J. H. King, Jr.
C. J. Tunis

Some Experiments in Spoken Word Recognition

Abstract: This paper describes some experimental work in the recognition of limited-size, but arbitrary, vocabularies of spoken words. The equipment consists of a filter-bank voice-spectrum analyzer providing real-time input of measurement data to an IBM 1620-II digital computer system. The computer implements various transformations on the input data and also implements various linear decision functions which are designed by means of adaptive algorithms. Recognition experiments have investigated the recognition capability of this system on arbitrary vocabularies of up to 30 words. Several normalizing transformations on the primary measurements were investigated.

Introduction

This work is concerned with machine recognition of spoken words. An experimental system capable of recognizing an arbitrary vocabulary of spoken words is described. The equipment consists of a voice spectrum analyzer acting as on-line input to an IBM 1620-II digital computer system. The basis of the spectrum analyzer is a contiguously tuned bank of bandpass filters whose instantaneous outputs are continuously compared in such a way as to locate the instantaneous peaks in the envelope of the speech spectrum. The output of the spectrum analyzer is a binary coded representation of the peaks of the envelope of the frequency spectrum as a function of time; it serves as input to the IBM 1620.

A programming system that allows application of various transformations to this input measurement space has been written for the digital machine. This transformed measurement space then is used as the input to the categorizer section of the recognition system. The categorizer section consists of linear decision functions in which the weights are obtained using an adaptive algorithm.

Filter bank analyzers have been much used (e.g., by Abramson, et al.,[1] Denes and Mathews,[2] Olson and Belar,[3] Davis, et al.,[4] and Talbert, et al.[5]) and so the primary measurement space used here is similar to that employed in those other works. Talbert, et al.[5] and Dammann[6] have reported on the use of adaptive linear decision functions in speech recognition studies, but the decision algorithm used here is different. Dersch[7] has used a distinctly different measurement space to accomplish the recognition of a vocabulary that is essentially the same as one used here. Rosenblatt,[8] who is responsible for much of the early work in adaptive networks, has also proposed a speech recognition machine using a multi-layer adaptive system. A good over-all survey of past and current work in the field of speech recognition has recently been published by Lindgren.[9]

This work extends the results existing in the literature in that it deals with significantly larger sample sizes than have commonly been used, with a limited number of different vocabularies, and with the effect of transformations of the primary measurement space on recognition performance.

In the first part of this paper the major functions necessary in any pattern recognition system are noted, aspects of decision theory as pertinent to the present work are summarized, and a theoretical model for speech synthesis is described. These considerations have provided the basis for the experimental approach taken in this work. It is shown that, specifically, a means is required for deriving the time variation of the speech waveform frequency spectrum envelope.

The second part of the paper describes two parts of the experimental system—the spectrum analyzer and the linear decision function—in some detail. Most of the circuitry of the spectrum analyzer is conventional, although some special circuits were designed for critical applications.

In the third part a rationale for a set of recognition experiments is developed and the experimental results using this apparatus are presented. Finally, the results are assessed with the view toward determining the direction of future work.

Theoretical basis for experimental system

• *Pattern recognition systems*

The automatic recognition of spoken words is here considered as one member of the broad class of pattern recognition problems, which have been much discussed in the technical literature.[10] There are three main sections in any pattern recognition system:

(1) The Measurement Section performs measurements on the primary input signal. These measurements may be preserved in analog or digital form.
(2) The Transformation Section manipulates the original measurements and converts them into different representations (or measurement spaces) that are more suitable for the particular decision function to be used.
(3) The Decision Section implements the decision functions used to classify the transformed measurements into the classes of input signals.

There is, of course, significant interaction between the design of these three sections; e.g., the choice of a particular measurement space may require the use of an extremely complex Decision Section. In addition, there are many purely hardware considerations that dictate the particular choice of measurement, transformation, or decision function to be used; certain measurement spaces may require much more digital storage than others. Considerations that lead to the choice of a particular initial measurement space for the recognition of speech sounds will be briefly discussed here. In later parts of the paper, the interaction between the three sections of a pattern recognition system will be further discussed.

• *Decision theory*

The decision problem involved here can be formulated as follows: Given a set of measurements, X, it is required to decide which word, out of a finite set of possible words, was uttered. The theoretical solution to this problem is contained in the statistical decision theory as formulated by Wald.[11] A decision in favor of a particular class of pattern (or word) $\mathfrak{S}_j$, is based on a set of weighted comparisons of the *a posteriori* conditional probabilities,

$$P(\mathfrak{s} \in \mathfrak{S}_j \mid X), \tag{1}$$

where $\mathfrak{s}$ is the unknown pattern (or speech utterance) received by the recognition machine and j is an integer ranging over the interval $1, 2, \cdots, p$ (where p is the number of classes of pattern to be recognized, or, in this case, the number of words in the vocabulary). It is well known that Bayes' theorem allows the computation of $P(\mathfrak{s} \in \mathfrak{S}_j|X)$ to be replaced by the computation of the conditional probability densities of $P(X|\mathfrak{s} \in \mathfrak{S}_j)$.

One approach of interest assumes the existence of a perfect measurement X_j for each class of utterance. It also assumes that the variety of measurements is the result of the perfect measurements being corrupted by additive noise. Abramson et al.[1] view this noise as resulting in a statistical spreading of measurements, X, about the set X_j. For the special case where the noise may be viewed as a multivariate, normal, random variable with zero mean and equal variance on each component, the optimum decision strategy, i.e., the computation of the conditional probabilities $P(X|\mathfrak{s} \in \mathfrak{S}_j)$, reduces to a comparison of a set of linear functionals, one functional associated with each class of utterance. There are many factors which conspire to preclude the possibility of there being a perfect measurement, X_j, if the transformation $X = v(t)$ is used, where $v(t)$ is the microphone voltage waveform, the most obvious being the variation in the speed of talking. A more promising "measurement space" for speech recognition is suggested by a consideration of a theoretical model for the generation of synthetic speech sounds.

• *A theoretical model for speech synthesis*

A simple model for human speech was originally formulated by H. W. Dudley.[12] It relies on the observation that during a human speech utterance (especially during the "voiced" portions), the acoustic energy is mainly concentrated in only a few relatively narrow regions of the frequency spectrum. It has also been observed that the locations of these energy concentrations occur in particular ways that are characteristic of the limited repertoire of vowel and consonant sounds that the human is capable of producing. Usually there are three distinct energy concentrations in the frequency range from 300 to 3,000 cps. In the "parlance of the trade," these energy concentrations are known as *formants*. During the utterance of a word the position and relation of the formants change, creating a characteristic "pattern" distinctive of the word and, to some extent, the speaker. The noteworthy success of a number of speech synthesizers (as determined by human recognizability of the produced sounds) attest to its validity.[13,14] A version of Dudley's speech synthesis model is described below and in Fig. 1.

A source of controls generates the signals U, Y, Z (functions of time).* For the synthesis of a word, a particular set of control signals U_i, Y_i, Z_i is generated. The U signal controls the glottal excitation source, $G(s, U)$, which is a time variable source of broadband energy (here $s = \alpha + j\omega$). The control signal, Y, similarly controls the "hiss" excitation source, $N(s, Y)$. The outputs of these

* Some liberties with mathematical rigor are taken here to simplify the description of a waveform that varies slowly with time. It will be observed that the error is not very great if U, Y, and Z are considered as essentially constant during several periods of the lowest frequency components of G and N.

