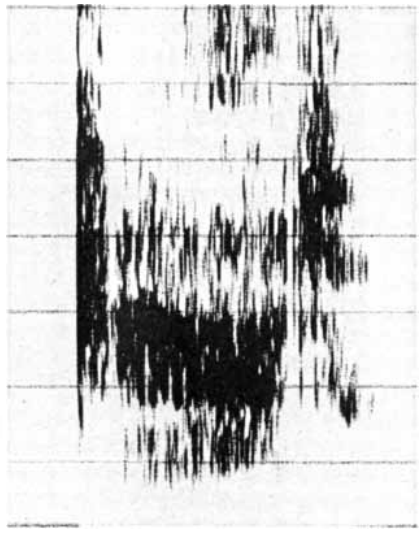
At the center of human language is the word. Sequences of words are generally arranged into phrases according to principles of combination known as syntax; moreover, such sequences are usually intended to mean something. The fact that words are ordinarily part of coherent discourse can help in the recognition of the words themselves by providing a context in which some words are likelier than others. Arranging for a computer to act as if it could "understand" word sequences is formidably difficult. The problem involves not only relations among words but also knowledge and reasoning about the nature of the world.

Although a capacity for understanding language may be the ultimate goal, the enterprise of speech recognition is really founded on the identification of words. The aspect of words that concerns us here is their sound. In this sense a word is an equivalence class of noises: the set of all sounds, however distinct in other ways, that represent (in the context of their utterance) the same lexical unit. The problem in word recognition is to find a mathematically defined space in which such a set of sounds can be effectively delimited. Because the amount of variation within the set of sounds corresponding to a given word is quite large, because the acoustic distinction between words can be quite small and because an adult speaker may know 100,000 words or more, the problem is a difficult one.

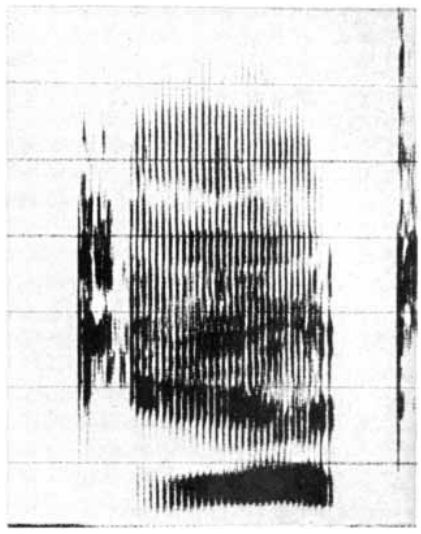
In order to understand the sources of variation in the sound of a word and the nature of the distinction between one word and another, it is necessary to grasp two things. First one must understand the basic medium of spoken communication: the ways in which acoustic disturbances of the air can be produced by the human vocal apparatus and perceived by the human auditory system. Second, one has to recognize that speech sounds are elements of a phonological system peculiar to a given language. The



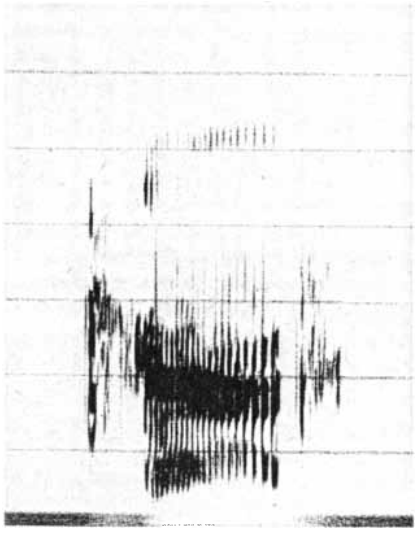
"CAT" SPEAKER 1 MICROPHONE



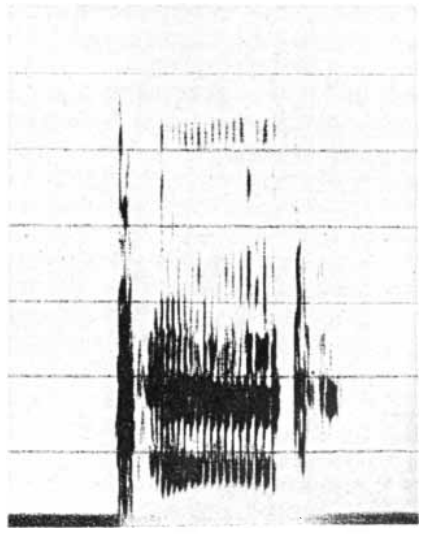
"CAT" SPEAKER 1 WHISPERED



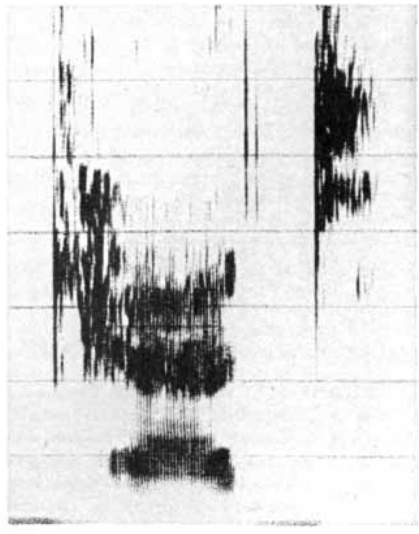
"CAT" SPEAKER 2 MICROPHONE



"CAT" SPEAKER 1 TELEPHONE



"PAT" SPEAKER 1 TELEPHONE



"CAT" SPEAKER 3 MICROPHONE

VARIABILITY OF HUMAN SPEECH, illustrated here by means of sound spectrograms, is one of the principal difficulties encountered in building an automated system for speech recognition. Spectrograms of distinct but acoustically similar words may be more alike than spectrograms of the same word pronounced under various conditions by different speakers. Automatic speech recognition must be able to attend only to relevant spectral differences (when they exist) and must disregard apparent differences that are linguistically irrelevant. The sound spectrogram represents a series of amplitude spectra over time. Time varies along the horizontal axis and frequency varies along the vertical axis. The darker the mark on the graph, the greater the amplitude of the waveform at that frequency and time.

```
NFRAM=  396
NO. OF WORDS=  5
CANDIDATES FOR WORD NO.  1  20 FRAMES
WHAT          1   .180
CANDIDATES FOR WORD NO.  2  29 FRAMES
IS            1   .270
CANDIDATES FOR WORD NO.  3  24 FRAMES
NINE          1   .343
ONE           1   .278
SIX           1   .370
SEVEN         1   .242
THE(ALT)      1   .314
CANDIDATES FOR WORD NO.  4  24 FRAMES
SOME          1   .296
FLIGHT        1   .341
ONE           1   .414
THREE         1   .390
CANDIDATES FOR WORD NO.  5  28 FRAMES
NINE          1   .378
OH            1   .390
TIME          1   .291
ONE           1   .411
SEVEN         1   .396
TIMES         1   .315
WHAT IS SEVEN SOME TIME
METRIC=  .1279442E 01
WHAT IS THE FLIGHT TIME
METRIC=  .1396801E 01
THE FLIGHT TIME IS FIVE HOURS AND TWENTY-FIVE MINUTES
```

Machine:	This is the Bell Laboratories flight information system. May I help you?
User:	I want one first-class seat on flight number three one to Denver on Sunday.
Machine:	I am sorry the flight is not available on Sunday.
User:	I would like to leave on Saturday.
Machine:	Flight number three one leaves New York at nine forty-five A.M., arrives in Denver at one ten P.M.
User:	What is the flight time?
Machine:	The flight time is five hours and twenty-five minutes.
User:	How many stops are there on the flight?
Machine:	This flight makes one stop.
User:	I want a nonstop flight.
Machine:	I am sorry there are no nine A.M. nonstop flights.
User:	I want to go at four P.M.
Machine:	Flight number one six leaves New York at four fifty P.M., arrives in Denver at seven P.M.
User:	What stops are there on the flight?
Machine:	This flight makes no stops.
User:	Is a meal served on the flight?
Machine:	Dinner is served.
User:	How much is the fare?
Machine:	The round-trip first-class fare is three hundred sixty dollars.
User:	What plane is on the flight?
Machine:	The aircraft is a Douglas DC ten.
User:	I prefer the Boeing seven four seven.
Machine:	The aircraft is a Douglas DC ten.
User:	Please repeat the departure time.
Machine:	The departure time is four fifty P.M.
User:	I will take flight number six five.
Machine:	I am sorry there is no flight six five to Denver.
User:	Please repeat the flight number.
Machine:	The flight number is one six.
User:	I want flight one six.
Machine:	Your reservation on flight number one six to Denver on Saturday is confirmed. Thank you.

