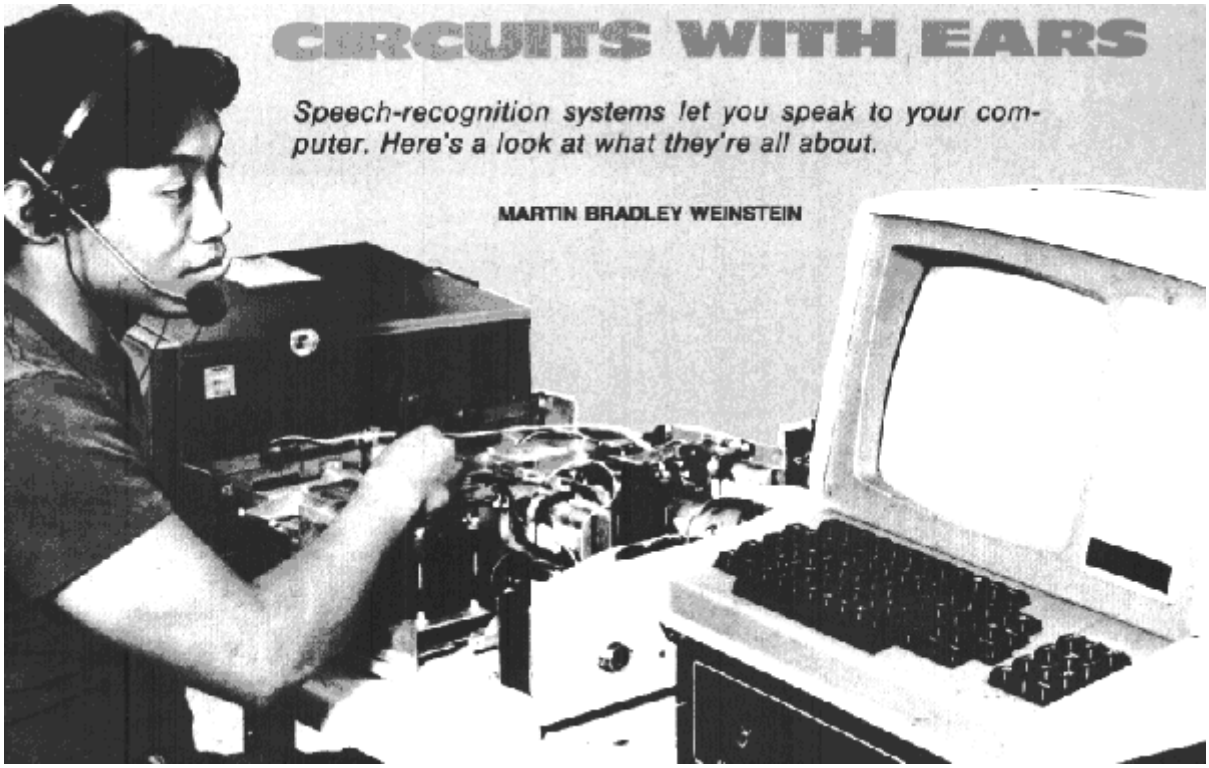
sp recog Circuits with ears

CIRCUITS WITH EARS

Dropbox



NEW TECHNOLOGY

Most of us remember HAL, the unlike star of Stanley Kubrick's 2001: a Space Odyssey. When the film was made in the late 1960's, the idea of conversing with a computer was pure science fiction. But as has happened so often, yesterday's science fiction is today's technology. Speech synthesis by computers is familiar to most of us by now, but what about the area of speech recognition? While speaker-independent recognition of connected speech by a computer is still a decade off, firm toeholds have been established in developmental areas important to its eventual success. We're going to take a look at some of the speech-recgnition systems now available, and what's being done to improve the state of the art.

A simplified explanation

Speech-recognition systems rely on the matching of a spoken word to a stored model of that word. In practice, the way words are modeled is the key of the success and accuracy of a system - as well as to its expense, speed, and more. Generally, the user of the system is asked to speak each word in the system's limited vocabulary several times. Those spoken samples are analyzed by a variety of techniques, and the samples of each word or phrase are compared to one another.Differences are minimized, similarities maximized, and the resulting model (called a template) is stored in the system's memory.

