# Eyes and Ears for Computers*

E. E. DAVID, JR.†, SENIOR MEMBER, IRE, AND O. G. SELFRIDGE‡

*Summary*—Attempts to mechanize character reading and speech recognition have greatly accelerated in the past decade. This increased interest was prompted by the promise of computer inputs more flexible in format than punched cards or magnetic tape. Research has shown that automatic sensing can be done reliably if the task is suitably delimited. Cleverly designed marks on standard forms can be both machine and man readable. A single type font or a few fixed ones are tractable if the print quality is controlled. Handprinting can be handled for careful writers, as can meticulous handwriting. Isolated spoken words taken from a small number of talkers and a limited vocabulary can be automatically recognized. Typical error rates for these machine-sensings run between 0.5 and 25 per cent. These results imply that reading unrestricted typestyles, handwritten scrawl, or recognizing conversational speech is beyond the reach of present methods.

From the engineering viewpoint, questions of values enter. Might it not be wiser to punch cards or tape while making copy rather than depend upon complex character recognition hardware? Is it useful to have voice input to a computer when a finger and typewriter are available? Answers to such questions will depend upon the specific application. Certainly, the utility of automatic sensing will depend upon what is to be done with the material after it enters the computer as well as the internal organization of the machine itself. Perhaps all would agree that we should have automatic inputs before the Russians, but it is not so clear that we need them soon as practical computer inputs.

AS MAN RUSHES to build his replacements, he notices an interim requirement for man-machine communication. In the meantime at least, computers must be able to, but cannot, understand the writing and talking of men. We are protected from technological unemployment so long as we are buffered by punched cards, magnetic tapes, and on-line or off-line printers. But the day will come!

Current computer inputs are peculiarly inflexible in format, and inconvenient in use. To communicate with the machine by these means we almost invariably must employ an extra man or man-plus-machine, who transcribes our words into shaped holes exactly positioned or magnetic pulses precisely timed.

Men profess to be ill-suited to such menial labor. Though women seem more easily subjugated, they too tend to be expensive, slow, and inaccurate. Therefore, we are increasingly calling on computers to handle those highly variable input signals whose relevant features are not simply related to their physical properties.

In the early days of computers we all learned a congeries of theorems by Turing and von Neumann which told us (or so we thought) that a computer could do anything we told it. We would merely (!) have to specify sufficiently accurately just what it was we wanted the machine to do. And it is true that some highly variable input signals can be categorized by elaborate, exhaustive programs, but it is just not feasible thus to program recognition of printing, speech, handwriting, radar and sonar signals, and objects in photographs (clouds in satellite weather pictures, for example).

Such material can be converted to a standard format, of course, by a suitable scanner, microphone or other transducer whose output is amplitude-quantized samples. We are not much better off, because we then have the same excess of data in two forms, neither related to the classificatory representation we want in the end. Rather, for instance, from speech waveforms we ask a phonetic or literal transcription, from printing and handwriting a literal ("alphanumeric") transcription and not a TV raster, from radar a target inventory, and from weather pictures a cloud map. Such representations imply abstraction of raw data into more meaningful perceptual coordinates. Computers able to convert inputs to such forms will be much more flexibly, powerfully, and economically coupled to the real world.

Though such abstraction is difficult, we already have given some of our machines limited ability to read printing in certain type faces [1], [2]. But reading scratchpad handwriting or transcribing conversational speech by machine is far beyond our ken. Also, it seems clear that, in any case, the computer needs much extra, expensive equipment to handle the patterned inputs we are discussing. Yet as computers get faster, and as we can program more sophisticatedly, input flexibility will become more crucial than it already is.

Computer engineers, in a spasm of interdisciplinary enthusiasm, are rediscovering that perception is hard to understand and even harder to simulate. Armed with the tautological knowledge that computers *can* do what we tell them to do, we tell them to perceive and then ask the psychologists to fill the gaps in our instructions. They don't know how to do it either, but *they've* known it longer. Pattern recognition, which refers to the actual categorization involved, the actual decision that some sound or sight is "mother," is learned by children at an early age. Computers are very young too and must learn pattern recognition; we do not say that pattern recognition must come full-blown from the machine unaided and untaught. Some men, at least, can construct and write pattern recognition programs. That is, men can do the learning and generalizing, and then

* Received by the IRE, August 11, 1961; revised manuscript received, January 8, 1962.

† Bell Telephone Laboratories, Murray Hill, N. J.

‡ Lincoln Laboratory, M.I.T., Lexington, Mass., operated with support from the U. S. Army, Navy and Air Force.

instruct the machine. But it does seem to us that in the long run the machine will have the ability to carry some of this burden itself and will, in fact, need to do so as it tackles subtler problems.

## Methods [3]

Automatic input (that is, without a man) is the classification of input signals (images or sounds) into categories. Basically, the objects or signals are presented one at a time to the input device which makes certain measurements on each. These measurements are inputs to a decision "logic" which assigns the input to one of the available categories [4]; that is, the machine measures and then decides.

What distinguishes the techniques we are discussing here from customary ones is that they concern recognition of complex variable patterns of many elements rather than detection of a few fixed elements. For instance, alphanumeric text from a typewriter comprises some 80 complex shapes. For machine recognition of a particular font, however, templates of these shapes can be prepared and input letters matched against these, reducing the recognition process to one of identity. Of course, the input letters must be (or must be transformed to be) the same size as the templates and must be properly oriented. *Template matching* then, involves samples or pieces of samples of actual or typical or ideal inputs in their original representations. The method gives accurate results for any finite alphabet of shapes each of whose examples are identical within the resolution of the processor.

