[10] M. R. Sambur, "An efficient LPC vocoder," in preparation.
[11] J. D. Markel, *Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation* (SCRL Monograph No. 7). Speech Communication Res. Lab., Santa Barbara, Calif., Oct. 1971.
[12] W. S. Meisel, *Computer-Oriented Approaches To Pattern Recognition.* New York: Academic, 1972, pp. 13–15.
[13] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.

Selection of Acoustic Features for Speaker Identification

MARVIN R. SAMBUR

Abstract—**The aim of this study was to determine a set of acoustic features in the speech signal that are effective for the identification of a speaker. The investigation examined a large number of theoretically attractive features. The analysis technique of linear prediction was incorporated to examine features that were previously ignored because their measurement was either too time consuming or not easily amenable to automatic measurement.**

A novel probability of error criterion was used to determine the the relative merits of the features. The experimental data base was collected over a $3\frac{1}{2}$ year period and afforded the oportunity to investigate the variation over time of the measurements. The measurements that were found to be the most important were the value of the second resonance (around 1000 Hz) in /n/, the value of the third or fourth resonance (1700–2000 Hz) in /m/ the values of the second, third and fourth formant frequencies in vowels, and the average fundamental frequency of the speaker.

A speaker identification experiment using only the best five features was performed. The test data consisted of the multisession data of 11 speakers, and the test data was kept independent of the design data. One error was made in the identification of these speakers for 320 separate identification experiments.

I. INTRODUCTION

A CRUCIAL ingredient in the success of any pattern recognition system is the selection of features that efficiently characterize the patterns of interest. This paper reports on a study [1] undertaken to determine a set of acoustic features in the speech signal that are effective for the identification of a speaker.

The investigation was conducted by first determining an initial set of acoustic parameters which, on the basis of theoretical considerations and past experimental work [2]–[4], might be suitable candidates for indicating the unique properties of a speaker's vocal apparatus, as well as some aspects of his learned pattern of speaking. The initial selection of features was also made to take advantage of the speech-analysis technique of linear prediction [5]. This analysis technique provided a quick and convenient measurement of many theoretically important speaker characterizing properties that have not been incorporated in recognition schemes because of the inefficient methods available for their measurement. These parameters include formant bandwidths, glottal source "poles," and the pole locations during the production of nasals and strident consonants. The initial list of speaker characterizing features also included the formant structure of vowels, the duration of certain speech events, the dynamic behavior of the formant contours, and various aspects of the pitch contour throughout an utterance. In all, a total of 92 features was examined in the study.

To determine the relative merits of the features, a novel probability of error approach was devised. This method evaluates the features in accordance with their relative contribution to the performance of a given speaker recognition scheme. The experimental data used in the evaluation were collected over a $3\frac{1}{2}$ year period and afforded the opportunity to investigate the variation over time of the features. The goal of this evaluation was an ordered list of speaker characterizing measurements that could guide an individual in selecting features to incorporate in a speaker recognition system. Before considering the merits of the features investigated, we shall discuss the method used in this paper to evaluate the measurements.

II. FEATURE EVALUATION

The problems associated with evaluating the relative effectiveness of a set of features can be best understood within the concept of a feature space. If N features are to be measured in the recorded speech of a talker, every replication of the experimental data by the speaker can be represented as a point in what is termed an N dimensional feature space. Fig. 1 illustrates an example of a two dimensional feature space representation of the measured data points for a number of talkers. The statistical nature of the ensemble of measured points for each individual can be considered to be governed by some underlining multidimensional probability distribution that is hopefully different for each speaker. The ability of a set of measure-

Manuscript received June 10, 1974.
The author is with the Bell Laboratories, Murray Hill, N. J. 07974.