TELEPHONE CONVERSATION between man and machine proceeds according to steps that can be displayed on the screen of a computer terminal. The computer counts the number of words in each sentence to be recognized and also divides the entire utterance into "frames" by taking a spectral sample every 15 milliseconds. "NFRAM" is the number of frames in the sentence. It exceeds the number of frames occupied by individual words because the speaker must pause briefly between the words. The candidate words listed for each position in the sentence have been found by comparison with word templates stored in the computer. Only those candidates appear that are grammatically possible in a given position and similar in spectral structure to the spoken word. Not all the candidates to be considered are listed. The numbers following each candidate word are measures of the distance between the word's template and the utterance. The shorter the distance is, the more similar the template is to the utterance. "METRIC" is the unrounded sum of the distance measures for a particular string of words. If the smallest possible METRIC (which necessarily consists of the most likely word in each position) is not allowed by the internal grammar, the grammatically correct string with the smallest METRIC is substituted. A synthetic-voice response to the question by the user is given over the telephone. The complete conversation is transcribed in the computer printout.

phonological system limits the ways in which the various words of the language can differ and controls in part the ways in which the pronunciation of any specific word in the language can vary.

During speech a flow of air from the lungs passes through the larynx, or voice box, into the throat and out through the mouth. If the velum (the flap of soft tissue at the rear of the palate) is lowered, the airflow also proceeds out through the nose; if the velum is raised, the nasal passages are blocked. The airflow can also be obstructed by closing the lips, by pressing the tongue against the palate or by closing the glottis, which consists of two parallel folds of soft tissue (the vocal cords) within the larynx.

The flow of air through the vocal tract can give rise to sound in three main ways. First, the vocal cords can be made to vibrate in somewhat the same manner as the double reed of an oboe or a bassoon. When the vocal cords are brought together, they stop the passage of air from the lungs, and so pressure builds up below them. The pressure forces the vocal cords apart, but the velocity of the rushing air then reduces the pressure in the space between them. The reduction in pressure and the elasticity of the tissues bring the vocal cords together again, in position for another buildup of pressure. The rate at which this cycle is repeated is the fundamental frequency of the voice, which is heard as pitch.

The second way of generating sound in the vocal tract is to form a constriction in the airway narrow enough to cause turbulence. For example, forcing air past a close contact between the upper teeth and the lower lip causes a turbulent flow that is perceived as the sound "f." Unlike the periodic sounds created by vibration of the vocal cords, the sounds generated by turbulent flow are aperiodic, or noiselike. It is possible for the vocal tract to create both periodic and aperiodic sounds at the same time. Combining vocal-cord vibration with the noise source of an "f" gives rise to the sound perceived as a "v."

A third kind of sound generation takes place when pressure built up behind a closure is abruptly released. Such bursts of acoustic energy occur in the pronunciation of consonants such as "p," "t" and "k."

These three sources of speech sound are shaped acoustically by the changing physical shape of the vocal tract. If the vibrations of the vocal cords were somehow vented directly to the outside air without first passing through the throat, mouth and nose, they would sound rather like a door buzzer and not like speech at all. On passing through the throat, mouth and nose cavities, however, the quality of the buzz is changed profoundly. It is the shape of the vocal tract, including the positions of the larynx, the

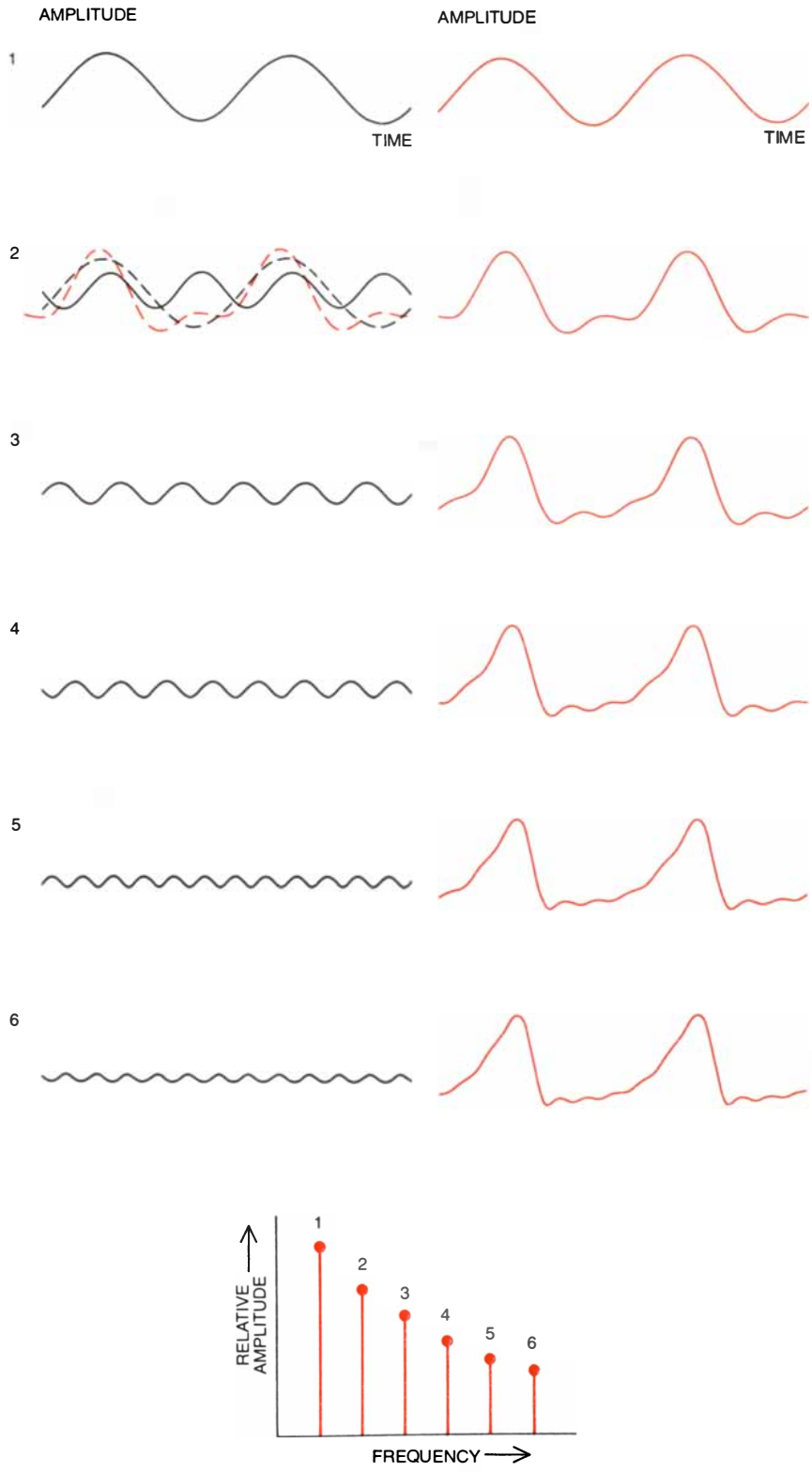
tongue, the lips and the velum, that distinguishes (for example) the "ee" sound in "me" from the "oo" sound in "you."

One way of understanding this acoustic transformation is the mathematical technique called Fourier analysis. In 1822 the French mathematician Jean Baptiste Joseph Fourier showed that any periodic waveform can be represented as the sum of an infinite series of sine waves. A periodic waveform is one that is repeated at uniform intervals. If the interval of repetition is t seconds, the fundamental frequency of the waveform is $1/t$ hertz. In the Fourier series for a periodic waveform, the frequencies of the component sine waves are harmonics, or integral multiples, of the fundamental frequency of the waveform being analyzed, and they must be assigned appropriate amplitudes and phases. The Fourier transform is a generalization of the Fourier series; it allows analysis of aperiodic waveforms. Thus the noisy hiss of an "f" sound can be represented as a sum of sinusoidal components all along the frequency continuum.