Unfortunately handwritten or handprinted letters, spoken words and even some printed text typically show wide variations from any set of templates which can be prepared in advance. We might provide a metric to measure the degree of fit of the unknown to every template and also supply transformations to normalize the unknown input to reduce the variability. These measures do not rescue the situation. The greater the variability, the less successful template matching becomes. By human standards, very little variability may reduce performance to unacceptable levels.

We conclude from history that template matching in its pure form, or modified with a metric and normalization, is adequate perhaps only for printing and typewriting. It was evident (after some years) that in many cases people do not work with templates, but prefer to abstract features. Parallel lines, cusps, or totally-enclosed spaces, for example, are features that may aid in separating alphanumerics.

Characterizing shapes by their distinctive features or properties may ease the problem of variability. Such properties tend to be invariant over a far wider range of either hand- or machine-printed letters than can be accommodated by a like number of templates. A list of the properties of known samples can provide a basis for classifying unknown ones. This *property-list matching*, however, does not eliminate the variability problem but transfers it to another level. The "invariant" properties must be recognized regardless of their size and orientation, despite the variability inherent in human motor performance, despite our individual personal habits, and despite the "noise" in reproduction devices. Yet property or feature extraction is a step in the right direction and has proved to be less fallible than a "grand" transformation intended to reduce all inputs to one of a finite set of precise templates.

If the property list is appropriately designed, recognition need not hinge strongly on any one feature, and may incorporate correlated or dependent properties for correcting errors in feature extraction. That is, the property-list technique may work well enough to be useful in situations where template matching would not be useful.

In either case, recognition performance can be improved by using context. Most obviously, letters and sounds make words, but even knowing digram letter or syllable frequencies can improve decisions about letters or phonemes. We may be more sophisticated and use words, perhaps incorporating their syntactical and grammatical restrictions; we might even use semantics. We guess that using more context will mean using much more storage—the number of English words being vastly greater than the number of letters and the number of semantic possibilities overwhelmingly greater.

Again, both template and property-list methods can be supplemented by limited "learning" or "self-organizing" or "adapting." In effect, properties in a list or elements of a template are examined statistically for their relevance to particular categories by noting their influence in determining correct and incorrect classifications. Those leading to correct decisions are reinforced; that is, their influence is increased, and vice versa. The "adapting" can be programmed into the recognition system itself if so desired [5]. Thereby, the machine may vary its parameters to take advantage of "good" features—"good" features being those which make for correct classification. Further the machine may be able to make large-scale adjustment in its parameters adaptively. For example, guessing that a piece of text is in English rather than in French would call up different *a priori* letter digram and word frequencies. The flexibility of computer inputs can be increased in this way, but the method has been explored only scantily.

The ability to cope with new input patterns hinges upon the relevance of the features incorporated in the early stages of the recognizer, or upon the relevance of templates which can be formed from the available elements. Yet to be explored at all is any method, other than random combinations of available elements, for formulating new features to be tested for relevance.

Thus the crucial, often the least mentioned, aspect of recognition is the selection of suitable features. In many studies, particularly involving alphanumerics, experi-

menters have been guided by their intuition. Others have taken the exhaustive (and exhausting) approach. The accomplishments of these methods ride on *ad hoc* inspiration and dogged persistence. Still other researchers have consulted generating mechanisms for speech and handwriting, searching for constraints to confine the range of choices. Thus in speech recognition, studies of vocal tract acoustics and speech synthesis have established the vocal resonances as central features. Also at the contextual level, language and grammar reveal how phonemes and words are contrived to make phrases, clauses, and sentences. At its best, successful *synthesis* reveals a vocabulary of "atomic" features. These then provide the acid to attack verbal or literal compounds. At its worst, synthesis misleads the unwary to accept oversimplified, overspecialized elements which have no counterparts in the league of everyday usage.

Yet to the extent that it reveals the structure of words, letters, or sounds, synthesis is not to be ignored, and it provides a rationale beyond the helter-skelter of randomness. Studies of human perception and sensory mechanisms too, hold a promise of "features-to-come," but the meat is yet to be put on the table even though the aroma whets the appetite. Here also, bona fide knowledge of behavior and physiology can aid in avoiding naïveté.

## Character-Recognition Devices

The goal here is to introduce alphanumeric, text into the computer, including *printed* text, handwriting, and handprinting. In the first the question of segmentation is usually not bothersome, since separate characters are completely isolated [6]. The characters are usually arrayed in horizontal lines, but subscripts and superscripts (as in reference notes and mathematical formulas) must be handled by any technique claiming to be complete. Fig. 1 shows a range of relatively easy to relatively challenging examples.

The first technical choice arises here: what range of fonts shall be admitted? There are many useful cases where just one font does suffice: an instructive example is that of the American Banking Association. Realizing that fonts intelligible to people are hard to make intelligible to computers, they settled on one that was intelligible to computers and only barely intelligible to people. (This example is also interesting because it is not a visual one for the machine. The font is printed in magnetic ink and scanned by magnetic sensors.)

If many fonts are to be recognized, we may suppose that they have been previously specified, and that successful recognition of the characters to some extent requires recognition of the font they come from. This case then must be distinguished from the most general one, when all fonts must be recognized, even those not met before—though perhaps we may exclude those like *The New York Times*, Fig. 1.