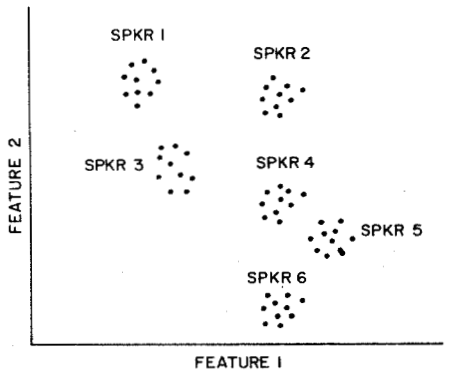

Fig. 1. Hypothetical two-dimensional feature space representation of the measured data of six speakers.

ments to characterize the voice of a talker is directly related to the distinguishability of the individual speaker distributions with respect to these features. The task of feature evaluation is concerned with finding a means of measuring the "distinguishability" of the distributions.

Intuitively, a good parameter for speaker recognition is one for which the individual speaker distributions are as narrow and as widely separated as possible. A statistic which previous investigators [2] have found useful in quantifying this desired property is the F ratio. The statistic is proportional to the ratio of the variance of the means of each speaker's feature distribution to the average value of the variance of each distribution. Thus the farther apart the individual distributions are with respect to their average spread, the higher the F ratio. Although the F ratio is an indication of a feature's effectiveness, it is not optimal in the sense that a feature with a high ratio necessarily contributes more to the performance of a recognition system than a feature with a lower ratio.

Another technique that has been used to select features for purposes of speaker recognition is that of discriminant analysis. This method involves the creation of new features that are linear combinations of the original features. The optimum linear transformation of the original feature space is determined by a combination of eigenvector analysis and F ratio techniques [8]. This method of feature selection suffers from many of the same drawbacks as the F ratio method, and will usually result in a selection of features that can not be interpreted in terms of meaningful characteristics of speakers [9].

The proposed method for feature selection is based upon the obvious fact that the goal of a speaker recognition system is to classify an unknown speaker correctly. This goal implies that the relative merit of a feature should be based upon its contribution to the performance of recognition. In practical terms, if a group of features, G, yields a smaller rate of error than another group of features, then the set G is necessarily a better set of features for recognizing speakers. The probability-of-error criterion can be viewed as a method of feature evaluation that employs "estimates" of each feature's relative performance to determine an ordered list of feature effectiveness.

To understand the probability-of-error criterion, it should be appreciated that the ultimate utility of a feature really depends upon the nature of the classification system that follows it. Only after a classifier is specified can the performance, and hence the relative merit of a feature be evaluated. In this paper, we shall define the relative effectiveness of a set of features as inversely proportional to the error performance of a prescribed classifier using the given feature set. This definition of feature effectiveness can be extended to a technique for ordering the effectiveness of a set of proposed features. A flow diagram description of the method is given in Fig. 2.

Assuming that the total number of features that are originally available is equal to N, the method begins by evaluating the effectiveness (error performance) of each of the N feature subsets with $N - 1$ members. The most effective feature subset is then determined, and the feature not included in this subset is defined as the least important feature. This feature is then eliminated or "knocked-out" from further consideration. The procedure continues until all the features are "knocked-out" from consideration. The ordered effectiveness of the features is then given by the inverse sequence of "knocked-out" features.

Two methods are suggested for determining estimates of the error performance of a set of features. One method is simply to experimentally determine the error rates over a selected test data base. This method requires no ad hoc assumptions about the underlying probability distributions of the feature set and will provide reliable estimates of the error performance of the feature set. However, if the experimental error determination is too costly or too time consuming to perform, a parametric method can be used. This method of estimating the error rates is linked to the fact that the development of a classification system from labeled samples requires a procedure in which the unknown multidimensional distribution associated with a set of features is estimated in some optimum fashion [6]. The classification rule is then determined by finding the appropriate (optimum) decision rules for the estimated densities. The procedure in which the unknown distributions are estimated is termed "learning" and the resulting classification system is called "recognition." In the probability-of-error criterion, the "best" estimated densities that correspond to the application of a particular classifier are utilized to compute the "best" estimates of the performance of a set of features linked to the given classifier.