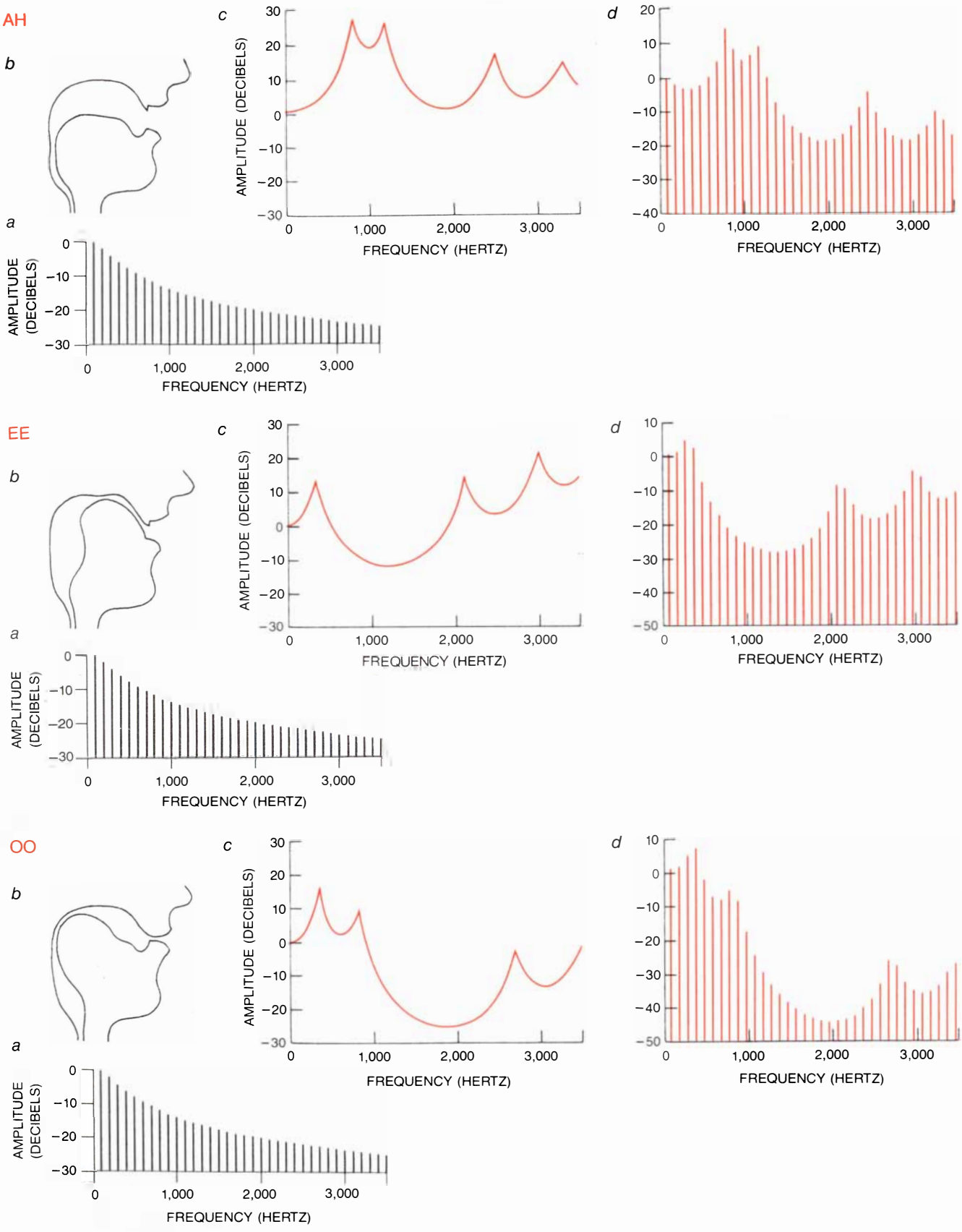
The most obvious way to represent sound waves is to graph the variation of air pressure with time. Fourier's result implies that the same information can also be displayed by a graph that shows amplitude and phase as a function of the frequency of the sinusoidal components. Because phase differences are of little perceptual significance, a speech sound can be represented in practice by its amplitude spectrum, a graph that shows the amplitude of the sine-wave component at each frequency.

What is the acoustic effect of the shape of the vocal tract on the sound emitted? When the sounds are represented by their amplitude spectra, the effects are clear [*see illustration on next page*]. The vocal tract acts as a filter on the spectrum of the sound source, enhancing some frequencies and diminishing others. The selective filtering can be described by a mathematical expression called a transfer function; a separate transfer function is associated with each position assumed by the organs of the vocal tract. The transfer function usually has several well-defined frequency peaks, called formants, in which most of the energy from the sound source is concentrated.

It is now possible to state with some precision why it is hard for a computer to make the translation from sounds to words, from an acoustic characterization of an utterance to a linguistic characterization of the intended message. One source of difficulty is that the organs of speech do not take up a series of fixed configurations corresponding to units of the message. Instead parts of the vocal tract are in constant motion along smooth trajectories. Some investigators



PRINCIPLE OF SUPERPOSITION allows temporal variation in the sound pressure of the signal to be represented as a spectrum of sound amplitude, or energy, at different frequencies. The amplitude spectrum is generally a more useful way of displaying acoustic information. The waveform (here a glottal wave) can be treated mathematically as a pattern that repeats indefinitely in the past and the future at a fundamental frequency. As the French mathematician Jean Baptiste Joseph Fourier showed, any such waveform can be decomposed into a series of sine waves at integral multiples of the fundamental frequency, with various amplitudes and phases. When the sine waves are combined by adding the amplitudes at each point, the result is the original waveform. When the amplitude of each sine wave that makes up the decomposition is graphed as a function of its frequency, the result is an amplitude spectrum.



VOWEL SOUNDS result from various configurations of the mouth, lips, tongue and velum (soft palate). The resulting shape of the vocal tract can be modeled by a series of resonating cavities that enhance energy at some frequencies and diminish it at others in predictable ways. Such filter-response characteristics can be represented by a transfer function (*c*) for each position of the model of the vocal tract (*b*). When the input sound energy is periodic (which is almost the case for vocal-cord vibration), both the input spectrum (*a*) and the output spectrum (*d*) are line spectra. In a line spectrum the sound energy is concentrated at harmonics, or integral multiples, of the vocal-cord frequency. An aperiodic sound source such as a whispered vowel has no discrete lines in its spectrum, but the shape of the output spectrum still matches that of the transfer function. Model vocal-tract configurations for vowels "ee," "ah" and "oo" are shown.

believe these motions "flow" through a sequence of target positions defined by linguistic units such as consonants and vowels. Others think even the simplest linguistic units are inherently dynamic. In any case the result is a complex and continual motion, which is inherited by the emitted sound in the form of a constantly changing amplitude spectrum. Such patterns of changing sound quality can conveniently be represented by a sound spectrogram, a graph in which time proceeds from left to right, frequency increases from bottom to top and amplitude increases from white through shades of gray to black.

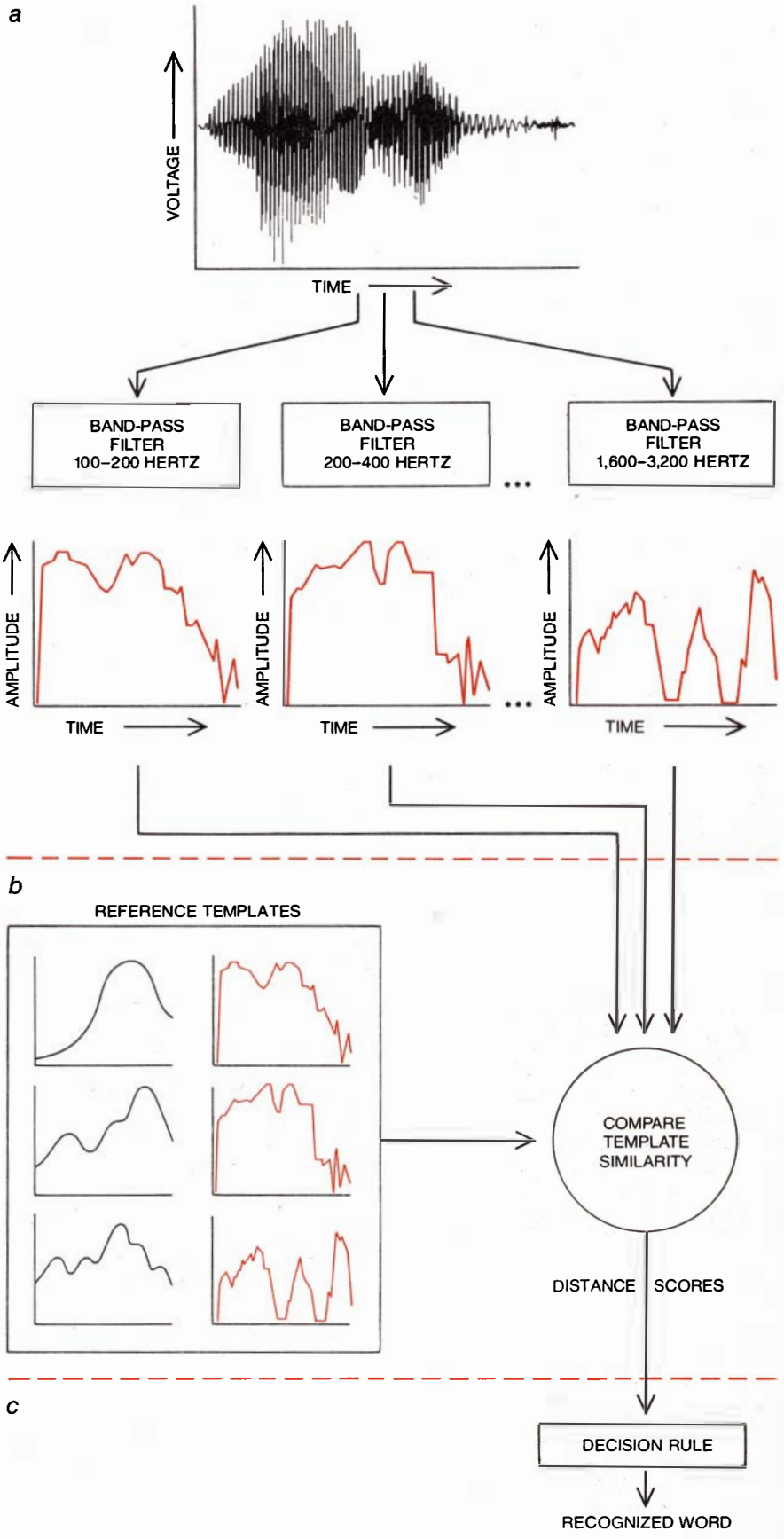
Motions of the vocal tract that correspond to linguistic units usually overlap and combine with their neighbors. For instance, in saying "coo" the lip-rounding of the vowel "oo" usually precedes the tongue motion of the initial consonant. Hence the acoustic effects of the two motions are combined from the beginning of the word. In fluent speech such amalgamation also applies between one word and the next. The effects are sometimes quite plain to the ear. When the "t" of "cat" combines with the "y" of "your," it makes the phrase "You gave the cat your dinner" sound like "You gave the catcher dinner."