The next big choice is the extent to which variation

IN THE BEGINNING WAS

CAPS

In the beginning was

CAPS AND SMALL

THE BELL SYSTEM

UNUSUAL SIZE

ENGINEERS &
SCIENTISTS

UNUSUAL SIZE AND SHAPE

Solid-State Parametric Amplifiers

Parametric excitation was illustrated in the previous section in terms of a periodic variation in the downward force acting on a pendulum bob. Excitation of this kind

* J. Herman and J. Litton, Jr., "Characteristic features of a broad band beam parametric amplifier," *Proc. Symp. on the Application of Low-Noise Receivers to Radar and Allied Equipment*, Lincoln Lab., Mass. Inst. Tech., Lexington, Mass., October 24–28, 1960; vol. 2, pp. 331–346; November, 1960.
** A. Ashkin, Bell Telephone Labs., Inc., Murray Hill, N. J. (private communication).
*** R. Adler, Zenith Radio Corp., Chicago, Ill. (private communication).

MIXTURE OF FONTS WITH ITALICS AND SUPERSCRIPTS


WHITE ON BLACK

*Technical Journal*

UNIQUE TYPE

The New York Times.

?

Fig. 1—Variations of type in common use complicate the automatic reading problem.

can be tolerated. We must expect a certain amount of additive or subtractive (or even multiplicative) noise, which should not (but usually does) seriously degrade performance. The noise is usually not independent from resolution cell to resolution cell, as has often been assumed for analytical purposes. In fact, noise comes in clusters like infections. There may be errors of registration, especially from typewriters, and occasionally errors of rotation. Probably the reasonable standard here is that errors from noise or misregistration should be no more frequent than typographical errors.

Now we reach the choice of techniques, as we described them above: template matching is well adapted to known fonts, especially when one can rely on good registration. Many commercial programs use what is essentially template matching. It cannot be said that these are entirely satisfactory, though some are interesting and ingenious.

Finally, there is a choice of mechanism. Though, of course, the critical element is some photosensitive material, it may be preceded by much or little optical processing, which can, in effect, do a large amount of parallel processing cheaply and fast. For example, in Fig. 2, two photosensitive elements can detect at once the degree of template matching of the image with a known character. Many people, however, have preferred scanning as a desirable compromise: here a vertical row of (typically 5 to 20) photocells sweeps horizontally along the line of type and the matching is done electronically. There are some advantages: misregistration in the optical example of Fig. 2 can scarcely be handled except by actual relative motion of the template and image, but electronically there are several plausible ways, for example, treating the lowest active channel as a base-line. Scanning does not restrict us to template matching but can extend fairly directly to property-list matching. The property list will naturally be biased by the fact of horizontal scanning, but should nevertheless be adequate. For either "matching," the stored patterns can be incorporated as logical circuitry. This acts as a "convoluted slot" through which the input pattern must pass. Further, extension to different fonts need not require physically new templates and one can, at least in principle, extend to completely new fonts.

There have been many character recognition experiments which illustrate these techniques and choices. Results from representative ones are listed in Table I [7]–[9].

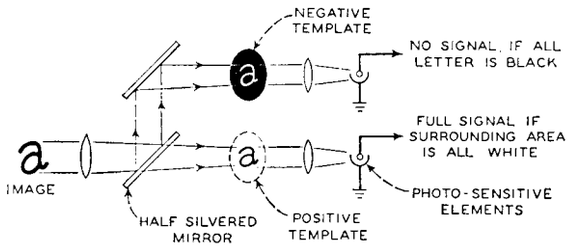


Fig. 2—Template matching is a useful technique where the type face can be specified in advance, but misregistration can reduce its effectiveness.

TABLE I

| Investigator | Input Representation | Vocabulary | Number of Writers Tested | Approx. Error Rate | Additional Facts and Comment |
|---|---|---|---|---|---|
| Baran, Estrin | *X-Y* matrix | Machineprinted numerals | One machine over ribbon life | 9 per cent | Maximize *a posteriori* probability (Bayes' Eqn.) |
| Highleyman | *X-Y* matrix | Machineprinted numerals | One machine, 50 heads | 0.6 per cent | Cross-correlation against probability matrices |
| Bledsoe, Browning | *X-Y* matrix | Handprinted alphanumerics | 1 person | 22 per cent | Matching 2-tuples of matrix elements against table |
| Lewis | 13 features | Machine and handprinted letters | 15 mixed people and machines | 18 per cent | |
| Bomba | Local features | Handprinted alphanumerics | 2 persons | 0 per cent | Decision tree |
| Doyle | Local features | 10 Handprinted letters | ≈20 persons | 12 per cent | Maximize *a posteriori* probability (Bayes' Eqn.) |
| Kamentsky | Presence or absence of intersections of character with scanning lines | Handprinted numerals | ≈20 persons | 1.7 per cent | Correlations of code lists which describe scan interactions |
| Sherman | Nodes describing line endings and junctions | 9 Handprinted letters | ≈20 persons | 5 per cent | Connection matrix describing graph which joins nodes; tested against master matrices |
| Marill, Green | Distance of character from field edge along selected line segments | 3 Handwritten letters | 10 persons | 3 per cent | Likelihood function assuming independent normal distribution of measures |
| Highleyman | *X-Y* matrix | Handprinted numerals | 50 persons | 6 per cent | Linear decision functions |

Notwithstanding these possibilities, commercial ventures now are concentrating on optical and electronic template matching and hope to achieve speeds startlingly small compared with the easily attained 50,000 characters per second from modern tape units. It does not seem, though, that this limitation is inherent in the techniques themselves because photosensitive elements can certainly have better than a microsecond response time. But 50,000 characters per second is on the order of 18,000 words or some 30 pages per second, which is dangerously rapid for paper.