The most widely used classifier in speaker identification schemes is the relatively simple linear classifier. The linear classifier determines the identity of the unknown speaker by selecting the speaker with the "closest" reference point to the test point. This type of classification can be shown to be based upon a decision rule that estimates the unknown feature distribution with the "best" fitting point Gaussian densities, and then uses the optimum decision rules for these densities [7]. Because of the widespread use of the linear classifier and the fact that the multidimensional Gaussian densities associated with this classifier do not make an unreasonable model of the densities

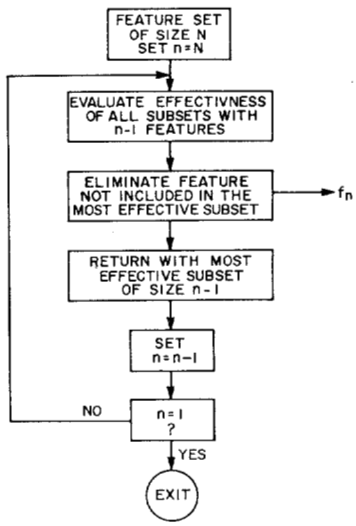

Fig. 2. "Knock-out" procedure for feature selection.

of typical speech parameters [8], [9], the Gaussian error rates were used to evaluate the relative merit of the proposed features. Note that the Gaussian assumptions that were used in the error criterion also apply to the application of the F-ratio method of feature evaluation and discriminant analysis [8], but the probability of error is a more meaningful measure of feature effectiveness.

Since normally distributed random variables are completely specified by their means and covariances, the estimated error rates can be calculated by measuring only the mean feature vector of the measurement set for each talker and the average covariance matrix W among all speakers. If $\boldsymbol{x}_{ij}$ is the p dimension feature vector (where p is the number of features available) corresponding to the jth sample from the ith speaker, and $\bar{m}_j$ is the mean feature vector for the jth speaker, then

$$W = 1/[n(L-1)] \sum_{j=1}^{L} \sum_{i=1}^{n} (\boldsymbol{x}_{ij} - \bar{m}_j)(\boldsymbol{x}_{ij} - \bar{m}_j)^T$$

where L is the number of speakers, and n is the number of samples. To overcome the difficulties associated with computing the error rates in a multidimensional feature space, the union bound on the probability of error was used [10]. Accordingly, for the ith speaker, the probability of error is bounded by

$$P(e_i) < \sum_{j \neq i} P_2(i,j)$$

where $P_2(i,j)$ is just the probability of error for a speaker-recognition system with only the ith speaker and the jth speaker present. For Gaussian densities

$$P_2(k,j) = \int_{d_{kj}/2}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(\frac{-x^2}{2}\right) dx = Q(d_{kj})$$

where

$$d_{kj}^2 = (\bar{m}_k - \bar{m}_j)^T W^{-1} (\bar{m}_k - \bar{m}_j).$$

The total error bound $P(e)$ is then given by the sum of the $P(e_i)$ over the set of speakers.

If the task is speaker recognition, the error bound $P(e)$ can be used as a measure of feature effectiveness. In addition, if the task is to verify the identity of the jth individual (speaker verification), the bound $P(e_j)$ is the appropriate measure of feature effectiveness.

III. EXPERIMENTAL DATA

The set of utterances to be used in a speaker recognition (or speaker verification) experiment should include a wide variety of speech sounds; vowels, fricative, stops, nasals, and diphthongs. The utterances should also be devised in such a way that they are easy to segment, natural to say, and usually spoken in just one way. Based upon these considerations Wolf (1969) used the sentences:

1) Cool shirts please me.
2) Pay the man first, please.
3) I cannot remember it.
4) Papa needs two singers.
5) A few boys bought them.
6) Cash this bond please.