Other variations in the sound of a word result from its position in a phrase, from its degree of emphasis and from the rate at which it is pronounced. The size and shape of the vocal tract vary from individual to individual, and habits of speech differ widely according to age, sex, geographic region and social background. Furthermore, the signal that reaches a speech-recognition device is influenced by various circumstances in addition to the sounds made by the speaker, such as room acoustics, background noise and the characteristics of the transmission channel.

For these reasons it is hard to divide a speech signal into chunks corresponding to the elements of the message the signal conveys, and it is difficult to translate pieces of acoustic information into information about the identity of pieces of the message. People find speech easy to understand, and so the needed information must be present in the signal. The trick is to find it.

A natural starting place is the recognition of words. Words are generally distinct from one another as elements of a linguistic system, and they constitute natural and relatively stable patterns for an automated speech-recognition system. Although speech is more than a sequence of words, it is at least such a sequence, so that a crucial function of a speech recognizer is identifying words. If a speech-recognition system can recognize words accurately, it will succeed; if it cannot, it will fail.

Most speech recognizers now in use are not capable of recognizing words in



STANDARD METHOD OF WORD RECOGNITION employs the basic principles of pattern recognition to discriminate among acoustic patterns. The speech waveform is measured and analyzed (*a*), in this case by filters that divide the signal into frequency bands, each band being an octave wide. The output of each filter is the energy in its band. The outputs are compared with stored reference templates, and distance scores are assigned to each template (*b*). A decision procedure then classifies the input utterance on the basis of the distance scores (*c*).

connected speech. Instead recognition is carried out on isolated words by a process of acoustic pattern recognition. Generally the user must "train" the machine by speaking into a microphone all the words the system is to recognize. In some cases the training is limited to one utterance of each word by a few of the speakers who will use the system. In other applications every potential user must say each word several times. The result of this training process is a set of stored "templates," which represent typical acoustic patterns for the words in the vocabulary.

When a word is presented for recognition, the machine analyzes the acoustic signal, compares the results of the analysis with the stored templates and decides which one of the templates most closely resembles the spoken word. The machine may also list other possible matches in decreasing order of similarity. Once a classification has been made the machine can respond to the user's utterance or issue an appropriate signal to some other device. Each stage of the template-matching procedure (analysis of the speech signal, comparison with the template and classification of the signal) can be carried out by a variety of techniques.

The aim of all methods of analyzing the speech signal is to characterize the temporal variation of the signal's amplitude spectrum. Perhaps the simplest method of estimating the spectrum is the zero-crossing count. This method consists in counting the number of times the voltage analogue of the speech signal changes its algebraic sign (from plus to minus or from minus to plus) in a fixed interval. The number of such axis crossings is related to the frequency.

One refinement of the zero-crossing method filters the speech signal into three frequency bands. The zero crossings are measured separately in each band to give rough estimates of the first three formant frequencies. Such measurements are useful in classifying vowel sounds, and for small vocabularies of easily distinguished words these measurements alone are sufficient for discrimination. The zero-crossing method is economically attractive because it can be accomplished by simple electronic devices.

A more elaborate procedure for spectral estimation is the filter-bank method. The speech signal is divided by filtering into between 20 and 30 frequency bands, covering the frequency range of human speech. The output of each filter is a measure of the energy in that frequency band. The energy levels are suitable for direct comparison with those of a template. The Fast Fourier Transform provides a general, computationally efficient method for estimating the amplitude spectrum of a signal from its time-domain waveform. This algorithm provides one of several ways to obtain filter-bank information in purely digital form.

Recently a new method for estimating the amplitude spectrum of speech, called linear predictive analysis, has been introduced. Actually statisticians have employed the method for some time under the name autoregressive analysis. The method predicts the amplitude of a speech wave at a given instant from a weighted sum (or linear combination) of its amplitudes at a small number of earlier instants. The coefficients, or weights, that give the best estimate of the true speech wave can then be mathematically converted into an estimate of the amplitude spectrum. For the analysis of speech linear predictive analysis is particularly appropriate because it is mathematically equivalent to treating the vocal tract as a pipe of varying circular cross section, or in other words as a sequence of resonant cavities. The model is quite faithful for nonnasalized, voiced speech. Because it is a model of the vocal-tract resonances and not of vocal-cord vibration, the linear-prediction spectrum is smooth. None of the pitch harmonics are in evidence. Consequently the formant structure of the speech wave, which is important for speech recognition, is brought clearly to the fore.

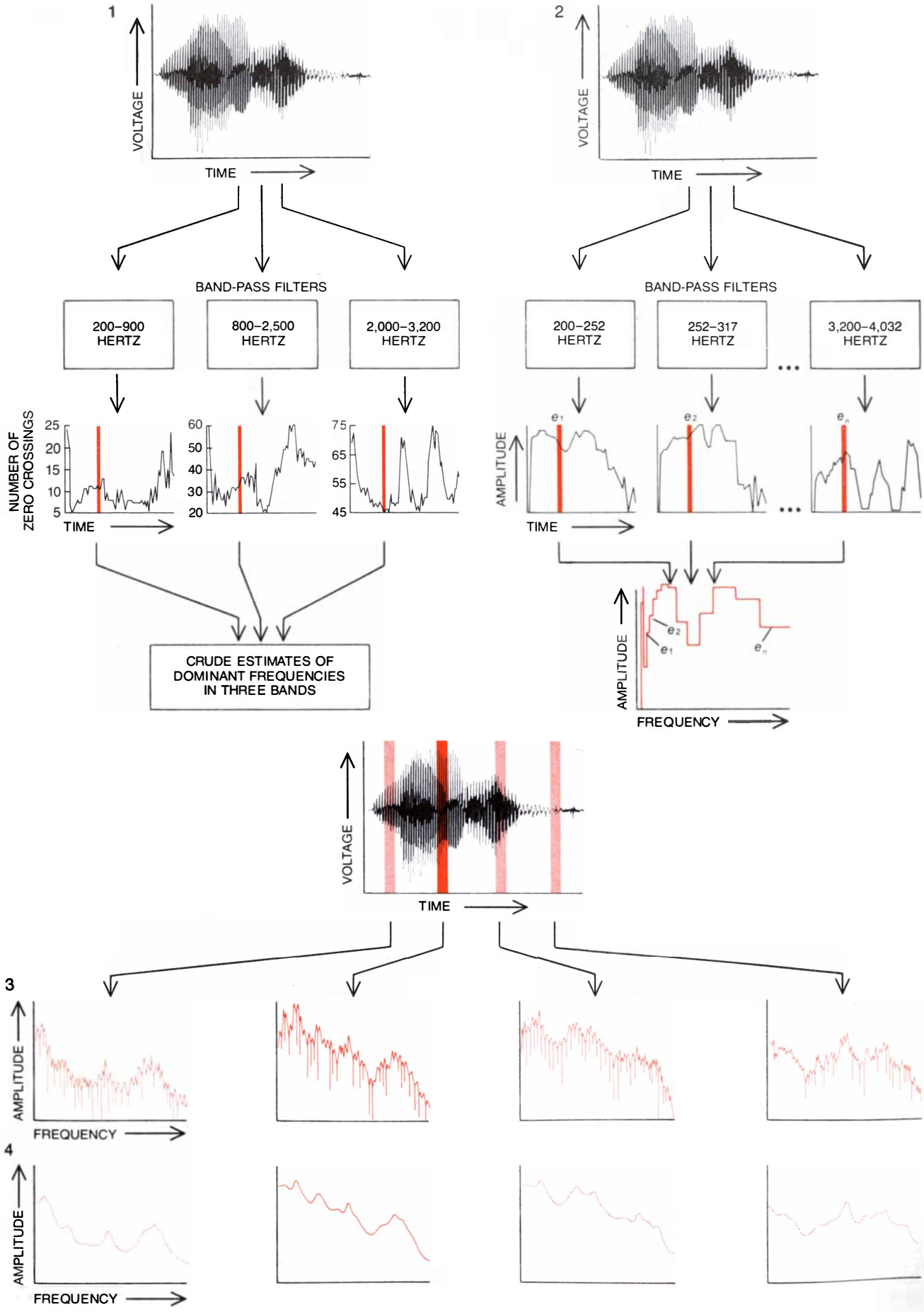
During the comparison, or template-matching, stage the phonological structure of a word can be exploited in an indirect way. A spoken word consists of a sequence of vocal gestures, which gives rise to a time-varying pattern of sound. The parts of the sound pattern rarely have the same durations in different utterances of the same word, but their sequence is more nearly constant. For example, the word "fable" begins with an "f" noise, which is followed by a pattern of moving formants that show the lips opening out from the "f" and closing again for the "b" while the tongue is moving through the first vowel; next there is a "b" lip closure, and finally there is another pattern of spectral motion as the lips open and the tongue moves into the final "l." On different occasions the timing of these patterns may vary considerably, but they must all be present in the described order if the utterance is to count as a reasonable rendition of the word "fable."