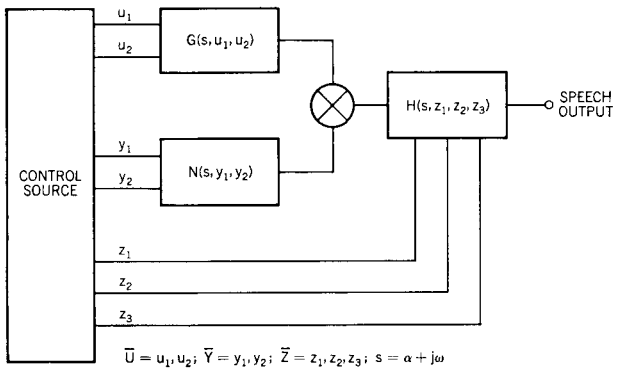


Figure 1 Model for synthetic speech production.

two sources are fed to the input of a time variable filter $H(s, Z)$ which is an analog of the vocal tract. The filter $H(s, Z)$ is operated on by the Z control signals.

It is the relatively slow variation of the pole locations of $H(s, Z)$ which, for the most part, determines the location in the frequency domain for the formant structure of the synthetically spoken word. The U and Y control signals determine the duration, intensity, and to some extent the formant structure of the voiced and unvoiced portions of the word, respectively.

The success of speech synthesizers based on Dudley's model has adequately demonstrated that it is the frequency spectrum (specifically, the energy concentrations in the frequency domain, or "formants") of the speech waveform $v(t)$, that is the information carrier.

- *A suitable measurement method*

If a continuous spectrum analysis of $v(t)$ with a frequency resolution sufficiently fine ($\sim$200 cps) to resolve the formants is performed, but not so fine as to resolve the discrete harmonics of $G(s, U)$, the result will be a representation of a sound that is dependent mainly on U, Y, Z. This representation should be very consistent for the same word because the U, Y, and Z functions are unique for each word. Such an analysis is done by the "sonograph," which is a type of spectrum analyzer manufactured by the Kay Electronics Co., Pine Brook, N. J. (The output of the sonograph is a continuous record of frequency and amplitude vs time; this record is called a sonogram.)

As a result of the spectrum analysis, a measurement $X(t)$ will be obtained. Here, $X(t) = x_1(t), x_2(t), \cdots, x_n(t)$, each $x(t)$ being the output of a bandpass filter. $X(t)$ may be considered an approximate representation of the output of a sonograph. More realistically, noise will be considered to be mixed with the transmitted signal $v(t)$. The result is that there will be some variation in the measurement X even for the same word. However, X should be a statistically invariant measure, unique for each word class in the vocabulary of the synthesizer.

If the various components of X are sampled and binary quantized at successive intervals during a speech utterance, each variation of X may then be viewed as a point in a multi-dimensional measurement space, M. It will be found, that due to the effects of noise, there is a distribution of the X's on M associated with each word class. Thus, the design of an optimum recognizer for this situation is again best considered from the point of view of statistical decision theory.

Although it is reasonable to assume there is a "perfect" signal (or measurement) for each word uttered by an ensemble of artificial synthesizers, the set of perfect measurements may not necessarily be known. For the case of additive noise of the character postulated above, there exist a number of effective algorithms for finding suitable decision boundaries, without the knowledge of the perfect X_i (this is discussed later). Some are variants of iterative routines which operate on a sequence of measurements, $X_1, X_2, \cdots, X_n$ where the class associations are known *a priori* and the number of measurements from each class is large enough to be representative. After a suitable number of iterations, the process either converges, or else the performance of the system ceases to improve (on the basis of the representative sample); e.g., the reject and substitution rates remain essentially constant for any further iterations.[15–17] Nonconvergence indicates either overlap of the distributions associated with certain or all pairs of classes or else linear inseparability.

If recognition of artificial speech utterances were the goal, the techniques outlined above would probably be quite suitable. However, there is one important difference between artificial and human-produced speech: there is a considerable variation in the speed of talking and the nature of sounds (expression, accent) produced by individual speakers. If an operation on the measurements, $X' = g(X)$, could be discovered that would eliminate the effects of variation in speed of talking, volume, and expression between speakers, the problem of recognition of human speech might still be approached from the perfect signal point of view.[2]

A similar problem (the lack of a "perfect" signal source) is encountered in the recognition of handwritten or printed characters. The problem is not present in the recognition of single-font typewritten material where the typebar itself is the "perfect" signal source.

The intent of this work has been to perform experiments investigating the utility of the measurements $X(t)$ for the recognition of human speech utterances. The foregoing analysis has already shown its relevance to the recognition of artificial speech.

A further aim is to investigate methods of designing the Transformation and Decision Sections of the recog-

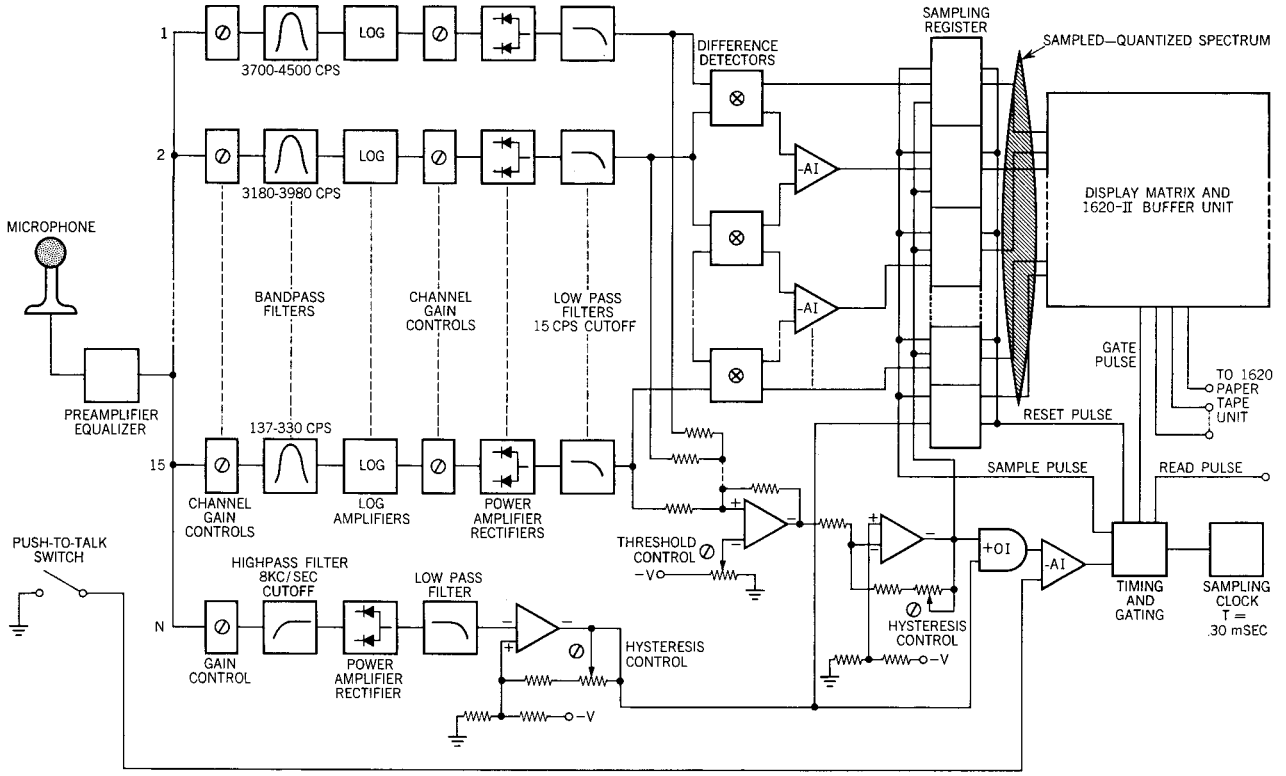


Figure 2 Schematic diagram of measurement apparatus for the experimental system.

nition system to tolerate or eliminate the variations in the measurements, X, when they are made on human speech waveforms. The following Section describes the method and circuitry used for extraction of the measurements from the speech waveforms, and the selection of transformation and decision sections of theoretical merit.