Once the "training" has been completed, any word or phrase spoken into the system is analyzed using the same techniques used in deriving the templates. That analyzed data is compared with the stored templates, and a score assigned to each match. If no score is high enough to be accepted as a fit, the system gives a "non-recognition" message or asks to have the word repeated. If more than one word is scored high enough so that there are several possible fits, the system can ask which is correct or ask to have the word repeated. However, in about 98% of all attemps, a single word is reconized uniquely.
Since it's important in matching a word to know precisely where a word begins or ends, there is usually some hardware or software incorporated to give that information. Also, there is usually some provision for normalizing the time distribution of the word - that is to say, the duration of the voiced sounds within the word. Without time normalization, variation in the ways we pronounce a given word would make matching it against its template very difficult.
Generally, both time-dependent and time-independent analyses are done. The time-independent analysis is usually concerned with the spectral distribution of the word. For example, a spectral distribution analysis (called a histogram) of the word six would show that the word has a lot of s sound within it, but not that the s sounds occurs twice, once at each end of the word. Rather, the spectral histogram would show how much energy appeared at any one frequency during the speaking of the word. In practice, narrow bands of frequency during the speaking of the word. In practice, narrow bands of frequencies are usually samped - although there is some progress in the Fourier analysis of speech through new hybrid analog/digital microprocessor technology, but thats's a subject best left until it can be covered somewhat more meaningfully.

How it works

Let's take a look at the elements of most of today's speech-recognition hardware in a little more detail. **(see fig. 1)** . The first step is to provide as favorable a signal-to-noise ration as possible. A noise-cancelling mike close to the speaker's mouth (often on a headset) and push-to-talk operation help to accomplish that.
Also, there is usually some preemphasis and shaping of the incoming audio to help eleiminate background noise and help accentuate some of the weaker segments of the speech spectrum. Some form of automatic gain control is usually used, either in the form of an analog compressor or as a part of the computer's task.