Because of differences in timing, the various parts of a word may be badly out of alignment with the corresponding parts of the template it is to be matched against. Since the order of events is fairly constant, the misalignment can be corrected by stretching the template in some places and compressing it in others so that a mathematically optimum match is found. Nonuniform temporal alignment is accomplished by means of a procedure called dynamic programming. Dynamic programming was developed by Richard E. Bellman of the University of Southern California School of Medicine for solving problems in the design of servomechanisms. It is a technique for mathematical optimization that is often carried out with the aid of a computer, but it should not be confused with computer programming itself.

Comparison implies some estimate of the degree of similarity between the sound of the input and the sound represented by the stored template. The final aspect of processing common to all word recognizers is a decision strategy, which is usually based on a statistical measure of closeness of fit. Each template is assigned a point in an abstract space; the position of the point is defined by the spectral characteristics of the template. The utterance to be classified is represented as a point in the same space. The recognizer calculates the distance in the space between the utterance and each of the templates. It then picks either the template closest to the utterance or the equivalence class of templates that is closest to the utterance in a statistical sense.

The performance of automatic recognition systems in identifying isolated words is poor compared with that of people. Even for the most powerful word recognizers the number of errors rises rapidly as the vocabulary increases to more than a few hundred words. The error rates get worse still when unknown speakers and acoustic conditions are introduced. In a recent experiment isolat-

METHODS OF ESTIMATING the amplitude spectra of short intervals of a word (here the word "language") all seek to highlight linguistically relevant information in a computationally efficient way. Zero-crossing counts exploit the fact that as the frequency increases, the number of times the voltage analogue of the acoustic signal changes its sign increases as well. In the band-pass-filter method the signal is divided into several frequency bands and the amount of energy in each band is measured. These measurements yield an amplitude spectrum for the interval. The Fast Fourier Transform is a general, computationally efficient algorithm for estimating the amplitude spectrum of the signal from its time-domain waveform. It is one of several ways of computing filter-bank information in digital form. The rough appearance of the spectrum is caused by pitch harmonics or other fine structure in the spectrum. The fourth method of spectral estimation, called linear predictive analysis, employs a model of the vocal tract to generate successive frequency spectra. Its advantage is that a smooth, continuous spectrum is generated for each sample. The spectra in dark color are all constructed from the same interval of the time-varying signal. Several other methods of spectral estimation are also in use.



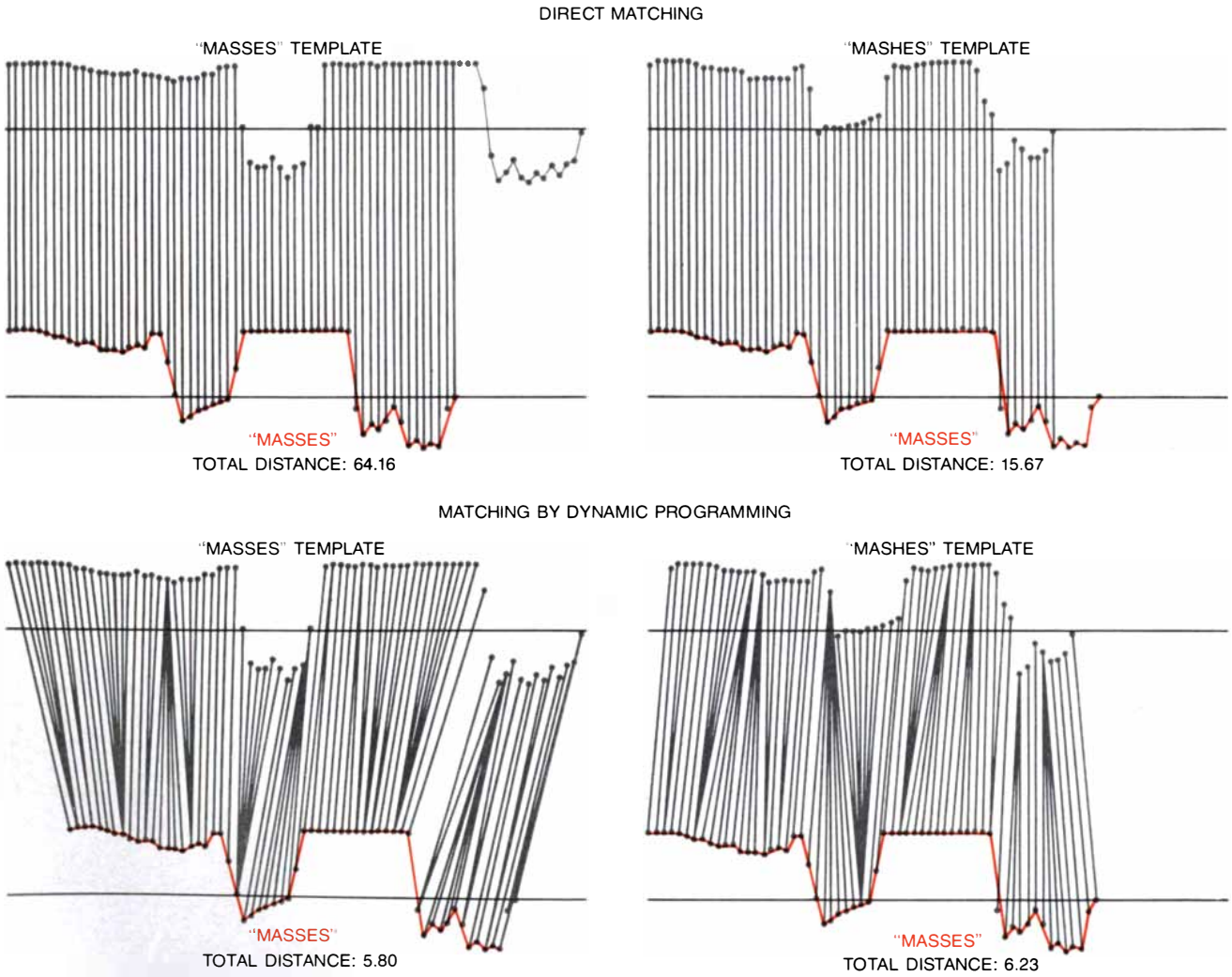
ed words from a 26,000-word vocabulary were spoken by a variety of speakers unknown to the listener; the words were identified with an error rate of less than 3 percent. Human word-recognition abilities are also remarkably tolerant of background noise: conversation can be understood even at a noisy party. No existing recognition system can approach this level of performance.

In attempts to recognize continuous speech the disparities between human and computer performance are even more evident. People generally find it easier to recognize words in context, but for an automated system the recognition of fluent speech is far more difficult than the recognition of words in isolation. One of the crucial problems is coarticulation, which causes the blending at the boundaries between words and makes the spectral patterns to be recognized highly complex and unstable. In fluent speech there are no clear acoustic signs of word boundaries and direct template-matching becomes extremely difficult. In essence every template must be aligned with every possible interval of the utterance by means of a variant of the dynamic-programming method.

The computational burden is somewhat reduced by the requirement that the intervals be contiguous, so that the end of one word meets the beginning of the next. Still, the combinatorial complexity of the process increases too fast for it to be considered a practical solution to the general problem of recognizing continuous speech. Direct template-matching can be useful only where the range of possible utterances is small. With present technology the technique can work in real time (that is, as fast as the utterance is spoken) for sequences as many as five words long, drawn from a vocabulary of about 20 words.

Instead of looking for every possible pattern everywhere in the signal, a continuous-speech-recognition system can search for linguistic units in a more constrained way, such as in sequence from the beginning of the utterance to the end. The speech signal is divided into intervals that correspond to specific acoustic patterns, and the intervals are classified in a way that matches the categories of a potential linguistic message as closely as possible. We shall call such techniques segmentation and labeling. The processes of segmentation and labeling can be carried out in many ways, and the intervals to be found can correspond to words or to smaller linguistic units such as syllables, phoneme pairs or phonemes.