An interesting exception to the previous paragraph is the IBM 1418 [10]. A vertical line of 17 cells moves through 12 positions to blanket a character. The cells give binary responses (setting the thresholds is tricky) which are stepped into a $12 \times 17$ shift register. The character is thus normalized since at some stage it is centered; errors of rotation are not allowed for. The machine uses a large number of empirically derived features as inputs to a binary decision tree. It can recognize some, though not all, of the numerical printing of IBM printers, with a rejection rate as low as one in $10^5$ or $10^6$, but more typically one in $10^3$ or $10^4$. Extension to full alphanumeric coverage should be merely a matter of time and demand. The speed is 480 letters per second or about four Kennedys [7], [11]. (As many as 7 pieces of paper per second can be handled, though not necessarily processed.) The system's speed is thus about 20-db below available tape mechanisms, which can handle a full length novel every 6 seconds or so.

It is proper to ask who or what would want to read very many full length novels at 6 seconds each. The point is that to tie up a large digital computer at much slower speeds is expensive; thus, we conclude that probably for some time the print-reading device will include a buffer and will run essentially off-line.

This conclusion will limit for some time effective application of adaptation and learning. These have a depth of structure difficult to handle without a full general-purpose computer, which, as we have seen, we are reluctant to tie up while pages are slowly turned.

The acquisition of more general devices will, therefore, depend greatly on the need. Leaving aside the question of very high speed, who needs all-font capabilities and is willing to pay for it in computer time or extra equipment? The intelligence services might be expected to need something to read *Izvestia* and *Aviation Week*, but so long as the computer is acting as a mere transcriber or storer, the effective bottleneck is the man who must read the material to understand it. (And we are a long way, indeed, from a computer's understanding any but the *most primitive* English.)

If the reader, working with the electronic scanner, could show and identify samples of every letter for a piece of text, the scanner could then proceed without him. In reading *Izvestia*, at least, this could relieve the man's boredom by occasionally replacing one level of tedium with another, and would not necessitate complicated learning programs. But even this much would be a very large and complex piece of electronic equipment indeed, and even so, it would incorporate little of the sophisticated learning techniques many of us espouse as intellectual goals.

A hopeful sign is the increased emphasis on multiprogramming. If we can find ways of interleaving large programs, we may be able to use the computer to supply the sophistication for character recognition while it is performing unconnected computations most of the time. Were this possible, it would be more feasible to take advantage of context to correct character misrecognition or typographical errors [12].

One branch of character recognition where context is clearly essential is handwriting. The segmentation problem for characters is very difficult and that for words is hard enough. Handprinting has received some fair amount of attention (chiefly because it avoids the segmentation problem) though the applications are obviously limited. Representative results are listed in Table I [13]–[18]. An easy solution here seems tantalizingly out of reach: an IBM exhibit at the Western Joint Computer Conference in Los Angeles, in 1961, had a *small* machine to recognize handprinting of numerals only, but they had to satisfy a set of rules about which the illuminating comment was often heard, "Well, I could live with those rules, but I doubt if most people would."

Pure handwriting recognition is being studied at Bell Telephone Laboratories by Harmon and Frishkopf [19], and nowhere else that we know of. Actually, they treat not just the completed writing, but the dynamic formation of the strokes, and their technique is not easily extendible to previously written material. An interesting program by Eden and Halle [20] of the Research Laboratory of Electronics at Massachusetts Institute of Technology generates handwriting, but the constraints inherent in it have not as yet been applied to handwriting recognition.

## Automatic Speech Input

Reduction of speech to machine-readable form is a problem of long standing. We do not know the origin of the idea and it is probably lost in antiquity. One early instrumentation of this notion was revealed by Paget [21], whose work in the 1920's on the basic nature of speech set the stage for much that followed. His contribution, among other important ones, demonstrated a correspondence between the individual phonetic symbols that represent a spoken language and the acoustic phenomena associated with the human voice. Thereby the voice-operated toy, "Radio Rex," became feasible. It consisted of a celluloid dog with an iron base held within its house by an electromagnet against the force of a spring. Current energizing the magnet flowed through a metal bar which was arranged to form a bridge with two supporting members. This bridge was sensitive to

500-cps acoustic energy which vibrated it, interrupting the current and releasing the dog. The energy around 500 cps contained in the vowel of the word "Rex" was sufficient to trigger the device when the dog's name was called. Modern speech recognizers utilize this same principle, elaborated and differently instrumented to be sure, but based upon similar phonetic-acoustic relations.

Such relations arise from the properties of the human vocal tract. In speaking, different positions of the tongue, lips, and jaw give the vocal tract diverse shapes. The shapes and gestures correspond to the various phonetic symbols. Each shape gives rise to a distinct frequency spectrum and each gesture to a spectral transition [22]. Acoustically, then, speech can be considered as a succession of spectral steady states and transitions. These features are in close correspondence with the phonetic content and we may see them clearly in a sound spectrogram which displays a "running short-time" speech spectrum, Fig. 3 [23]. Note that over most of the time scale, the spectrum contains three or four concentrations of energy, which arise from the natural modes of the vocal tract when it is excited by short pulses of air from the vocal cords. These energy concentrations are known as formants.