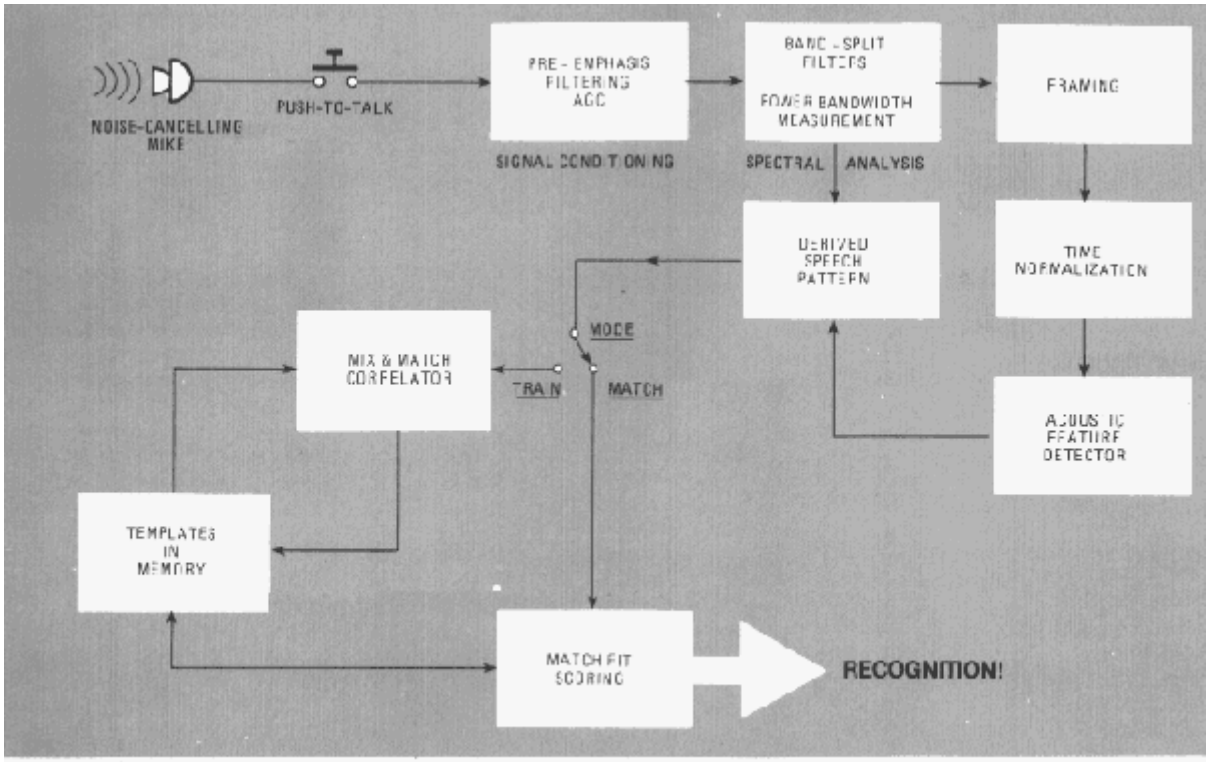


FIG. 1—GENERALIZED block diagram of a speech-recognition system shows how a spoken word is processed so that it can be recognized by a computer.

How it works continued

Since spectrum analysis is time independent and since it can be used to indicate whether or not speech is present the incoming speech is first analyzed for energy content in each voice frequency spectrum sub-band of interest. While the energy content in each sub-band is significant, the amplitude variations of speech overall are generally of no help in analyzing speech; instead, zero-crossings have been found to convey the most significant speech information. Those are counted to give frequency information, although in some methods the interval between zero crossings is counted instead.

In addition to the energy content in each frequency sub-band, some measurement of the rate of change of speech spectrumenergy (rapid for explosive sounds for example, and gradual for vowels) might also be made. Once the end of the word has been detected, the word is framed, defining its beginning and end, and time-related acoustic phenomena are analyzed. An acoustic-feature detector extracts key features, including pauses, vowels and vowel-like sounds, formants, and so on. Then the word is divided into a number of equal parts (Threshold Technology, for example, uses 16 samples that are spaced equally in time) to obtain a time normalized pattern of those key features.

Those patterns are compared with the templates stored in memory; the algorithms that are used for those comparisons are a key difference between the various speech-recognition systems. In all Systems, the input word is compared against the stored vocabulary, and the similarities and differences are weighted into a correlation score. Those scores might be expressed as a product, a vector distance, a probability evaluation, or a figure of merit. The score is a numerical characterization of how good the match is.

Most systems require that the match or "fit" exceed some minimum value in order to be valid. Larger vocabularies, or more critical applications, often require a higher minimum value.

Speaker dependence

Let's consider the problem of recognizing more than one voice. For the speaker-dependent recognition systems available today (speaker-dependent means that the system can only effectively recognize words spoken by the person who trained it), there is an easy answer: trade-off vocabulary for more voices. A system capable of recognizing one speaker and a vocabulary of eighty words could just as well accommodate two speakers, each training it to a list of forty words-or eight speakers and ten words, five speakers and sixteen words, and so on.

Bell Labs has successfully made speech-recognition systems capable of recognizing isolated "utterances" spoken by designated speakers. Those systems use eighth-order LPC (Linear Predictive Coefficient) analysis. You may recognize LPC analysis as the technique used by Texas Instruments to translate speech into much-compressed data, and back again, in their Speak & Spell and elsewhere.

The object of the Bell Labs investigation is an automatic directory assistance system, but they found that the limited vocabulary and speaker dependence of contemporary speech recognizers made the recognition of spoken names impractical, if not impossible.

Limiting the vocabulary to the "names" of the letters used in spelling names makes the task more manageable, but there are still drawbacks. One is that the names of the letters are short compared to most words and so they don't give a recognizer much to go on. There are also many letters whose names sound a great deal like each other.

Bell Labs found an answer. They decided that even if they don't know for certain what a given letter is, it's enough to know that it's one of say five probable candidates. A string of six letters gives enough information that an exact match to a recorded directory listing can be made most of the time, at least under experimental conditions. But what about speaker independence?

Slurring

In the same way that a system maximizes the similarities and minimizes the differences between successive samples of a spoken word during training, samples of the same word spoken by different individuals pmduce an even broader template. In that way, differences between one speaker's articulation of a word and that of another are slurred together. By extension, a system could become speaker-independent if any such thing as a "universal" template (an absolute set of similarities in the ways all people say a word) could be found.

Unfortunately, slurring also blurs the recognition capabilities of a system by making dissimilar words sound more like each other. It may become impossible to discriminate between similar-sounding words. Just as today's speaker-dependent systems are evaluated in terms of their accuracy - a 98% matching rate, for example-future speaker-independent systems may be rated both for overall accuracy and for the percentage of the population that the accuracy figure applies to.