The easiest way to achieve automatic segmentation and labeling is to require the user to pause briefly between words. The pauses that appear as intervals of



COMPARISON STAGE of word recognition is carried out by compressing and stretching stored templates according to an optimization process called dynamic programming. For each stored template, dynamic programming seeks to associate every frame of the input word with some frame of the template in such a way that a distance measure of overall fit between the input and the template is minimized. The nonuniform time alignment of the stored template with the spoken word allows for variations in the rate of speech and in the relative lengths of the vowels and consonants in a word. Here matching the templates (*black*) to the input (*color*) without dynamic programming yields a misidentification, indicated by the distance scores, that is corrected when the compression and expansion procedure is applied. Dynamic programming is often done with the aid of a computer but should not be confused with computer programming.

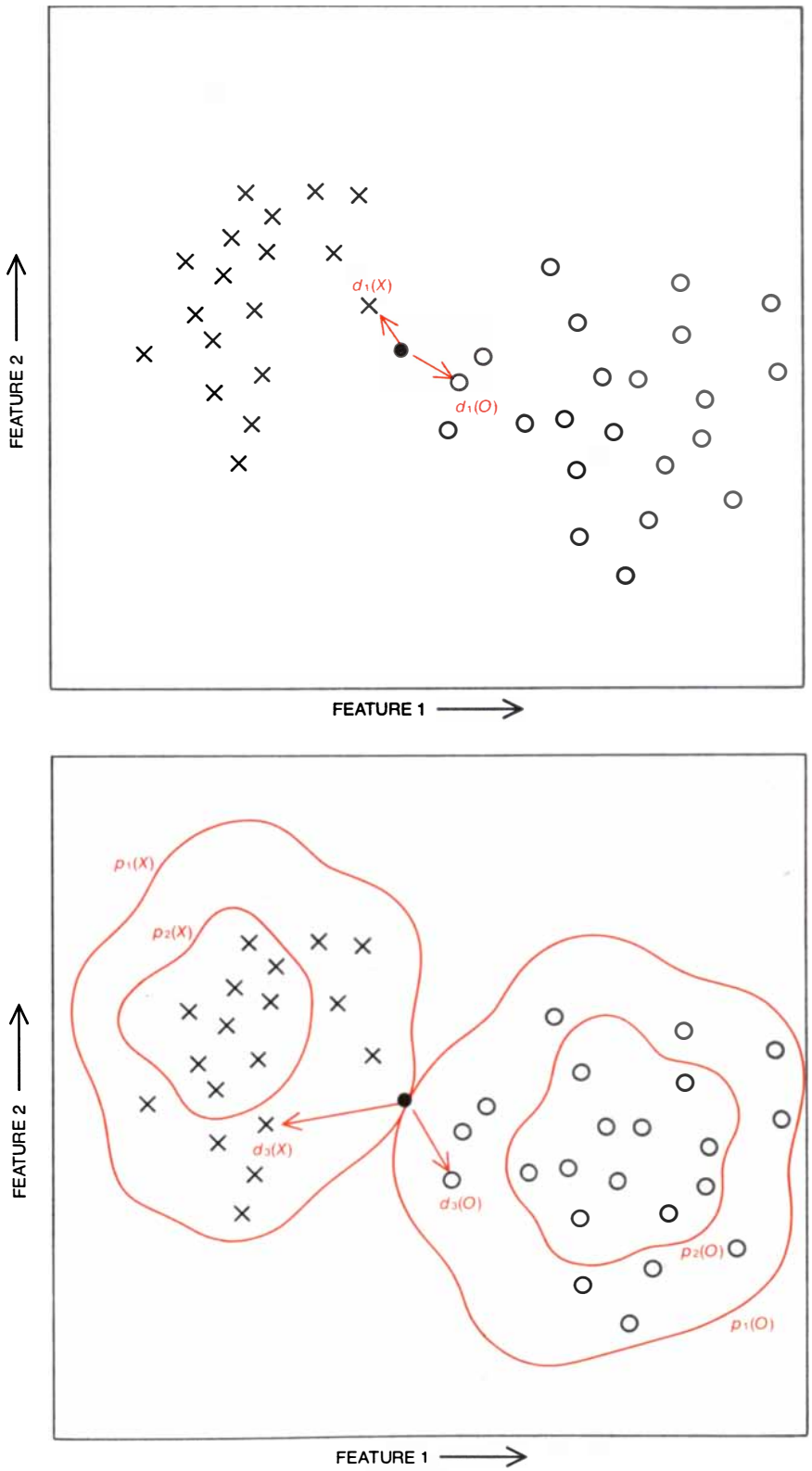
low sound energy are a reliable indication of word boundaries. Once the words have been segmented they can be analyzed independently. Although this method works well, it does not really address the question of fluent speech recognition. Other methods of segmentation and labeling are available.

Discontinuities in the spectrum, peaks and valleys in the energy of certain frequency bands and other acoustic signs provide clues to articulatory events: the closing or opening of the vocal tract or the beginning or ending of laryngeal vibration. This suggests that segmentation and labeling might be carried out on the basic phonological units of which words are constructed.

Blending and the diffusion of acoustic information across boundaries affect the acoustic shape of the smaller speech units even more than they affect words. As a result such units are difficult to identify by template-matching, and segmentation errors would probably be at least as frequent for the smaller speech units as they would be for words. Nevertheless, there may be reason to favor segmentation into smaller units as the vocabularies of speech recognizers become larger.

There are some 300,000 words in English, far too many for all of them to be tested by template-matching. Moreover, it is difficult to allow for the effects of blending at the boundaries of words when word templates are employed. English syllables number some 20,000, which is still too many for them to be identified easily and reliably. In addition the effects of blending at boundaries are even more disruptive to template-matching with syllables than they are with words. In contrast, there are only about 40 English phonemes (basic linguistic elements such as consonants or vowels), and the phonemes can be further decomposed into about a dozen phonological features that specify distinctive characteristics of vocal-tract shape and larynx control. Such features can also be combined directly into syllable-like units. As the set of linguistic units is reduced in number, however, the relation of the units to patterns of sound becomes more abstract, more complex and less well understood. Segmenting and labeling such small speech units by currently available techniques leads to high error rates. Still, if constraints imposed by the linguistic code can compensate for the errors, or if more reliable methods of analysis can be found, the small number of the basic phonological units will give them a decided advantage as the fundamental elements of a recognition system.

There is one difficulty that is common to all segmentation and labeling procedures: the probability of error is much higher in making a number of independent classifications than it is in making



CLASSIFICATION OF AN INPUT SOUND consists in finding the shortest distance in a space of acoustic features from the input (represented by a dot) to a stored template or class of templates (represented by *X*'s and *O*'s). The simplest decision strategy picks the closest template (*upper graph*), and so the input is classified as the sound "ah" (an *X*). When several templates represent linguistically equivalent sounds (as when the computer must recognize the voices of several speakers), the decision strategy may take account of entire classes of templates. One method calculates the distance from the input to the third-nearest neighbor in each class (*lower graph*); here the input is classified as the sound "aw" (an *O*). Under certain conditions it is possible to draw equal-density contours along which the number of template samples per unit area is constant. The highest-density contour that passes through the input can then be found; because $p_1(X)$ is greater than $p_1(O)$ the input is classified as the sound "ah" (an *X*).

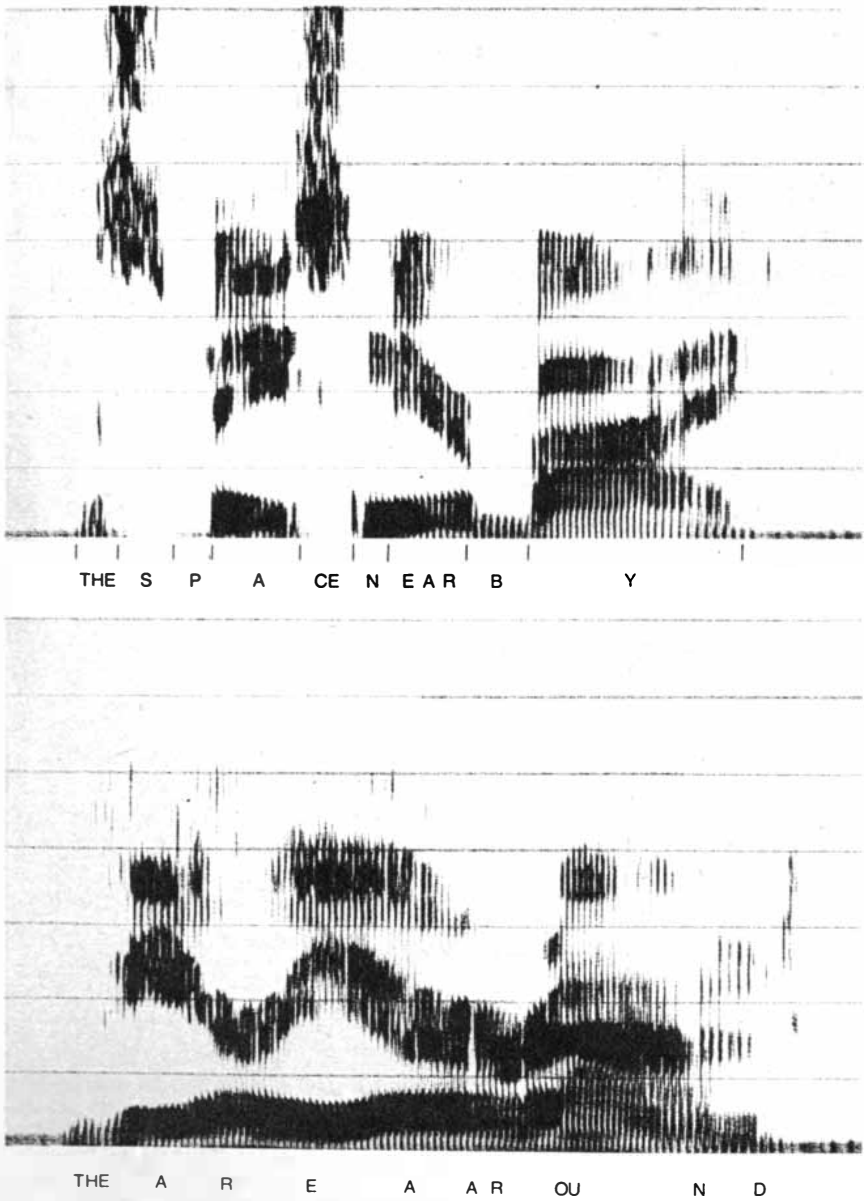
a single classification. In a three-word phrase, even if the probability of recognizing the correct word in any given position is .8, the probability of recognizing the entire phrase correctly is only about one-half ($.8 \times .8 \times .8$).