At other times, the spectrogram contains no clearly-defined structure, but is noise-like. This characteristic arises when the vocal tract is excited with a random disturbance created by air-flow turbulence at a constriction in the vocal tract. An example is the final sound in the word *this*. Sounds with vocal cord excitation are known as *voiced;* those with turbulent excitation, as *unvoiced;* those with no excitation, as *silent.*

Speech spectrograms may be described quantitatively by dividing the time-frequency plane into rectangles $\Delta f \times \Delta t$ in size, and specifying numerically the energy in each. Intelligible speech can be recovered from such a matrix of numbers, and so in that sense it is a "complete" template representation [24]. Also complete, but more succinct, are abstract descriptions based upon time-dependent parameters, that is, properties such as formant locations, over-all intensity and voiced-unvoiced status. Sufficient resolution for the time-frequency matrix representation requires, *inter alia*, about 1000 degrees of freedom ($\Delta f \times \Delta t$ rectangles) per second of speech, while the parametric version has succeeded with as few as 300 per second. In terms of adequately quantized values, these numbers become 3000 and 1000 binary digits. Such representations achieve a great economy over a speech-time waveform which typically requires 10,000 degrees of freedom and 40,000 bits per second. In storing speech data for recognition, this feature is clearly of great importance, not only for efficient storage, but also because the spectrographic and parametric forms of speech-acoustic data are basic to phonetic interpretation. Most experimental speech recognition systems rest upon these fundamentals.

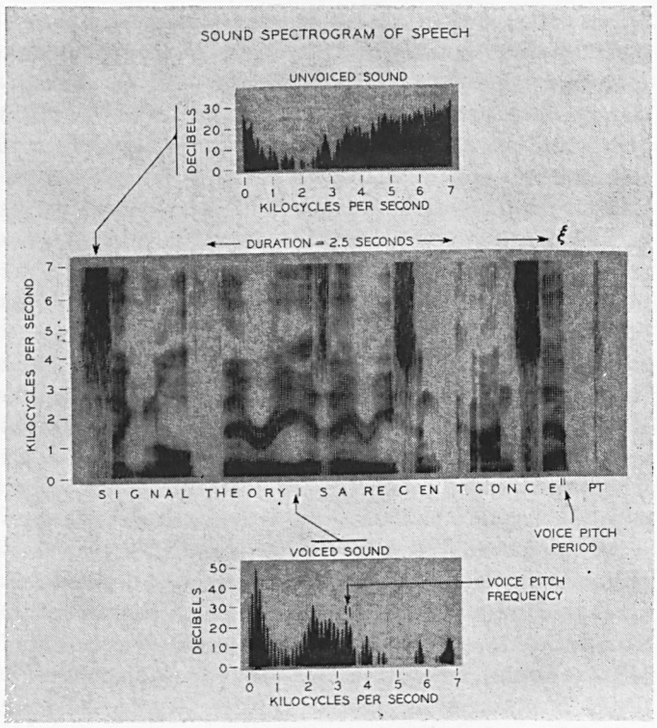


Fig. 3—Sound spectrograms of speech show spectral steady states and transitions which are intimately related to the phonetic content of the utterance.

Typical recognition paradigms use either spectral templates or parametric property lists. Utterances to be identified are analyzed and represented in the same form, and then compared to the stored data. The closest fit, according to some metric (correlation or rms error, for instance) fixes the identity.

Schemes of this kind work less than perfectly. The same phonetic sound spoken by various talkers can give rise to spectrograms whose general features are qualitatively similar, but whose frequency-time matrices differ in detail. This fact is all too clear in Fig. 4. These differences reflect variations in vocal tract dimensions (which determine mean formant frequencies), the range of vocal pitch found among the population, the different gestures of articulation from person to person, the diverse timing and intensity patterns typical of the individual, and so on. Some of these factors are discounted in a parametric representation, but here too, any simple description taken over a number of talkers is far from invariant.

Another difficulty arises from the different dialects in common use. Here the same *words or phrases* spoken by different talkers will have different phonetic content. Thus transliteration from a sequence of phonetic elements to English words may involve complex linguistic structure.

Still another difficulty is that acoustically, speech is continuous in time, whereas its phonetic or literal representation consists of discrete symbols or words. How is a recognizer to reconcile or match a continuous acoustic