Speaker independence is the first priority in improving coming generations of speech-recognition systems according to most manufacturers we talked to. One promising approach involves producing speaker-adaptive systems that in some way modify stored templates to help adjust them into a closer match with the particular voice characteristics of the speaker. For example, a brief initial sample of the voice might determine if it is that of a man, woman, or child and whether it is basso, alto, tenor, soprano, etc. The spectrum's sub-band energy distribution would obviously shift slightly as the pitch of the speaker's voice shifts, and weighting factors could be introduced into the analysis to help correct for differences between speakers.

Connected speech

We have seen that time analysis of speech for today's isolated word-recognition systems requires proper framing of the word, which means recognizing its beginning and its end. But normal speech is connected speech, with the end of one word often indistinguishable from the beginning of the next.

IBM has been working on the problem of recognizing words and phsases in the midst of continuous speech. Using a large mainframe computer and some advanced techniques including spectrographic analysis, they've been able to take text derived from a 1000-word vocabulary and with a speaker reading the text at a normal pace transcribe the spoken text into printed copy with better than 90% accuracy.

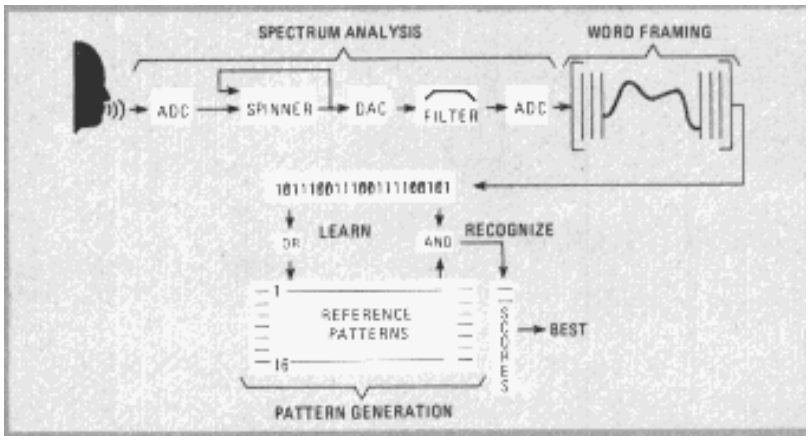
What's here and what's coming

Speech-recognition equipment is available at every level from single boards an experimenter can connect to his computer to huge mainframe systems like those in use at Bell Labs and IBM. A great deal of attention is being given to continued developments in the sophistication and accuracy of voice-entry terminals, which accept spoken rather than keyboard entered data.

Threshold Technology, Incorporated and Centigram Corporation are two of the leaders in speech terminals. A newcomer to that area is one of the pioneers in speech recognition for experimenters, Heuristics Incorporated. New on the experimental end is the Cognivox by Voicetek, and the VET/1 and VET/2 from Scott Instruments

Commercial speech-recognition Systems are also made by Verbex Corporation (formerly Dialog Systems Incorporated), Scope Electronics Incorporated, and Interstate Electronics, as well as Perception Technology Incorporated.

MIKE



How MIKE operates

Mike lears and recognizes patterns derived from spectrum-analysis data. When learning a word, Mike stores patterns in memory for future reference. When attempting to recognize a word, Mike compares the incoming pattern to each reference pattern and generates a set of closeness of fit scores. Above a certain threshold, the highest score is taken to indicate successful recognition. The spectrum analysis is performed every 25 milliseconds to measure the energy in 19 logarithmically spaced frequency bands over the 300-hz to 3.000 hz range. Mike's approach to that analysis is unique. The data to be analyzed is spun past a single filter 16 times., each time a a different frequency, so that

the frequency of interst matches the center frequency of the filter. That is in contrast to the conventional approach, which involves using 16 indivually tuned filters opeating in parallel.

The spectrum-analysis data is digitized and passed to the word-framing process. When a sufficient level of spectral activity is detected, the beginning of a word is marked. When that activity falls below a threshold, the end of the word is marked. Since Mike is an isolated-word recognition device, a silent interval of approximately 100 milliseconds is required at the end of a word to frame it adequately.