These sentences were also used for our experimentation. Since it is quite possible that an individual's voice may change from day to day, the data were collected over a period of time. The initial session consisted of 21 adult, American male speakers who ranged in age from 22 to 42 years. The recordings of the speech data were made in an anechoic chamber. In order to insure uniformity of the stress patterns in each sentence, the speakers were instructed to first listen to a particular sentence and then to say the sentence in the same manner as that on the program tape. The six sentences were presented to the subject in random order, at intervals of ten seconds, so as to avoid the undesirable effects of "list intonation." A total of ten repetitions for each sentence was recorded.

The second session took place $2\frac{1}{2}$ years after the original session. The third and fourth session were conducted at one month intervals after the second session. The speakers in these sessions consisted of three speakers from the original group and one additional speaker. The recordings were made in the same anechoic chamber, but the speakers were not guided by a program tape in making the recordings. Instead, the speakers were given a stack of cards, and asked to say the sentence in sequence on each card in a natural manner. The sentences were ordered in a random manner and each speaker was instructed to pause between each sentence. Again a total of ten repetitions was made for each utterance.

The fifth session took place $3\frac{1}{2}$ years after the first session and consisted of the 4 speakers used in the second, third and fourth sessions and 7 additional speakers from the original group of 21 speakers. The recordings were made in a similar fashion as those performed in the second through fourth sessions.

IV. MEASUREMENT INVESTIGATED

A. Vowels (Formant Structure and Glottal Source)

Every word contains at least one vowel sound, and the manner in which an individual produces these sounds is an integral part of his unique voice characteristics. The production of vowel sound is influenced by the shape of the speaker's vocal tract and the properties of his glottal source. The formant structure of the vowel spectrum is directly related to the unique shape of the vocal tract [4] and supplies important information about the speaker's identity. In addition, the location of any real axis poles and extraneous wideband poles in the vowel spectrum are related to the individuals glottal source and may also be important speaker characterizing properties. However, before the development of linear prediction analysis, the measurement of formant bandwidth, glottal source "poles" and higher formant frequencies was avoided because the available measurement techniques were either not easily amenable to automatic methods or the schemes were too time consuming. This study provides the first investigation of the speaker recognition potential of these features.

A 12-coefficient linear prediction analysis was used as a means of extracting formants and glottal "poles." The prediction coefficients were computed using the method of Atal and Hanauer [5]. The determination of the exact nature of the computed poles (glottal "poles" or formants) was automatically made on the basis of a combination of bandwidth considerations and data on the expected frequency regions of the first five formants of each analyzed vowel. The features used in the study were the computed poles at the target locations of selected vowels in CVC environments. For purposes of automatic detection, the target position was defined as that point in the vowel segment for which the second formant reached a stationary point (i.e., $dF_2/dt = 0$).[1] Fig. 3 shows the defined target position for a typical CVC utterance. The location of the vowel segment was done manually by viewing the speech waveform.

The vowels analyzed were /ae/ (*c*ash) and /I/ (th*i*s) in sentence 6, and /i/ (n*ee*ds) and /u/ (tw*o*) in sentence 4. In general, the error criterion showed that the formant frequencies were more significant than formant bandwidths and glottal "poles" for speaker recognition, but the other parameters were sometimes quite important for speaker-verification. For example, some speakers had a consistently measured glottal "pole" in the region of 1000 Hz in their production of the vowel /i/ which almost perfectly discriminated them from other speakers. In addition, one talker was found to have an extremely high second formant bandwidth (≈400 Hz) which was a consistent feature of his vowel production.

To interpret the results of the feature evaluation for the formant frequencies, it should be noted that the probability

[1] A 12 pole LPC analysis with a 200 point frame length was computed consecutively throughout the entire vowel segment.