Design of experimental system

- *The measurement apparatus*

The method chosen for the analysis of the speech waveform is based on techniques similar to those used by others for this same purpose.[2–4] Functionally, the first part of the system (Fig. 2) consists of a microphone (transducer for producing an electrical analog of the acoustic pressure waveform) and a preamplifier-equalizer. The preamplifier-equalizer has an amplitude vs frequency response that is the inverse of the average variation in amplitude vs frequency of normal speech. The second part consists of a bank of contiguously tuned bandpass filters (the outputs of which are envelope detected) which perform the function of gross spectrum analysis on the speech waveform, $v(t)$. The frequency increment between filters and the bandwidth of each are adjusted so that the nominally closest spacing of two formants can be just resolved and so that the sensitivity to discrete harmonics is minimized. Since the main interest is in the location in the frequency spectrum of each formant, and not in its absolute intensity, a considerable simplification is allowed in the hardware. Again the footsteps of others are followed here. As shown in the schematic of Fig. 2 the instantaneous magnitudes of the envelope from each bandpass filter are compared by means of a set of difference detectors. By the suitable ANDing together of the bi-polar outputs of the difference detectors, a representation of the instantaneous local spectrum maxima is obtained.[3] These maxima presumably correspond to the formants. If the output of the set of AND gates is periodically sampled and stored during a speech utterance, the nominal result is a quantized record of the formant structure of a spoken word.

Specifically, a binary array, or matrix, is formed, in which the rows correspond to discrete contiguous frequencies, and the columns, scanning from left to right, correspond to successive intervals of time. If the convention is adopted that a ONE bit corresponds to a local instantaneous maximum in the spectrum, then the analyzer should produce one-bit arrangements (resulting from a speech utterance) that correspond directly to the spectral maxima as displayed on a sonogram.

The envelope detectors are actually full wave rectifiers. Low pass filters, which follow, serve the function of removing the beat note (about 100 cps) between the discrete harmonics of the speech spectrum; they also improve the signal-to-noise ratio at this point in the system by removing some of the effects of unsteadiness in the voice.

This analyzer differs from some others[2-4] in that there is no AGC (automatic gain control) in the microphone preamplifier-equalizer. Most simple AGC systems for speech suffer from the deleterious effects of loop delay (overshoot and distortion). Of the two, distortion would probably be the most harmful in that it could alter the gross envelope of the spectrum. A somewhat unorthodox method of compressing the output waveform of each bandpass filter circumvents this difficulty. This is accomplished by means of an active network which has a transfer function approximately defined by:

$$e_{out} = \begin{cases} |\log (1 + e_{in})| & \text{for} \quad e_{in} > 0 \\ -|\log (1 - e_{in})| & \text{for} \quad e_{in} \leq 0. \end{cases} \tag{2}$$

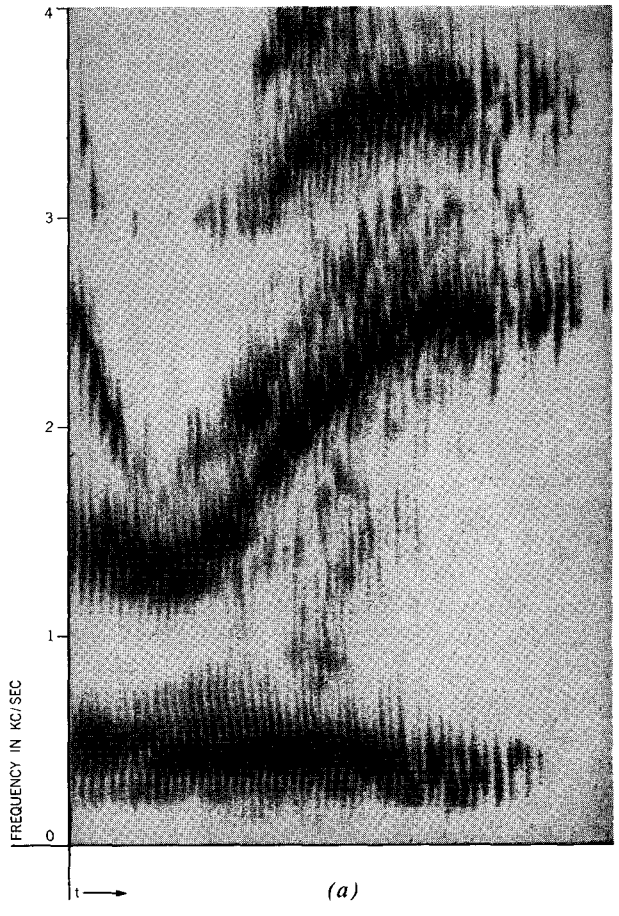
Admittedly, this operation also produces distortion, but this occurs after the frequency-selective filtering. The result is that, for a pure sine wave signal into the system, the difference voltages supplied to the difference detectors are closely proportional to the ratio of the outputs of the respective filters over about a 30 dB range, thus making the sensitivity of the system less dependent on the absolute input level. For a speech waveform input the result is similar, but now the distortion of the waveforms due to the discrete harmonics degrades the performance of the envelope detectors somewhat.

It is well known that during the noise-excited portions of speech most of the energy is concentrated above the frequency range of the three principle formants, although the formant resonances are nevertheless still excited. During voiced portions most of the energy is concentrated within the range of the three principle formants, i.e., below 4 kc/sec. This immediately suggests that a simple highpass filtering technique might be adequate to distinguish (detect) noise excitation. This has been the standard approach and is the one that is employed here. Thus, by comparing levels of energy in the highpass range with the average level in the formant range, the system is able to distinguish, with some degree of reliability, between noise-excited and glottal-excited (or both) portions of spoken words.

- *Choice of decision function*

The binary quantized representation of the spoken word produced by the measurement apparatus is used as the input data to the categorizer portion of the recognition system studied here. This measurement space has some

Figure 3 Typical sonogram and display matrix.



correspondence with the sonogram of a spoken word; Fig. 3 shows an example of the degree of correspondence that is achieved. Note that a "light on" condition in the display matrix corresponds to a relatively darker region of the sonogram.

Changes in the speed of talking have the effect of changing the relative time scale of the display matrix. A given word spoken rapidly will produce a foreshortened pattern, whereas the same word uttered slowly and deliberately will produce a horizontally elongated pattern (more samples) but one which would have essentially the same topological features.

From this observation it might be concluded that if all words were simply normalized to the same length (number of samples) by a uniform stretching operation, this "timing" problem might be eliminated.[2] In theory this should work quite well, since the problem would be reduced to one closely analogous to that of the recognition of a set of "perfect" signals combined with additive noise. The one disadvantage here is the considerable amount of data processing involved in each normalization. It should be noted that this normalization (transformation) would not be effective in reducing other disturbances due to speaker differences, particularly those due to anatomical differences in the vocal tract cavities. Nevertheless, normalization of this sort can be thought of as a method for producing a reference standard measurement space if one were dealing with only single-speaker word recognition.

There are other methods currently in vogue in the field of character recognition for eliminating the effects of such distortions as stretching, skewing, magnification, etc. Many of these may be referred to as "feature detection" methods, although they variously go by the names: *n*-tuple detection, zoned *n*-tuples, stroke detection, lakes and bays, etc. Feature detection performs a transformation on the primary measurement space, with the intention of producing a secondary measurement space in which the effects due to the aforementioned distortions are largely normalized out. These techniques have found the greatest application so far in multifont-character, hand-printing, and handwriting recognition.[18,19]

Another point worth mentioning here is that transformations may be either of the discrete or continuous type. Linear transformations are not necessary since they are automatically incorporated into any linear decision function. Continuous nonlinear transformations have the greatest theoretical merit (as discussed later) because they do not influence the size of the recognition unit (word, syllable, etc.). However, practical considerations place constraints on the complexity of nonlinear transformations that may be employed. Therefore, it becomes almost a necessity to investigate the usefulness of various "feature detection" transformations which are discrete in nature but easy to implement.