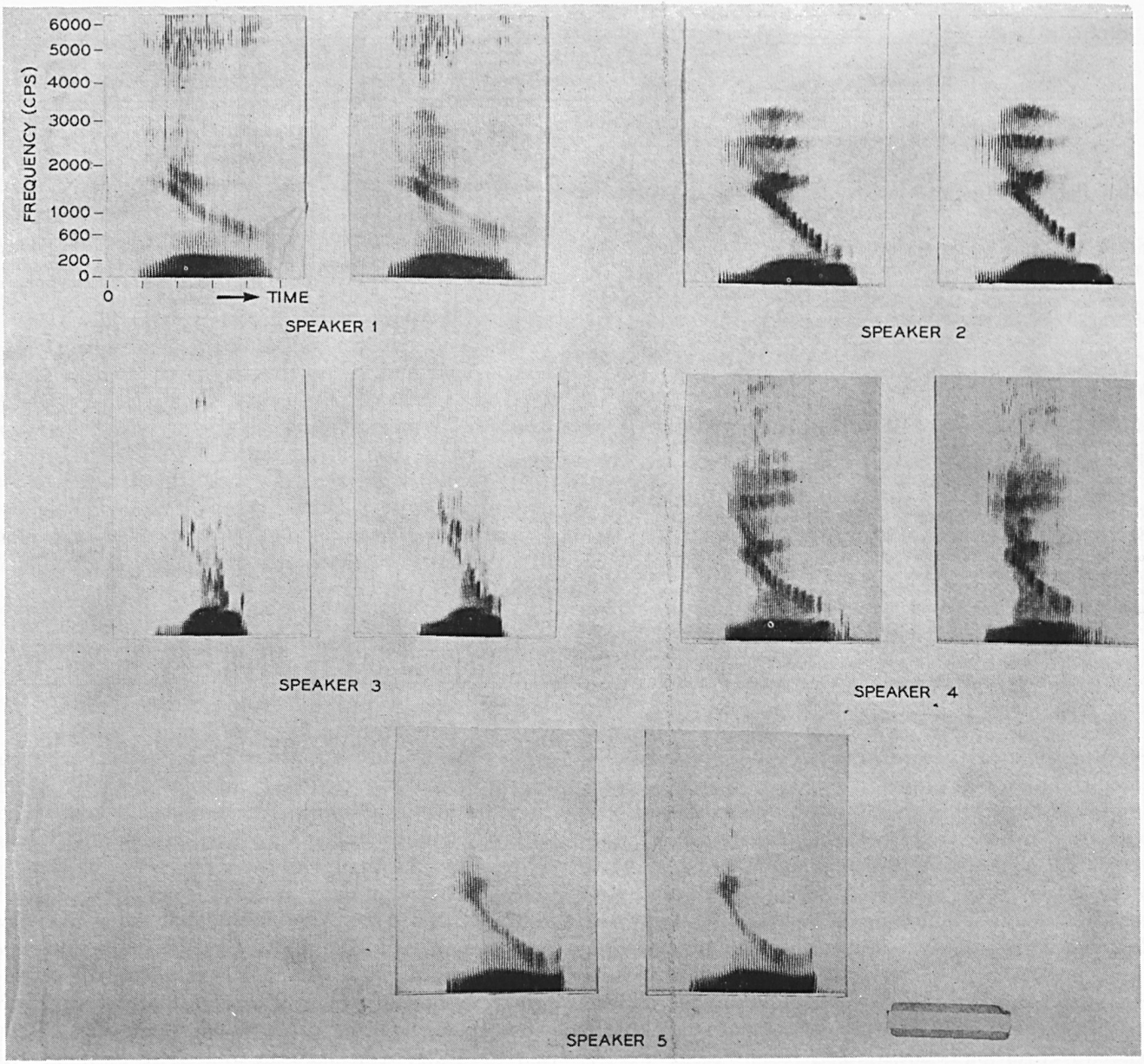


Fig. 4—Sound spectrograms of five talkers each pronouncing the word *you* twice, show that variations in spectral pattern from talker-to-talker are much greater than between successive utterances from the same talker.

flow at its input with the discrete acoustic segments in its memory?

These thorns have been avoided, rather than plucked out, in most speech recognition devices by imposing the following constraints:

1) The stored spectral-temporal patterns and the samples to be recognized are taken from the same talker, or a small number of talkers.
2) The recognition vocabulary is limited to a few words or sounds such as the 10 decimal digits.
3) Each word or sound must be spoken in acoustic isolation.

These crutches aid greatly in retrieving respectable performance from a difficult situation. Typical recognition results are listed in Table II [25]–[37]. Examination of this table shows that the rate of correct identification can approach 100 per cent when the input is protected from the rough and tumble of everyday talk.

As far as we know, no one has yet used such a scheme for practical computer input. Yet there seems no technical reason why speech recognition could not serve under constrained conditions like those where character readers have proved useful [38]. One obstacle lies in the amount and complexity of equipment required. Spectrographic analysis calls for an extensive bank of contiguous band-pass filters or their equivalents [39]. Their rectified and smoothed outputs may require further processing, normalization with respect to intensity and time for example, before being routed to the comparator.

TABLE II

| Investigator | Speech Representation | Vocabulary | Number of Talkers Tested | Approx. Error Rate | Additional Facts and Comment |
|---|---|---|---|---|---|
| Kersta | Selected entries from $\Delta f$-$\Delta t$ matrix (200 cps $\times$67 ms) | 10 digits | 9 men, 5 women | 0.2 per cent | Spectrograms quantized into 2 levels |
| Davis, Biddulph, and Balashek | Formants 1 and 2 as a function of time | 10 digits | 1 talker | 2.0 per cent | Correlation metric |
| Fry and Denes | Selected entries from $\Delta f$-$\Delta t$ matrix | 14 speech sounds in 139 words | 1 talker | 28.0 per cent (Sounds)<br>56.0 per cent (Words) | Phoneme digram frequencies used to supplement primitive recognition from acoustic data |
| Olson and Belar | $\Delta f$-$\Delta t$ matrix | 10 words or syllables | 1 talker | 2.0 per cent | |
| Dudley and Balashek | $\Delta f$-$\Delta t$ matrix | 10 digits | 2 men | 5 per cent | Temporal sequence disregarded |
| Mathews and Denes | $\Delta f$-$\Delta t$ matrix | 10 digits | 6 men | 6 per cent | Spectral pattern time and amplitude normalized |
| Hughes | Spectral features | 11 sound categories in 100 words | 4 men, 3 women | 30 per cent | Feature selection based on linguistic analysis |
| Shultz | Spectral features | 10 digits | 25 men, 25 women | 3 per cent | |
| Petrick and Willett | $\Delta f$-$\Delta t$ matrix | 10 digits | 1 talker | $<$1.0 per cent | Spectral patterns time normalized |
| Forgie and Forgie | Spectral features | 10 vowels | 11 men, 10 women | 7 per cent | |
| Keith-Smith and Klem | $\Delta f$-$\Delta t$ matrix | 10 vowels | 11 men, 10 women | 6 per cent | Statistical decision procedure used to select relevant spectral features |
| Sebestyen | $\Delta f$-$\Delta t$ matrix | 10 digits | 10 speakers | $<$1 per cent | |
| Suzuki and Nakata | Formants 1 and 2 | 5 vowels in consonant contexts | 5 speakers | $\approx$20 per cent | Additional experiments on vowels in bisyllable words and short sentences yield higher error rates |