Noise-canceling and time-base normalization are integral parts of the word-framing process. During silent intervals, constant (ambient) noise is measured; during word framing, that constant noise signal is subtracted from the input signal. When a word or segment of sound has been isolated, it is normalized to a fixed time-duration to compensate for different speaking rates.

The pattern-generation process further operated on the framed word to extract features on interest and to reduce it to a string of approximately 240 bits. The pattern is then generated using a proprietary mapping algorithm.

In training Mike, patterns are logically or'ed with the patterns of previous repetitions of the word being learned. Typically, two or three repetitions of each vocabulary word suffice for reliable recognition. When Mike is attempting to recognize, patterns are compared by and'ing them in turn with each of the previously learned reference patterns. The matching ones are tallied to form a set of scores for each comparison.

Mike recognizes a word if its score is both above a threshold and greater than the next highest score by a prescribed increment. A code indicating the identity of the recognized patterns is transmitted to a host device. If a word is framed but does not meet the recognition criteria, a no-recognition code is transmitted.

Centigram's recognition approach is patented in the United States (Patent nummer 4,087,630) and patents have been applied for in 15 other countries.

VOICE-RECOGNITION SYSTEMS

For more information, circle the corresponding number on the Free Information card inside the back cover.

Threshold Technology
1829 Underwood Boulevard
Delran, NJ 08075
CIRCLE 91 ON FREE INFORMATION CARD

Centigram Corporation
155A Moffett Drive Park
Suite 106
Sunnyvale, CA 94086
CIRCLE 92 ON FREE INFORMATION CARD

Heuristics Incorporated
1285 Hammerwood Avenue
Sunnyvale, CA 94086
CIRCLE 93 ON FREE INFORMATION CARD

Interstate Electronics
707 E. Vermont Street
Anaheim, CA 92803
CIRCLE 94 ON FREE INFORMATION CARD

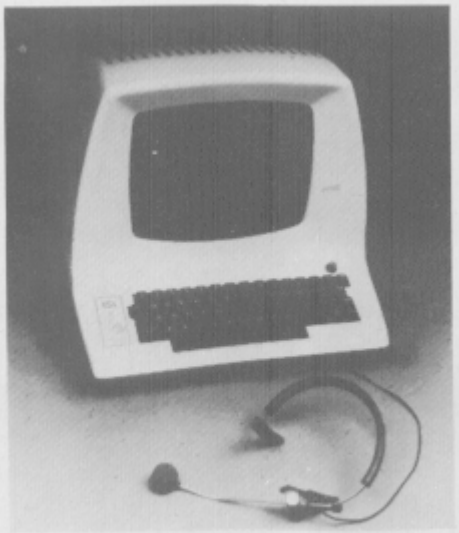
Scope Electronics Incorporated
1860 Michael Faraday Drive
Reston, VA 22090
CIRCLE 95 ON FREE INFORMATION CARD

Verbex Corporation
2 Oak Park
Bedford, MA 01730
CIRCLE 96 ON FREE INFORMATION CARD

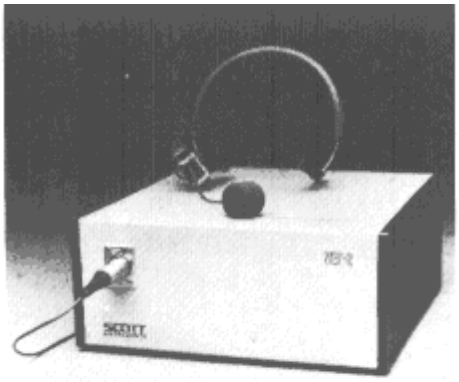
Perception Technology Incorporated
95 Cross Street
Winchester, MA 01890
CIRCLE 97 ON FREE INFORMATION CARD

Scott Instruments
815 North Elm Street
Suite 5
Denton, TX 76201
CIRCLE 98 ON FREE INFORMATION CARD

Voicetek
PO Box 388
Goleta, CA 93017
CIRCLE 99 ON FREE INFORMATION CARD



HEURISTICS *model 5000* speech-recognition system transforms Lear Siegler's *ADM-3A* into a voice-actuated terminal.



VOICE-ENTRY TERMINAL, the *VET-2* from Scott Instruments, is compatible with a *TRS-80 Model 1*.

Radio Electronics JUNE 1981
R-E