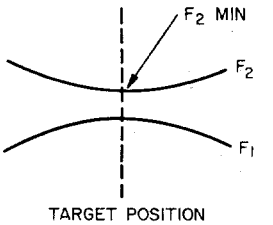

Fig. 3. Definition of target position as location of stationary point in F_2 contour.

of error criterion is intrinsically related to the interspeaker variability (the separation between each speaker's feature distribution) and inversely related to the intraspeaker variability (spread of the speaker's distribution) of a particular feature. The experimental data indicated that the higher order formant frequencies demonstrated the most interspeaker variability. This result is consistent with theoretical observation that length perturbations in the vocal tract are most significantly characterized by changes in the frequencies of the higher order formants [1]. Since the most common physical difference between speakers is the length of the vocal tract [4], this experimental result is to be anticipated. The intraspeaker variability of each formant frequency parameter was seen to be dependent upon the transitional movement of the particular formant contour in the vicinity of the target location, and the ability of the analysis technique to accurately measure the formant frequency. The more rapid the transitional movement, the more probable that the measured target frequency of the formant will be seriously perturbed by slight changes in the speaker's pronunciation, and hence a greater opportunity for intraspeaker variability. The accuracy of the linear prediction analysis was experimentally found to be dependent on the separation between adjacent formants, the bandwidths of the formants, and the amplitude of the formants.

For the front vowels analyzed (/i/, /ae/, and /I/), the error criterion indicated that the best mixture of high interspeaker variability and low intraspeaker variability was to be obtained for the second and fourth formant frequency measures. For the back vowel analyzed (/u/), the error criterion showed that the frequency of $F3$ was the most effective feature.

The error analysis also indicated that the true statistical nature of the feature set was not reflected in only one session of test data. This conclusion was reached in all 92 measurements that were made.

B. Nasals

The nasals /n/ and /m/ were examined in the words *needs*, *remember*, and *man*. The nasal spectrum is closely tied to the nasal cavity and certain attributes of this spectrum have been effectively incorporated in speaker-recognition. However, the measurement of the actual frequencies of the spectral peaks corresponding to the poles of the transfer function has been avoided because of

the difficulties caused by the presence of zeros in the transfer function and the generally higher formant damping of the nasal consonants. Since these pole frequencies are quite significant in specifying the unique attributes of the speaker's nasal cavity [2], [11], it was decided to use linear prediction analysis to measure the pole frequencies.[2] To test the ability of the linear prediction method to extract the pole frequencies, the experimental data obtained by Fujimura [11] for nasals were reanalyzed using the prediction method. Fujimura obtained accurate pole-zero data for the nasals by an elaborate analysis by synthesis scheme. In comparing his results to those obtained by linear prediction, it was found that the prediction method is quite accurate except in regions of pole-zero interplay. In these regions, the frequency of the measured pole was slightly perturbed from the vowel of the true pole and the measured bandwidth was 3–5 times greater than the true bandwidth.

The use of the error criterion showed that the nasal parameters were an especially rich source of recognition and verification features. The most promising measurements were the formant frequencies that were not affected by any pole-zero interplay. The best measurements were the value of the formant frequency near 1000 Hz in /n/ and the value of the third or fourth resonances (1700–2300 Hz) in /m/.

Although the nasal parameters were among the best recognition features, it should be noted that the nasal parameters are particularly vulnerable to physiological changes. Fig. 4 illustrates the variability of the nasal measurements when the speaker is suffering from a mild head cold.

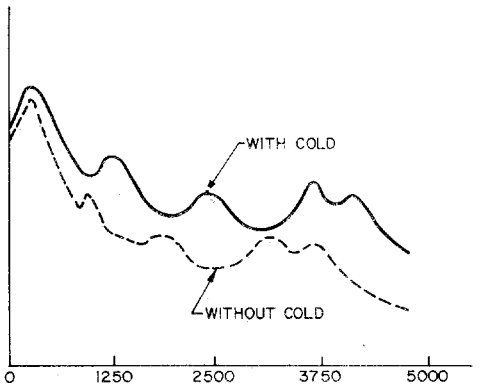

Fig. 4. Comparison of the linear prediction /n/ spectrum of a speaker with and without a mild head cold.