For a parametric representation, this processing may include complex logic aimed at selecting significant spectral peaks, separating voiced from unvoiced speech and both from silence, etc. Time multiplexing may be convenient at some point in this chain to avoid multiple processors and comparators. Similar multiplexing must be adapted to read-out from the storage. If the stored data are to be adapted to different talkers or groups, then the equipment must operate in a "store" or "learning" mode in addition to the recognition mode. Suffice it to say, the engineering design problem for practical speech recognition is not simple.

Another unavoidable thorn is reliability. How is the performance implied in Table II affected by degradation of the input speech? One should not forget in practice that there are extraneous noises, like airplanes or other talkers or typewriters, and that people talk loudly or softly, holding the microphone close or far away, hoarsely or musically, and so on. We have barely started evaluating such difficulties; neither have we paid much attention to the permissible tolerances.

All these difficulties may well be eased by further research on speech and perception. More versatile speech recognizers may distinguish not only particular spectral features but particular dynamic transitions as well. Linguistic information concerning speaker dialects, intonation, stress, and timing may all aid in reconciling acoustic patterns with their literal counterparts. Surely the complexity of recognition paradigms will not decrease.

We should reiterate that mere extensions of the techniques in the works cited are almost certainly inadequate for handling conversational speech, which we believe to be much harder than handling isolated words.

We cannot easily predict the utility of speech recognition as a computer input. At present the amount of necessary hardware seems forbidding and the competition is simplicity itself, a keyboard and a finger, but the need for speed, for rapid feedback, for technological bravado surely dictate the availability, at least, of a speech typewriter in the next ten years. It will probably turn out that its utility will hinge on what is to be done with human utterances after they enter the machine.

## Conclusion

We believe that progress towards useful and effective eyes and ears for computers has been slow, possibly slower than necessary (especially for print readers); partly because we seriously underestimated the difficulties; and there is an inherent question of values. What is the utility of a speech typewriter? Telephones now have dials instead of operators, and one big customer has vanished, alas! Why can't printing devices produce machine-readable tapes while they are printing? Present-day character readers provide a ponderous way of reading carefully-controlled print. All agree we should get these exotic sensors before the Russians, but we are not sure that we obviously need them soon as practical, useful, economical computer inputs. For most of us, contact with the computer is, as it were, by 24-hour mail;

telegraphy or teletype would be a bountiful improvement and telephones altogether too confusingly fast.

But in another vein, as these techniques appear, they will uncover their own applications just as computers themselves have. The commercial interest nowadays is largely a competitive one—successive advertising disclosures arouse successive new bursts of research activity. At some critical point the techniques will draw support from performance rather than promise, though we can at present only ill-describe such future performance.

Costs will be large. But remember that one reading input device can (potentially) replace hundreds of people reading onto keypunches. A real-time speech typewriter can replace 5 people if it works 24 hours a day; it will be a large piece of equipment that can be amortized on $20,000 a year.

The real difficulty then will be the input-input problem: given words thundering in, what do we do besides store them? The state-of-the-art of automatic indexing and abstracting is even more inadequate and as we said before, it will be a long time before we handle English in all its full meaningful glory. But we are a step ahead. The problems we discuss today are being solved. Here at least we can feel fairly sure that we have identified one of tomorrow's problems.

## References


[1] References in this paper will attempt to be representative, rather than exhaustive. The interested reader is referred to M. L. Minsky, "A selected descriptor-indexed bibliography to the literature on artificial intelligence," IRE Trans. on Human Factors, vol. HFE-2, pp. 39–55, March, 1961, for a comprehensive listing.

[2] D. H. Shepard, P. F. Bargh, and C. C. Heasley, Jr., "A reliable character sensing system for documents prepared on conventional business devices," 1957 IRE WESCON Convention Record, pt 4, pp. 111–120.

[3] M. L. Minsky, "Steps toward artificial intelligence," Proc. IRE, vol. 49, pp. 8–30; January, 1961.

[4] The term "logic" refers to the traditional techniques that are closely akin to truth-table-like manipulation of binary signals. Today the more meaningless term "process" is appropriate.

[5] O. G. Selfridge, "Pandemonium: a paradigm for learning," *Proc. Symp. on Mechanization of Thought Processes*, H.M. Stationery Office, London, England; 1959.

[6] Note that a typewriter leaves a blank rectangular annulus around each character, but that typically other printing techniques do not, even combining several characters into one character; *e.g.*, fi, ffl, &c.

[7] P. Baran and G. Estrin, "An adaptive character reader," 1960 IRE WESCON Convention Record, pt. 4, pp. 29–41.