One way of offsetting this effect is to introduce constraints imposed by the linguistic code, such as allowable sequences of words in a sentence or allowable sequences of syllables in a word. An area of mathematics called formal-language theory provides several methods for specifying and using such constraints. By applying some of the elementary principles of formal-language theory it is possible to write precise and efficient descriptions, or formal grammars, of linguistically possible sequences of sounds and words. One can also write computer programs that utilize these grammars to recognize formally correct linguistic sequences.

One simple way of exploiting grammatical structure makes use of a mathematical construction called a state diagram. A state diagram defines every possible sentence the machine can recognize. Each path from the starting point of the diagram to the end points represents an acceptable sentence. From acoustic measurements the recognizer assigns a probability to each transition in the diagram. A probability can then be calculated for each path by forming the product of the probabilities of all the transitions that make up the path. The sentence chosen is the one represented by the path with the highest probability. This technique can significantly reduce the error rate in sentence recognition: it can choose a word with a relatively low probability in a given position in order to enhance the likelihood that the overall transcription is correct.

Such a reduction in the error rate was demonstrated in a phoneme-based system for the recognition of fluent Japanese, which was tested at Bell Laboratories and at the Nippon Telegraph and Telephone Electrical Communication Laboratories. The segmentation and labeling of phonemes was correct only 60 percent of the time. Syntactic processing, however, led to a 70 percent success rate for the recognition of sentences with an average length of 25 phonemes. Although 70 percent recognition is not adequate for reliable communication, the result is remarkable in view of the small probability of finding a correct sentence without syntactic processing: it comes to about one chance in three million.

A state diagram can also improve the efficiency of continuous-speech recognition by nonlinear time alignment. Instead of matching every template to every interval in the input sentence, the recognition system tests only those templates that fit admissible sequences described by the state diagram. This procedure eliminates much wasted computation, since only a small subset of the words in the vocabulary can appear at a given position in a sentence. A device employing syntax-directed time alignment can recognize connected sentences of more than 20 words composed from a vocabulary of more than 100 words.



SEGMENTATION OF ACOUSTIC PATTERNS into words or other linguistic units is difficult because of the temporal smearing of speech sounds. Certain sounds, however, exhibit more spectral discontinuity than others. The alternation of the consonants and vowels in the phrase "The space nearby" shows up in relatively clear discontinuities. The smooth sequence of blended vowels in the phrase "The area around" makes segmentation more difficult.

So far we have described the phonological symbols that correspond to the acoustic reality of speech and the grammatical organization of the symbols into words and phrases. These form the linguistic code of speech. The purpose of the linguistic code is to convey meaningful messages: semantic information. Hence semantic information imposes additional constraints on the way the symbols of a language can be combined to form messages.

A machine that processes the semantic information encoded in speech attempts a much more complex and subtle task than a machine that merely recognizes words. In order to deal with meaning a machine not only must recognize acoustic patterns but also must manipulate abstract representations of reality. In other words, it must simulate at least some important aspects of human intelligence.

At Bell Laboratories we have incorporated a rudimentary semantic proc-

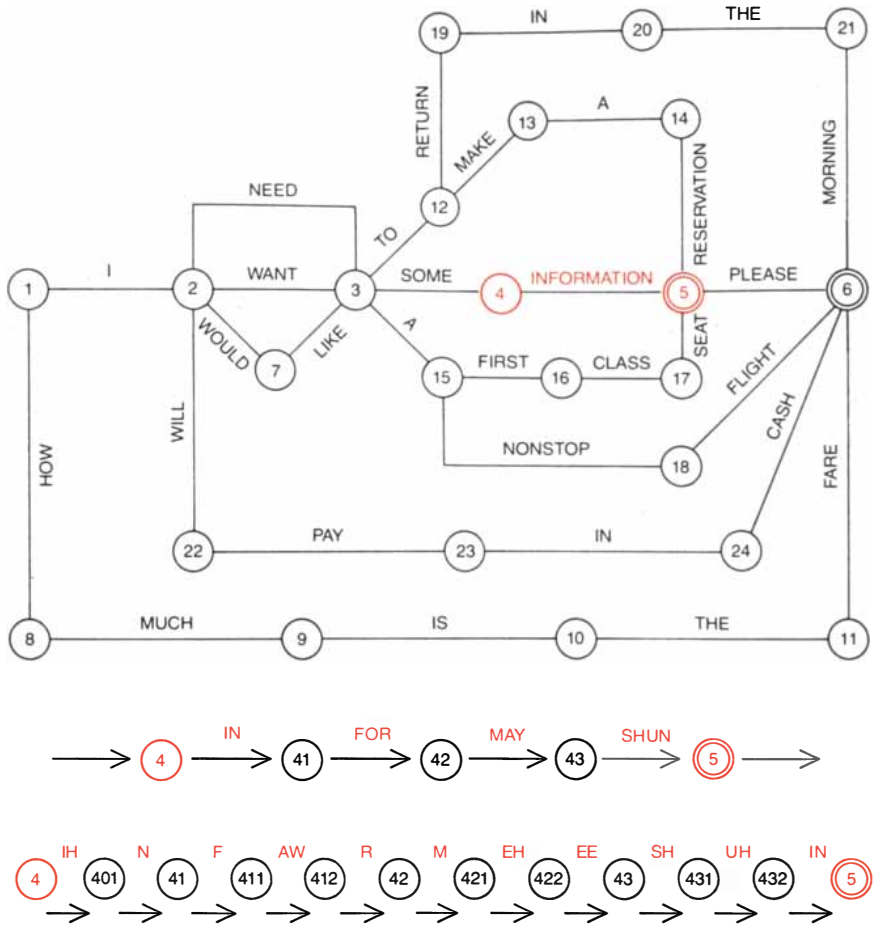
essor in a system designed to emulate the entire process of human speech communication. The user communicates with the system by telephone. The computer, which is intended to function as an airline ticket agent, responds in a synthetic voice. The integration of the necessary functions into a single device has enabled us to study the interaction of the subsystems and their control.

As a complete simulation of human communication the Bell Laboratories machine is the most advanced system known to us. The individual components, however, are less advanced than those of experimental systems in other laboratories. There are speech-recognition systems that work with vocabularies much larger than the 127 words our machine recognizes, and there are systems with a more flexible syntax. There are more sophisticated semantic processors that accept typed input instead of speech. There are processors that respond faster than ours does. A question that is asked in 10 seconds receives a reply on our system after about 50 seconds. We hope to improve the performance of all the building blocks of our system.

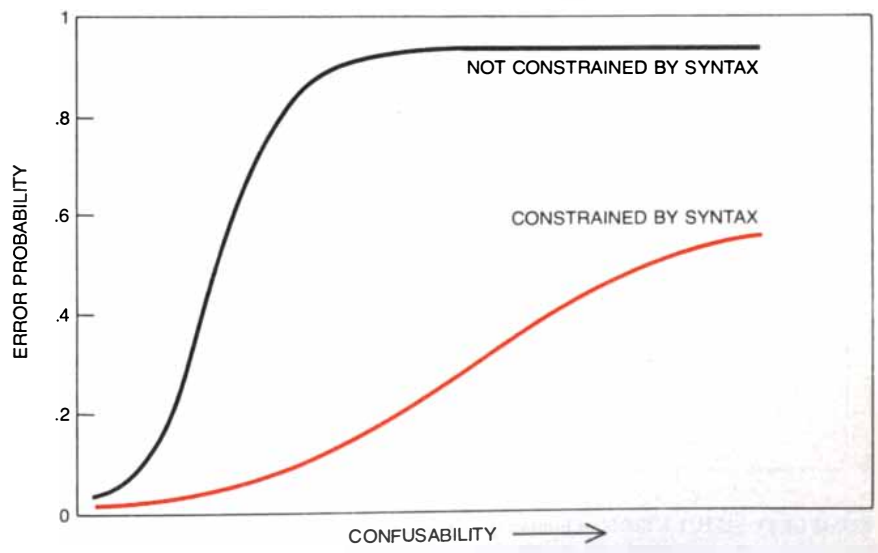
In the airline-information system the acoustic processor and the syntactic parser are coupled, so that the acoustic processor tests each hypothetical word-identification made by the parser for agreement with spectral information. The rest of the system, with the exception of two memory units that are shared by all the components, is devoted to semantic processing.