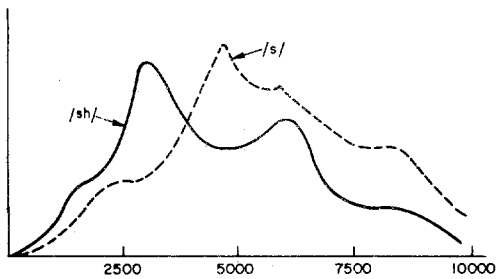

Fig. 5. The linear prediction spectra of /s/ and /sh/ for a typical speaker.

C. Strident Consonants

The formant structure of the strident consonants is influenced by anatomical details around and forward of the alveolar ridge, and hence should display some recognition and verification potential. The stridents were examined in the words *this* (/s/) and *cash* (/sh/) in sentence 6. As in the case of nasals, the measurement of the poles of the transfer function is not easily amenable to automatic extraction because of the influence of zeros and the generally higher damping of the poles. Again linear prediction analysis was used as a means of characterizing the stridents in terms of a set of computed poles. Since the high frequency energy of the stridents is quite important and the stridents have only about 5 poles in the region 0–10 kHz, these sounds were analyzed by first sampling the waveform at 20 kHz and then computing a 10-coefficient spectrum in the middle of the manually located frication region of the strident. A typical computed spectrum for /s/ and /sh/ is depicted in Fig. 5.

The error criterion showed that the stridents were not as significant as the nasals and vowels sounds in characterizing a speaker. However, the frequency of the computed pole near 4300 Hz in /s/ and the value of the computed pole frequency near 3400 Hz in /sh/ did, however, provide some identification potential. These computed poles correspond to the value of the most prominent poles in the spectrum of each strident.

D. Fundamental Frequency

Fundamental frequency parameters have been found to be valuable recognition features by previous investigators [2], [8]. In this study, we examined the Fo contour in the sentence "Cash this bond please." The fundamental frequency was measured by a technique that computed the intervals between zero crossing in the 200 Hz low pass filtered speech waveform [2]. An examination of the actual contours used by a number of speakers throughout this sentence led to the formulation of the stylized intonational pattern of the pitch contour depicted in Fig. 6. Within each of the four delineated voiced sections in the analyzed utterance a set of parameters was employed to characterize the fundamental frequency contour. In the first section (during the /ae/ in "cash"), the parameter RFO represents the slope of a linear mean square fit to the pitch contour from the start of voicing to the maximum Fo (FOMAX) in this section. F1F0 is the slope of the best linear mean square fit from the point of maximum Fo to the end of voicing and F2F0 is the slope during the voiced portion of the word "bond." The average fundamental frequencies associated with the voiced sections and the overall average pitch were also used as recognition parameters.

[2] The location of the nasal sound was performed manually.

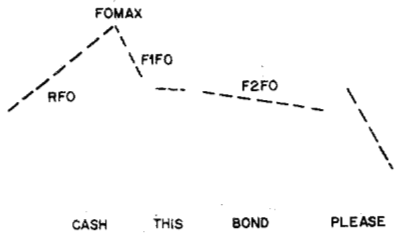

Fig. 6. Stylized intonation pattern for the FO contour in the sentence "Cash this bond, please."

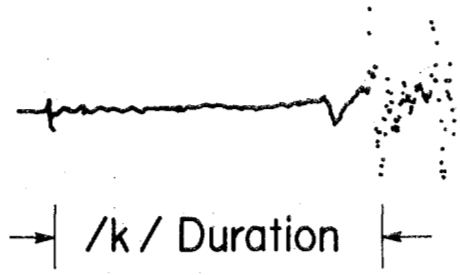

Fig. 7. Fraction and aspiration noise for the plosive /k/ in the word *cash*.