[8] W. H. Highleyman, "An analog method for character recognition," IRE Trans. on Electronic Computers, vol. EC-10, pp. 502–512; September, 1961.

[9] P. M. Lewis, "The characteristic selection problem in recognition systems," IRE Trans. on Information Theory, Special Issue on Sensory Information Processing, vol. IT-8, pp. 171–178; February, 1962.

[10] J. J. Leimer, "Design factors in the development of an optical character recognition machine," IRE Trans. on Information Theory, Special Issue on Sensory Information Processing, vol. IT-8, pp. 167–171; February, 1962.

[11] Our President is reputed to set the reading pace of Washington, D. C., with a reading speed of 1200 words per minute.

[12] See, *e.g.*, G. A. Miller "Language and Communication," McGraw-Hill Book Co., Inc., New York, N. Y.; 1951.

[13] W. W. Bledsoe and I. Browning, "Pattern recognition and reading by machine," *Proc. Eastern Joint Computer Conf.*, Boston, Mass., December 1–3, 1959, pp. 225–232.

[14] J. S. Bomba, "Alpha-numeric character recognition using local operations," *Proc. Eastern Joint Computer Conf.*, Boston, Mass., December 1–3, 1959, pp. 218–224.

[15] W. Doyle, "Recognition of sloppy, hand-printed characters," *Proc. Western Joint Computer Conf.*, San Francisco, Calif., May 3–5, 1960, pp. 133–142.

[16] L. A. Kamentsky, "The simulation of three machines which read rows of handwritten Arabic numbers," IRE Trans. on Electronic Computers, vol. EC-10, pp. 489–512; September, 1961.

[17] H. Sherman, "A quasi-topological method for the recognition of line patterns," *Proc. Internatl. Conf. on Information Processing*, Paris, June 15–20, 1959, pp. 232–238; UNESCO, 1959.

[18] T. Marill and D. M. Green, "Statistical recognition functions and the design of pattern recognizer," IRE Trans. on Electronic Computers, vol. EC-9, pp. 472–477; December, 1960.

[19] L. S. Frishkopf and L. D. Harmon, "Machine reading of cursive script," *Proc. 4th London Symp. on Information Theory*, Butterworths, London, August 29–September 2, 1960, pp. 300–316; 1961.

[20] M. Eden and M. Halle, "The characterization of cursive writing," *Proc. 4th London Symp. on Information Theory*, Butterworths, London, August 29–September 2, 1960, pp. 287–299; 1961.

[21] Sir Richard Paget, "Human Speech," Harcourt Brace and Co., pp. 79–80; 1930.

[22] C. G. M. Fant, "Acoustic Theory of Speech Production," Mouton and Co., The Hague, The Netherlands; 1960.

[23] This spectrogram was made by a 300-cps bandwidth scanning spectrum analyzer whose output was recorded on electrically sensitive paper; see E. E. David, Jr., "Signal theory in speech transmission," IRE Trans. on Circuit Theory, vol. CT-3, pp. 232–244; December, 1956.

[24] E. E. David, Jr., "Naturalness and distortion in speech processing devices," *J. Acoust. Soc. Am.*, vol. 28, pp. 586–589; July, 1956.

[25] E. E. David, Jr., "Artificial auditory recognition in telephony," *IBM J. Res. & Dev.*, vol. 2, pp. 294–309; October, 1959.

[26] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, pp. 637–642; November, 1952.

[27] P. B. Denes, "The design and operation of the mechanical speech recognizer at University College, London," *J. Brit. IRE*, vol. 19, pp. 219–229; April, 1959.

[28] H. F. Olson and H. Belar, "Phonetic typewriter," *J. Acous. Soc. Am.*, vol. 28, pp. 1072–1081; November, 1956.

[29] H. W. Dudley and S. Balashek, "Automatic recognition of phonetic patterns of speech," *J. Acoust. Soc. Am.*, vol. 30, pp. 721–732; August, 1958.

[30] P. B. Denes and M. V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Am.*, vol. 32, pp. 1450–1455; November, 1960.

[31] G. W. Hughes, "On the recognition of speech by machine," Sc.D. dissertation, Dept. of Elec. Engrg., MIT, Cambridge, Mass., 1959.

[32] G. L. Shultz, "Investigation procedures for speech recognition," *Proc. Seminar on Speech Compression and Processing*, AF Cambridge Res. Ctr., Bedford, Mass., September 28–30, 1959, Tech. Rept. No. 198; 1959.

[33] S. R. Petrick and H. M. Willett, "A method of voice communication with a digital computer," *Proc. Eastern Joint Computer Conf.*, New York, N. Y., December 13–15, 1960, pp. 11–24.

[34] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Am.*, vol. 31, pp. 1480–1489; November, 1959.

[35] J. E. Keith-Smith and L. Klem, "Vowel recognition using a multiple discriminant function," *J. Acoust. Soc. Am.*, vol. 33, p. 358; March, 1961.

[36] G. S. Sebestyen, "Recognition of membership in classes," IRE Trans. on Information Theory, vol. IT-7, pp. 44–50; January, 1961.

[37] J. Suzuki and K. Nakata, "Recognition of Japanese Vowels," *J. Rad. Res. Lab.*, vol. 8, pp. 193–212; May, 1961.

[38] Probably the real reason is economic: the marginal utility of a speech input adapted to just one talker, when he does not suffer from grippe or laryngitis, is just too small.

[39] There are only three computer installations that have gone as far as a real-time microphone-with-filter-bank input, we believe.