It is one of the purposes of the present work to postulate and compare the effectiveness of various transformations in a speech recognition system. This is done by comparing the performance of recognition systems which differ only by the use of the different transformations, in a standard recognition experiment.

It will be recalled that the application of a transformation of either of the kinds mentioned above can be viewed as a method of producing a measurement space in which the "perfect signal" situation (corrupted by noise) exists. It has been stated that in this situation a linear decision function, in which one linear functional is associated with each class, is the optimum one. This is the decision function used in the present work.

• *The linear decision function: adaptive algorithm*

The decision function to be used here has already been alluded to. If

$$W_i \cdot X + t_i \geq W_j \cdot X + t_j + \epsilon \quad \text{for all} \quad j \neq i, \quad (3)$$

then the measurement X is assigned to class i. Otherwise, the pattern is rejected; ϵ is a fixed positive constant chosen in advance and results in reject zones or regions in the measurement space.

The adaptive procedure used to determine the set of vectors and constants, W_i and t_i is given here. It is a variant of procedures reported previously in other works. Let $\mathfrak{s}_1, \mathfrak{s}_2, \cdots, \mathfrak{s}_n$ be a sequence of words of the vocabulary, of which $\mathfrak{s}_k$ is a particular member. Each word of the vocabulary should occur many times in this sequence. Let $X_1, X_2, \cdots, X_n$ be the corresponding sequence of vectors arising in measurement space due to the sequence of words $\mathfrak{s}_1, \mathfrak{s}_2, \cdots, \mathfrak{s}_n$.

Let T_{i1} be any vector (say the zero vector) and v_{i1} be any number (say, zero), where i denotes the class of word and $i = 1, 2, \cdots, p$. We define sequences of a set of vectors $T_{i1}, T_{i2}, \cdots, T_{i(n+1)}$ and constants $v_{i1}, v_{i2}, \cdots, v_{i(n+1)}$ iteratively as follows: If $\mathfrak{s}_k$ is from $\mathfrak{S}_i$ and

$$T_{ik} \cdot X_k + v_{ik} > T_{jk} \cdot X_k + v_{jk} + \theta \text{ for all } j \neq i, \quad (4)$$

then,

$$T_{i(k+1)} = T_{ik} \text{ and } v_{i(k+1)} = v_{ik},$$

$$T_{j(k+1)} = T_{jk} \text{ and } v_{j(k+1)} = v_{jk};$$

but if

$$T_{ik} \cdot X_k + v_{ik} \leq T_{jk} X_k + v_{jk} + \theta \text{ for any } j \neq i, \quad (5)$$

then

$$T_{i(k+1)} = T_{ik} + B X_k \quad \text{and} \quad v_{i(k+1)} = v_{ik} + B,$$

$$T_{j(k+1)} = T_{jk} - X_k \quad \text{and} \quad v_{j(k+1)} = v_{jk} - 1,$$

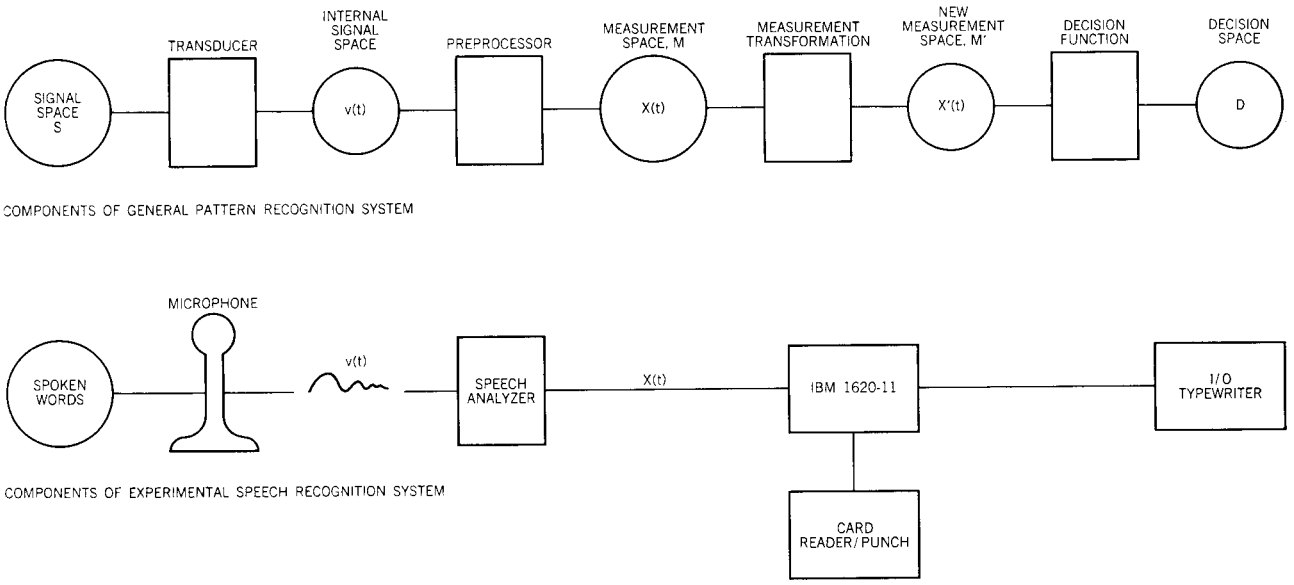


Figure 4 Block diagrams showing correspondence between experimental system components and those of a general pattern recognition system.

where B is the number of linear functionals indexed by j for which Eq. (5) is satisfied and θ is a fixed positive constant chosen in advance.

The sets of vectors $T_{i(n+1)}$ and constants $v_{i(n+1)}$ are a tentative choice for W_i and t_i, respectively. The process can be repeated on the sequence of words $\mathfrak{F}_1, \mathfrak{F}_2, \cdots, \mathfrak{F}_n$ (sometimes called the analysis sample or training sample) until no further corrections are made in a pass. This condition is called convergence, and can be expected to occur only for certain distributions of the measurements, X, in the measurement space.

The vector W_i and constant t_i, chosen in this way will be "good" choices, if the sequence of measurements due to the words in the analysis sample is representative of those measurements to be encountered in the later environment of the recognition system.

There is another adaption algorithm which has been found useful[15–17] for obtaining a set of linear functionals useful for recognition using the decision function of Eq. (3). It is an iterative routine which attempts to find, for a sample sequence of measurement vectors $X_1, X_2, \cdots, X_n$, a set of linear functionals, $L_1, L_2, \cdots, L_p$ such that for all $X \in \mathfrak{S}_i$,

$$L_i > \theta \quad \text{for} \quad i = 1, 2, \cdots, p$$

and

$$L_i < -\theta \quad \text{for all} \quad j \neq i.$$

- *The experimental system*

The Transformation and Decision Sections of the recognition system explored here are simulated using an IBM 1620 computer. The 1620 operates fast enough so that recognition can be accomplished on-line (for the vocabulary size used here), and the flexibility provided by a general purpose digital computer allows the desired investigation of various transformations and vocabulary sizes.

Thus, the speech analyzer described here was connected on-line to an IBM 1620-II computer. A programming system for the 1620 was developed that allows the various transformations and decision functions to operate on the original measurement space. The system is capable of (1) applying these various transformations, either singly or in series, to the original space, and then (2) applying the decision function (or categorizer) to the transformed measurements.

In addition, by preserving in punched card form the primary measurements resulting from spoken utterances, a universe of patterns may be built up for one or many speakers allowing various comparative experiments to be performed. This system and its correspondence to the parts of a general decision-theoretic recognition model are shown in Fig. 4.