The semantic processor incorporates a world model, whose state can be changed as a conversation progresses, and a memory module, which cannot be altered. The world model is based on a set of concepts, each of which can take on a number of values. Among the concepts are "destination," "departure day" and "departure time." During a particular conversation these categories might be assigned the values "Boston," "Tuesday" and "5:00 P.M.," whereas another state of the world model might correspond to the values "Chicago," "unknown" and "unknown." The semantic processor determines a new state from its present state, from the words in the input sentence and from the transitions in the state diagram that were employed in generating the sentence. The need for the latter two sources of information reflects the fact that semantic content is a function both of the words and of their relations in a sentence.

The memory units store two kinds of information: facts and procedures. Facts are of two types. Airline schedules are stored as a portion of the *Official Airline Guide,* but relations among the concepts in the *Airline Guide* must also be stored. If the system is asked the elapsed time of a flight, it can calculate



FINITE-STATE GRAMMAR is computationally the most straightforward means of imposing syntactic (or word-order) constraints on the recognition of sentences. The grammar diagrammed here would force the computer to classify every sequence of acoustically possible words as one of the 26 sentences that can be traced through the state diagram, starting at state 1 and ending at state 5 or 6. For instance, one possible sentence is "I would like a first-class seat, please." The principles of the diagram can also be extended to levels of analysis lower than the level of the word, such as the syllabic and phonemic analyses of the word "information." The grammars of experimental recognition systems allow for billions of sentences.



CONFUSABILITY OF A SPEECH SIGNAL is a complex function of the size of the input vocabulary, the acoustic similarity of the elements to be discriminated, the number of speakers to whom the system must respond and the amount of noise in the communication channel. Errors tend to become more frequent as confusability increases. Syntactic contraints can significantly reduce the effect. This error pattern is as true for human listeners as it is for machines.

the time from the listed departure and arrival times; for it to do so, however, the time zone of each city must be available. (In the *Official Airline Guide* all times are local.)

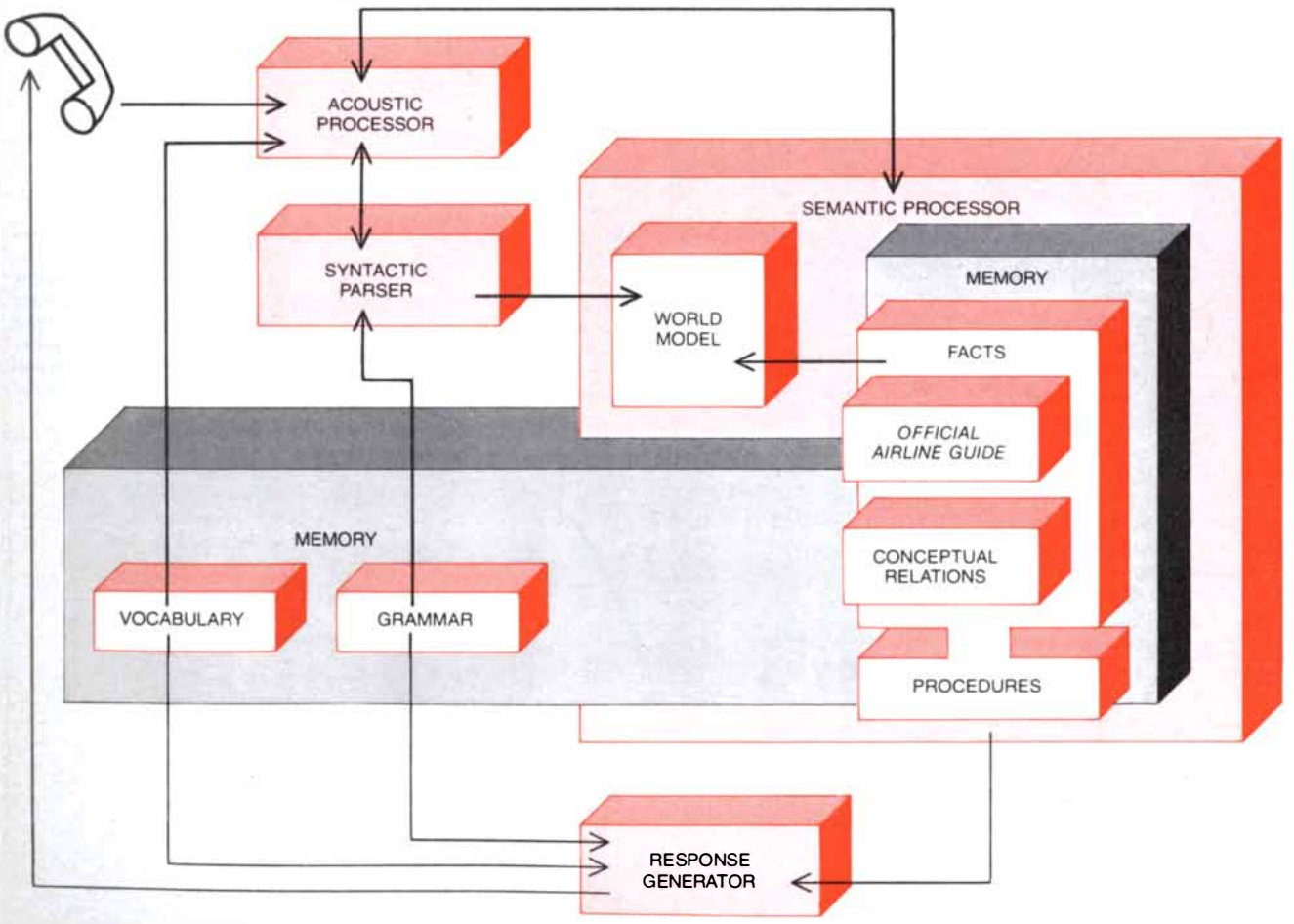
Procedures are special-purpose programs that use stored facts to derive new information from an input and from the current state of the world model. For example, one program is a perpetual calendar, which can find the day of the week for any given date. The conversion is needed because a question may specify only a departure date, whereas the *Official Airline Guide* is organized by day of the week.

When an internal instruction calls for a reply to the user, the system activates a linguistic encoder. The semantic analyzer tells the encoder what concepts are to be communicated from the world model. Then the encoder retrieves grammar and vocabulary from memory and transforms the concepts into a sequence of symbols. The speech synthesizer transforms the sequence into speech [see "The Synthesis of Speech," by James L. Flanagan; SCIENTIFIC AMERICAN, February, 1972].

In what ways can the art of speech understanding be advanced? We see two basic aims. For the near term it is important to seek a better grasp of the fine structure of speech communication. This should include detailed information about the kind of signal analysis done by the human ear and a better understanding of the relation between sound symbols (such as phonemes and syllables) and actual sounds. More efficient ways of exploiting this information must be developed and incorporated into recognition systems.

For the long term several areas of investigation may bring significant advances. We have stressed that the speech code includes a number of coexisting kinds of structure, such as phonology, syntax and semantics. A general theory of such complex codes is needed, particularly so that the interactions of the levels can be coordinated and controlled. It is also desirable to gain a better understanding of the processes by which people acquire a language. Although present speech recognizers are "trained," the training is rudimentary and cannot be altered through "experience." We believe this lack of adaptive abilities is a serious disadvantage. The best design strategy is not to program a computer directly with the wealth of descriptive detail that constitutes a natural language but rather to give it the basic set of expectations and abilities that are needed to learn a language.

It is hard to predict how well these investigative strategies can ultimately succeed in approximating natural speech communication. Whatever the rate of progress, this goal will continue to be pursued. Some success is guaranteed, and wisdom will be required in its application.



COMPLETE SIMULATION of human speech communication is attempted by an automated system constructed by the authors and their colleagues at Bell Laboratories. Functional relations among major parts of the system are shown here in a block diagram. The user asks for information about airline timetables over the telephone, and the computer replies in a synthesized voice. Heavy arrows trace the flow of information related to speech recognition. The generation of a response is traced by lighter arrows. Memory modules concerned with semantic processing include facts and procedures related to flights and reservations. Nonsemantic memory stores vocabulary templates and grammatical rules used both in speech recognition and in speech synthesis. The semantic processor also includes a world model, which is constantly updated with data based on the user's questions and on the information in the semantic memory.