Rationale for a suitable set of recognition experiments

This section presents the reasons for the choice of the particular experiments reported here. The hypothesis has been made that the spectrum analysis of a speech waveform provides measurements that contain, if they are not themselves, statistically invariant measures of the spoken word. If this is so, then for single-speaker word recognition and

for a particular vocabulary, the performance level (error rate) of the speech recognition system using this measurement space (or some transformation of this space) should be relatively independent of the particular speaker. To verify this, experiments designed to yield the following results should be conducted:

(a) Performance level of the system using a fixed vocabulary for a number of different speakers singly.
(b) Performance level of the system using one speaker (or a fixed set of speakers) for a number of distinct and arbitrary vocabularies of equal size.

The method used to obtain the "performance level" is as follows: The measurements, X, resulting from a large number of utterances of the vocabulary involved, by the speaker concerned, are recorded in binary form on punched cards. This sample is divided into two parts, the analysis sample and the recognition sample. The analysis sample is used to obtain the linear functionals L_i by one of the adaptive methods described earlier; the recognition sample is used to test the system. The "performance level" is the error rate of the system on a recognition sample, after the W_i and t_i have been determined using the analysis sample. The size of the analysis sample should be large enough to represent the utterances in the recognition sample. If the measurement apparatus is stable, and the measures are truly statistically invariant, the W_i and constants t_i obtained from the analysis sample of one speaker should allow good performance on the recognition sample of another speaker. Thus, further experiments should be conducted yielding:

(c) Performance level of the system adapted on an analysis sample of one speaker and tested on a recognition sample of a different speaker.
(d) Performance level of the system adapted on an analysis sample of a set of speakers and tested on a recognition sample of the same set of speakers.

From an information-theoretic point of view it would be expected that for the same information channel capacity, the error rate will increase as the number of possible messages (vocabulary size) is increased.[15] The following experiment should be done to verify this:

(e) Performance level of the system using one speaker for vocabularies of increasing size.

From this same consideration, the lowest error rate should be obtained by using the largest possible message length as the unit of recognition.[15] The largest possible message length, in this case, is a single word of the vocabulary. Certain of the transformations of interest in this work essentially perform the recognition of subunits of the word. Transformations 1 and 2, to be described later, recognize topological features in the original measurement space and "characteristic sounds," respectively. Thus, the following result will be of interest:

(f) Performance levels of systems using different units of recognition (as performed by the Transformation Section) for the same analysis and recognition samples.

The above result is of further interest, for certain transformations may produce a measurement space that is closer to the perfect-signal space than the original measurement space (as noted earlier under "Choice of Decision Function") which might more than offset the effects due to changing the size of the unit of recognition.

Experimental results

The experiments reported here were carried out according to the rationale of the previous section. In some of these experiments (1 through 6) primary measurements, X, were used and the Decision Section was composed of fifteen linear functionals derived by the method shown in the part of the paper entitled "Linear Decision Function, Adaptive Algorithm." The analysis sample used for each speaker is noted in each table. In all the recognition experiments the results are reported in terms of a "forced decision" substitution rate; that is, no reject errors are permitted ($\epsilon = 0$ in Eq. (3)). The intent is to allow better comparison of these results to those of other workers, for there are many possible ways of introducing reject criteria into a decision function, and comparisons of reject rates may be misleading.

The other experiments (7 through 9) involved processing the original measurements, X, in such a way that the basic units of recognition were smaller than the word.

The vocabularies used throughout the experiments were the following:

Vocabulary 1: *one, two, three, four, five, six, seven, eight, nine, zero, minus, plus, times, over, total.*
Vocabulary 2: *clear, patterns, weights, date, I.D., learn, print, code, punch, process, read, speaker, cards, report, run.* This list was chosen without regard for phonetic content; it is simply a number of words used in the programming system for the IBM 1620.
Vocabulary 3: This vocabulary is the conjunction of Vocabularies 1 and 2.

- *Group A: Primary measurement space*

Experiment 1: Fixed vocabulary, different speakers. The most extensive experimentation was performed on Vocabulary 1. This particular set of words was chosen to allow the direct dictation of simple arithmetic problems to the recognition system. This is merely a fringe benefit gained from the use of a general purpose digital computer,

WORD	RECOGNIZED AS														
	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	ZERO	MINUS	PLUS	TIMES	OVER	TOTAL
ONE	89								1						
TWO		90													
THREE			90												
FOUR				90											
FIVE					90										
SIX						90									
SEVEN							90								
EIGHT								90							
NINE									90						
ZERO										90					
MINUS											90				
PLUS												90			
TIMES											1		89		
OVER														90	
TOTAL				1											89

(a)

WORD	RECOGNIZED AS														
	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	ZERO	MINUS	PLUS	TIMES	OVER	TOTAL
ONE	75														
TWO		74	1												
THREE			75												
FOUR				73										2	
FIVE					75										
SIX						75									
SEVEN							75								
EIGHT								75							
NINE									75						
ZERO							1			74					
MINUS											74		1		
PLUS												75			
TIMES									1				74		
OVER	1			1				1						72	
TOTAL															75

(b)

WORD	RECOGNIZED AS														
	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	ZERO	MINUS	PLUS	TIMES	OVER	TOTAL
ONE	114		1												
TWO		115													
THREE			112	1				1							1
FOUR		1	1	111											2
FIVE					113				1						1
SIX						115									
SEVEN				1			114								
EIGHT		1						113							1
NINE			1						112				2		
ZERO			1							113					1
MINUS											113	1	1		
PLUS												115			
TIMES	1		1		1				4		1		107		
OVER				1				2		1				110	1
TOTAL		1			7				1						106

(c)

Figure 5 Confusion matrices for three speakers of Vocabulary 1. Experimental conditions are given in Table 1. (a) Speaker A, (b) Speaker B, (c) Speaker C.

but it provides an interesting demonstration of speech recognition. Table 1 compares system performance for three different speakers of Vocabulary 1. The confusion matrices for these speakers are shown in Fig. 5.

Analysis sample sizes used in this experiment varied from 35 to 50 alphabets and recognition sample sizes from 75 to 115 alphabets. (The word "alphabet" is used here to denote one set of utterances of the words in a vocabulary.) In all cases, of course, the analysis and recognition samples were distinct. The number of bits in this original measurement space is 320; the space is in the form of a matrix having 16 rows corresponding to the 16 frequency bands, and a maximum of 20 columns.

The forced decision substitution rate for the three speakers ranged from 0.2 to 2.4%. Speakers A and B were male whereas Speaker C was female. The difference in performance on Speakers A and B is statistically significant; to what this difference should be attributed is not yet clear, but it has been suggested that there might have been more speed-of-talking variation with Speaker B. This speculation is unconfirmed since we have not as yet run an experiment with a destretching normalization.[2,6]

The order of magnitude poorer performance with the female voice can be at least partially attributed to the higher frequency of the voice fundamental. Visual examination of the quantized spectrograms on the display matrix revealed that the location of the lowest frequency formant using this technique is much more erratic with female voices. This result demonstrates the fundamental weakness of the short-term Fourier analysis in locating the poles of the vocal tract.

The error rates with Speakers A and B compare favor-

Table 1 Effect of different speakers on recognition of Vocabulary 1. Analysis sample and test sample were obtained from same speaker in each case. Measurement space contains 320 bits/word. Analysis $\theta = 200$.

			"Forced decision" test results			
Speaker	*Number of analysis sample alphabets*	*Number of test alphabets*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
A	35	90	1347	3	99.8	0.2
B	50	75	1116	9	99.2	0.8
C	40	115	1683	42	97.6	2.4

Table 2 Recognition of different vocabularies as spoken by same person. Analysis and test samples were both obtained from Speaker B. Untransformed measurement space contains 320 bits/word. Analysis $\theta = 200$ for Vocabularies 1 and 2, and 100 for Vocabulary 3.

			"Forced decision" test results			
Vocabulary	*Number of analysis sample alphabets*	*Number of test alphabets*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
1	50	75	1116	9	99.2	0.8
2	45	80	1190	10	99.2	0.8
3	45	30	888	12	98.7	1.3

Table 3 Recognition of Vocabulary 1 with recognition and analysis samples obtained from different speakers. Measurement space contains 320 bits/word. Analysis $\theta = 200$.

			"Forced decision" test results			
Speaker	*Number of analysis sample alphabets*	*Number of test alphabets*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
Analysis: A Test: B	35	20	163	137	54	46
Analysis: B Test: A	50	20	255	45	85	15

ably with any that may be found in the literature on speech recognition. Although the experiments thus far have been on a limited number of speakers, the size of the samples used allows confidence that these results represent the true recognition capability of such a system for single speakers.

Experiment 2: Same speaker, different vocabularies of equal size. Table 2 compares system performance for the equal-size Vocabularies 1 and 2, as spoken by the same person. The Table also shows the performance when the conjunction of Vocabularies 1 and 2 (Vocabulary 3) was spoken by this person.

The forced decision substitution rate on Vocabulary 2 was 0.8%, identical with the recognition result on Vocabulary 1. As noted, this second vocabulary was somewhat arbitrarily chosen. It seems, therefore, that the recognition performance is not a function of the particular vocabulary used. This experiment tends to confirm the generality of the measurement space and its applicability to the recognition of arbitrary sounds.

Experiment 3: Analysis and recognition samples from different speakers. Table 3 shows the recognition performance on Speaker B when the analysis sample is from Speaker A and the performance on Speaker A when the analysis sample is from Speaker B. This experiment was intended to indicate the degree of invariance of the measurements, $X(t)$ with respect to different speakers.

As shown in the Table, the analysis sample of one

speaker is not a good statistical representation of the test sample of a different speaker. It must be noted, however, that the performance of the system was significantly better than chance, implying that there is some degree of statistical invariance in the measurements $X(t)$ from one speaker to another.

Experiment 4: Analysis and recognition samples from same set of speakers. Table 4 shows the recognition performance on Vocabulary 1 for Speakers A and B after the system has been adapted using the analysis sample of Speakers A and B. For comparison, this table includes the performance data on Vocabulary 1 when the analysis and recognition samples were obtained from Speaker B, alone. In the case of the pair of speakers, the substitution rate was 0.8%, the same as that obtained for Speaker B, alone. It should be noted that, in this case, 15 of the 19 errors originated from the utterances of Speaker B. This experiment indicates that the original measurement space may allow the recognition of a limited vocabulary uttered by an ensemble of male speakers.

Experiment 5: Same speaker, vocabularies of different size. Here the system performance in recognizing the 30-word Vocabulary 3 was compared with its performance on the 15-word Vocabularies 1 and 2. The data for this experiment have already been shown in Table 2.

The forced decision substitution rate on the 30-word vocabulary was 1.3%. On the individual 15-word vocabularies the rate was 0.8%. It is seen that the recognition performance did not drastically deteriorate as the number of words in the vocabulary was increased and suggests that good performance on vocabularies larger than 15 words is achievable.

Experiment 6: "Sample on Change." The measurement apparatus used here samples the speech waveform at a fixed rate, in this case at 30 msec intervals. Some workers have suggested that it is more efficient, in terms of total measurements made, to sample the speech waveform only when there is a detected change in its frequency composition.[3] This method of sampling has been simulated using the IBM 1620; Table 5 shows recognition results when this method of sampling was used with the same original measurements, X. In this experiment the original analysis and recognition samples were the same as those used in Experiment 1.

It should also be noted here that the measurement apparatus as originally constructed sampled the spectrum analysis on the basis of detected changes of state of the input lines to the sample register (see Fig. 2). It was experimentally determined that a sampling dead-time of about 15–30 msec produced patterns of the best uniformity. Shorter dead time resulted in large variations in the pat-

Table 4 Recognition of Vocabulary 1 with recognition analysis samples obtained by joining individual samples from Speakers A and B. Results with Speaker B alone are shown for comparison. easurement space contains 320 bits°word. Analysis θ = 200.

			"Forced decision" test results			
Speaker	*Number of analysis sample alphabets*	*Number of test alphabets*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
A + B	85	165	2456	19	99.2	0.8
B	50	75	1116	9	99.2	0.8

Table 5 Effect of "sample on change" transformation on recognition of Vocabulary 1 as spoken by Speaker A. Measurement space contains 320 bits/word max. Analysis θ = 100 for transformed results; θ = 200 for others.

			"Forced decision" test results			
Transformation	*Number of analysis sample alphabets*	*Number of test alphabets*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
Sample on change	35	90	1323	27	98.0	2.0
Fixed sampling rate	35	90	1347	3	99.8	0.2

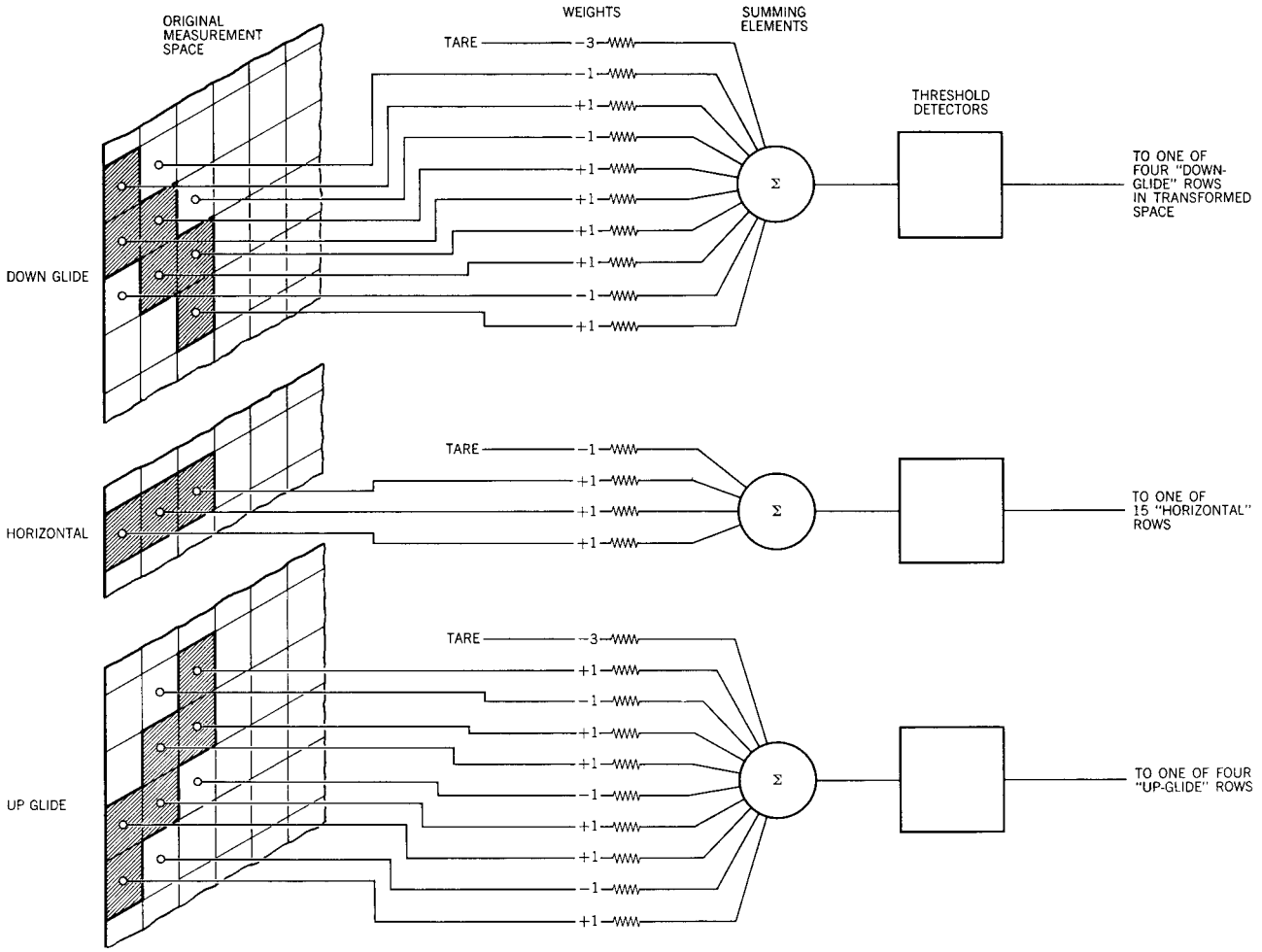


Figure 6 Threshold logic for topological feature transformation. Feature shapes are outlined on the drawing. If the sum of the weighted inputs to a summing element is ≥ 1, the threshold detector will set a bit that indicates identification of a feature into a predetermined output-matrix cell. Starting with columns 1, 2, 3, of the input matrix, the system checks for "up-glides and "down-glides" in each of four overlapping sets of rows, and for "horizontals" in each of 15 rows. It repeats the process in columns 2, 3, 4; 3, 4, 5; etc.

terns of any given class; also, the number of samples exceeded 20($n \cong 320$ bits) for some utterances. Preliminary recognition experiments gave relatively poor results using this sampling scheme, and it was soon abandoned. It will be observed that this result is in contradistinction to the philosophy of some other workers.[3] It must be emphasized, however, that the "sampling on change" performed here by computer processing of the original measurement is different only in detail from the "sample on change" schemes described in the literature.

It is difficult to say whether the poorer performance was due entirely to the greater variability in the number of samples (stretching) for each word class, or partly to the fewer average number of samples. It would seem that most of the performance degradation could be attributed to the greater variability of the resultant measurements.

• *Group B: Transformations of primary measurement space*

To date, three transformations have been investigated. Transformations 1 and 2 are of the type that recognize subunits of the spoken word; they are described below. Transformation 3 has been used to simulate operations that could be performed by a modified measurement apparatus; this method produces fewer bits in the measurement space. It has been used in series with Transformations 1 and 2. In this set of experiments only one speaker's (Speaker A) performance was tested. The analysis and test samples were derived from the same primary measurements used for the analysis and test sample in Experiment 1.

Transformation 1. This transformation applies a layer of feature detectors (see Fig. 6) to the original measurement

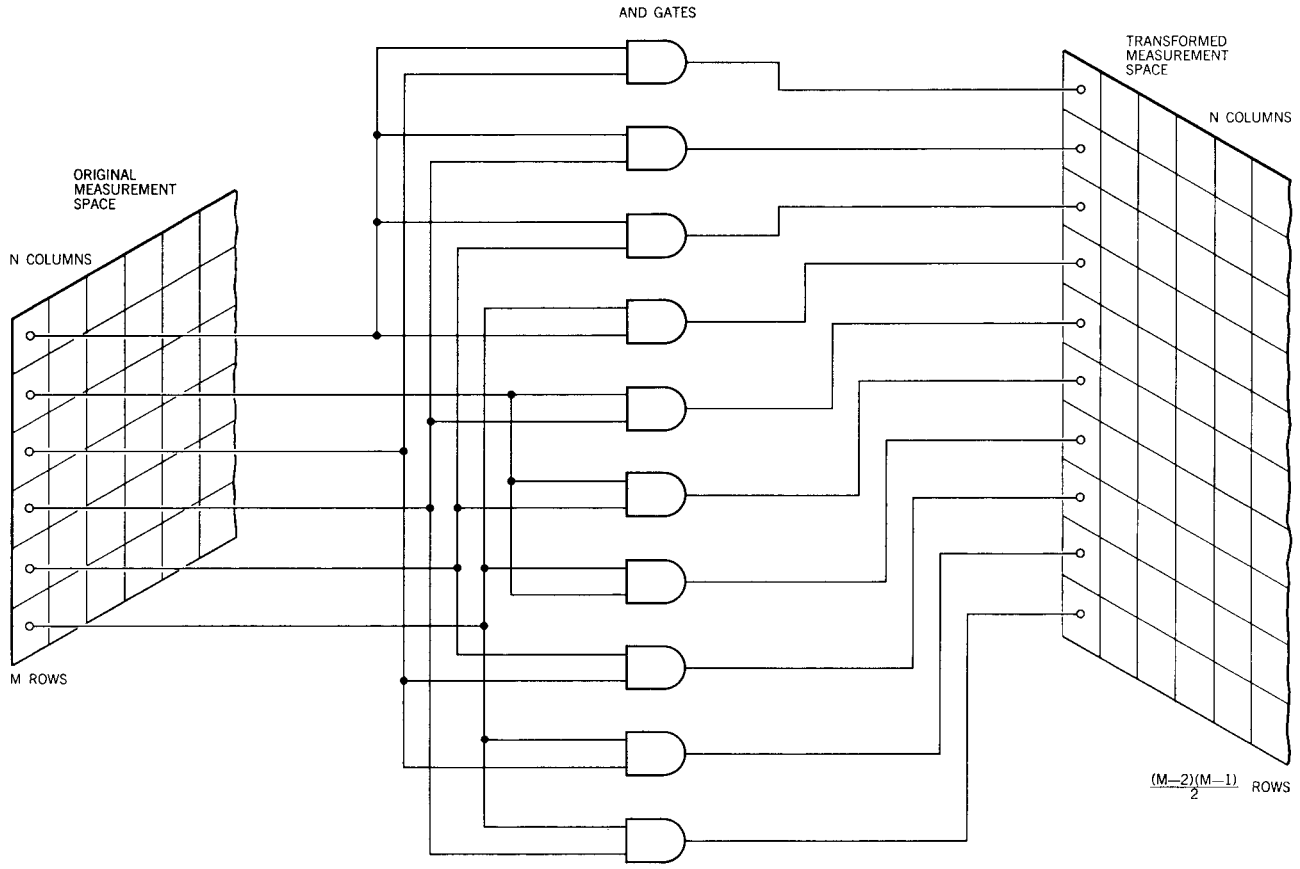


Figure 7 Example of logic for performing a 2-tuple transformation. A six-row input matrix is shown here for simplicity. Each cell in a column is paired with all other cells in the column except for those that are adjacent to it. The same logic is performed, in turn, on each column. In the experiment reported here, the 2-tuples were applied to rows 5–15 of the basic 16-row input matrix.

space, resulting in a new feature space. There were two kinds of feature detectors applied to the original space, those which detected formant glides and those which detected steady formants.

Transformation 2. This transformation applies "*n*-tuple" detectors to each column of the original measurement space. The sets of *n*-tuples are used to detect the occurrence of particular sounds during the word (e.g., combinations of formant frequencies). Both 2-tuple (Transformation 2a) and 3-tuple (Transformation 2b) detectors were used (see Fig. 7).

Transformation 3. This is a transformation designed to reduce the number of columns in the measurement space presented to the decision function. The number of columns hitherto have been dictated by the duration of the word and the sampling rate. This particular transformation arbitrarily compacts the measurement space (the *n*-tuple or feature space) to 3 columns. To accomplish this the *n*-tuple or feature space is divided as near as possible into 3 equal portions; the columns within each portion are ORed together to preserve the occurrence of a particular feature within any portion of the word. Thus, the features or *n*-tuples are zoned into three relative portions of each word: early, middle, and late.

Experiment 7: Fixed vocabulary, single speaker, zoned feature detectors with Transformations 1 and 3. In this experiment the primary measurements are successively transformed by Transformations 1 and 3 resulting in a three-zoned feature measurement space. The number of bits in this measurement space is 72.

Experiment 8: Fixed vocabulary, single speaker, zoned column 2-tuples with Transformations 2a and 3. The application of Transformations 2a and 3 in series yields a three-zone measurement space of 144 bits, since there are forty-eight 2-tuples in each zone.

Experiment 9: Fixed vocabulary, single speaker, zoned column 3-tuples with Transformations 2b and 3. This experiment was identical to Experiment 8 except that 3-tuples

were utilized instead of 2-tuples. Eighty-eight different 3-tuples were used, giving 264 bits in this measurement space. The performance of the various transformed measurements is compared in Table 6. The transformations used in these experiments recognize subunits of the word and are independent of the length of the word, itself.

It should be emphasized that these experiments were done using the same analysis and test samples of spoken utterances throughout. Thus, the only variable in each of these experiments is the transformation on the original measurement space.

None of the transformations provided a system with a performance as good as that obtained by using the original measurement space. From the point of view of information theory this is not a surprising result, but it was of interest to see how close to this performance a transformed measurement space could come. It will be observed from the table that Transformations 2a and 3 in series provide a result comparable to that obtained with the original measurement space even though a significantly smaller number of bits is used.

Conclusions

A frequency-quantized, continuous, short-term, spectrum-analysis technique capable of extracting statistically invariant properties of human speech has been described. The effectiveness of this measurement depends to some extent on the particular speaker, as evidenced by the order of magnitude poorer performance on Speaker C (a female). Apparently a high-pitched voice (female) makes for considerably less reliability in detection of the formants by this technique.

The transformations investigated here yielded substitution rates from 3 to 10 times higher than those obtained using the original measurement space, but with a significantly reduced number of bits per word in some cases. Transformations that will eliminate the effects of variations in speed of talking and other differences in formant structure between speakers must still be found.

Finally, it may be concluded that the techniques investigated in this work are adequate for achieving a forced decision recognition rate of at least 98% over a range of male speakers and for arbitrary vocabularies of up to thirty words.

Since the running, short term spectral envelope is not always a reliable method of locating the vocal tract poles (formants) alternative methods should be explored. A few techniques exist that have been briefly experimented with by others and show some promise of better accuracy and reliability.[21] However, some of them do not operate in real time. Other areas which require additional work include the problems of word segmentation, and discrimination of voiced, unvoiced, and mixed speech from background noise.

Given better formant locators, speed invariant representations of the spoken word (such as that obtained from its representation as a trace in a formant frequency space) and other transformations should be investigated.[4]

Acknowledgments

Some of the ideas embodied in the experimental apparatus grew out of discussions between the authors and G. L. Clapper. Indeed, if it were not for some of Mr. Clapper's earlier successes in this same field, this work probably would not have been undertaken. The authors wish to thank Mrs. E. Ide and C. Kiessling for their joint efforts in developing a programming system for the IBM 1620-II suitable to our special requirements; R. Ramm for the complete design of the input buffer unit to the 1620 and his assistance in the design of the operational amplifier used extensively in the preprocessor; and L. LaBalbo for his technical assistance in constructing most of the experimental hardware.

References

1. N. M. Abramson, W. E. Dickinson, and F. B. Wood, "The Application of Decision Theory to Voice Recognition Machines With an Illustrative Example," IBM Technical Report 16.01.071.003, December 1959.
2. P. Denes, and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," *J. Acoust. Soc. Am.* **32**, 1450–1455 (1960).

Table 6 Comparison of system performance with various transformations of measurement space. In all cases the analysis sample and the test sample were 35 and 90 alphabets, respectively; all were obtained from Speaker A. Analysis sample $\theta = 100$ for all except original measurement space where $\theta = 200$.

		"Forced decision" test results			
Transformation	*Size of measurement space in bits/word*	*Number of correct recognitions*	*Number of substitutions*	*Percent correct*	*Percent substituted*
None	320	1347	3	99.8	0.2
1 + 3	72	1336	14	99.0	1.0
2*a* + 3	144	1342	8	99.4	0.6
2*b* + 3	264	1325	25	98.1	1.9

3. H. F. Olson, and H. Belar, "Syllable Analyzer, Coder and Synthesizer for the Transmission of Speech," *IRE Transactions on Audio* **AU-10**, 11–17 (1962).
4. K. H. Davis, R. Biddulp, and S. Balashek, "Automatic Recognition of Spoken Digits," *J. Acoust. Soc. Am.* **24**, No. 6, 637–642 (1952).
5. L. R. Talbert, G. F. Groner, J. S. Koford, R. J. Brown, P. R. Low, C. H. Mays, "A Real-Time Adaptive Speech-Recognition System," Technical Documentary Report No. ASD-TDR-63-660, Stanford Electronics Laboratory, May, 1963.
6. J. E. Dammann, "Application of Adaptive Threshold Elements to the Recognition of Acoustic-Phonetic States," *J. Acoust. Soc. Am.* **38**, 213–223 (1965).
7. W. C. Dersch, "A Decision Logic for Speech Recognition," *Symposium on Living Systems—The Key to New Technology*, Wright Air Development Division, Dayton, Ohio, September 13–15, 1962.
8. F. Rosenblatt, "A Description of the Tobermory Perceptron," *Collected Technical Papers, Vol. 2*, Cognitive Systems Research Program, Cornell University, 1963.
9. N. Lindgren, "Machine Recognition of Human Language—Parts I and II," *IEEE Spectrum* **2**, 114–136 (March, 1965) and 45–59 (April, 1965).
10. M. Minsky, "A Selected Descriptor-Index Bibliography to the Literature on Artificial Intelligence," *IRE Transactions on Human Factors in Electronics* **HFE-2**, 39–55 (1961).
11. A. Wald, *Statistical Decision Functions*, J. Wiley and Sons, New York, 1950.
12. H. W. Dudley, "The Carrier Nature of Speech," *Bell System Technical Journal* **19**, 495–515 (1940).
13. H. W. Dudley, "The Automatic Synthesis of Speech," *Proc. Natl. Acad. Sci.* **25**, 377–383 (1939).
14. E. H. Rothauser, "Dependence of Speech Quality on Transmitted Information Rate in a Band Compression System," IBM (Austria) Internal Publication.
15. J. S. Griffin, J. H. King, and C. J. Tunis, "A Pattern Identification System Using Linear Decision Functions," *IBM Systems Journal* **2**, 248–267 (1963).
16. H. J. Greenberg and A. G. Konheim, "Linear and Non-Linear Methods in Pattern Classification," *IBM Journal* **8**, 299–307 (1964).
17. A. Novikoff, "On Convergence Proof for Perceptrons," *Proceedings of the Symposium on Mathematical Theory of Automata*, April 24–26, 1962, Polytechnic Press, Polytechnic Institute of Brooklyn, Brooklyn, N. Y.
18. L. A. Kamentsky and C. N. Liu, "Computer Automated Design of Multifont Recognition Logic," *IBM Journal* **7**, 2–13 (1963).
19. E. C. Greanias, P. F. Meagher, R. S. Norman and P. Essinger, "The Recognition of Handwritten Numerals by Contour Analysis," *IBM Journal* **7**, 14–21 (1963).
20. R. M. Fano, *A Statistical Theory of Communications*, published jointly by M.I.T. Press and John Wiley and Sons, Inc., New York, London, 1961.
21. W. H. Huggins, "A Phase Principle for Complex-Frequency Analysis and Its Implications in Auditory Theory," *J. Acoust. Soc. Am.* **24**, 582–589 (1952).

Received March 16, 1965.