The error analysis showed that the average frequency parameters were more important than the slope features for speaker recognition. The error analysis also indicated that the fundamental frequency of a speaker was quite variable across 5 recording sessions, and this variability diminished the total identification effectiveness of the measurements. The variation of the fundamental frequency with time was as much as 20 Hz for some speakers.

The average fundamental frequency parameters demonstrated their most effectiveness in classifying talkers into three groups composed of individuals with comparatively low fundamentals (about 100 Hz for men), those with a rather typical average fundamental (about 125 Hz) and those with a relatively high average fundamental (about 160 Hz). It was usually the case that the average fundamental frequency of a given individual could radically vary across sessions within each of these classifications but would rarely change from one classification to another.

E. Timing Measurements

The learned behavior of an individual is reflected in the temporal aspects of his speech and provides a source of identification parameters [3]. The timing measurements investigated were the slope of a linear mean square fit to the second formant in the manually segmented diphthong /aI/ and the duration of the frication and aspiration noise of the plosive /k/ in *cash*. Fig. 7 illustrates a typical waveform for the /k/.

The manual measurement of the duration of the noise in the production of this /k/ ranged from 40 ms for one individual to 127 ms for another. This parameter was also quite stable within and across recording sessions and turned out to be an effective feature. The $F2$ slope in /aI/ was also quite variable among speakers and demonstrated excellent identification potential.

IV. OVERALL FEATURE RANKING

The error criterion was used to order the entire set of 92 parameters that were investigated. The first 38 of these features are listed in Table I. The notation used to describe the features in the table consists of an abbreviation for the speech event plus the measurement. Thus NF2 is the value of the second formant frequency in the nasal /n/.

When interpreting this ranking of features, it is important to keep in mind that the ordering is established in accordance with the measurements of a given group of speakers; the speech characteristics of another group may result in a different ordering of features. For example, a group composed of both female and male speakers may result in higher relative ranking of fundamental frequency information than the group investigated here. In any case, the ranking shown in Table I affords a general idea of what features are important in recognizing an unknown speaker. These important features include the value of the second resonance (around 1000 Hz) in /n/, the value of the third or fourth resonance (1700–2200 Hz) in /m/, the values of the second, third and fourth formant frequencies in vowels, the average fundamental frequency of the speaker, and measurements related to the dynamic properties of the talker's voice patterns that reflect his learned behavior of speaking.

TABLE I

Ordering of Features

	Feature	Speech Event		Feature	Speech Event
1.	NF2	/n/	20.	THISFO	FO
2.	UF3	/u/	21.	MANF2	/m/ in man
3.	IF2	/I/	22.	MANB3	/m/ in man
4.	K	duration of /k/	23.	EEF1	/i/
5.	REMF3	/m/ in remember	24.	EEF4	/i/
6.	NF6	/n/	25.	EEF3	/i/
7.	REMF4	/m/ in remember	26.	SHF2	/sh/
8.	CASHFO	FO	27.	AEF2	/ae/
9.	IF4	/I/	28.	AEF4	/ae/
10.	AI	F2 slope in /aI/	29.	AEF1	/ae/
11.	REMFI	/m/ in remember	30.	SF2	/s/
12.	AVFO	FO	31.	UF4	/u/
13.	SF3	/s/	32.	IF1	/I/
14.	UF2	/u/	33.	BONDFO	FO
15.	EEF2	/i/	34.	REMF6	/m/ in remember
16.	NF1	/n/	35.	IF5	/I/
17.	MANF4	/m/ in man	36.	MANB4	/m/ in man
18.	UF1	/u/	37.	AEF3	/ae/
19.	NF3	/n/	38.	SHF1	/sh/

V. RECOGNITION EXPERIMENT

As a by product of the feature evaluation method, an estimate of the error bound for a linear classifier can be obtained. Fig. 8 shows the theoretical error bound as more and more features are sequentially incorporated in the classification scheme. If only the first five features are used, the curve predicts that the error will not exceed

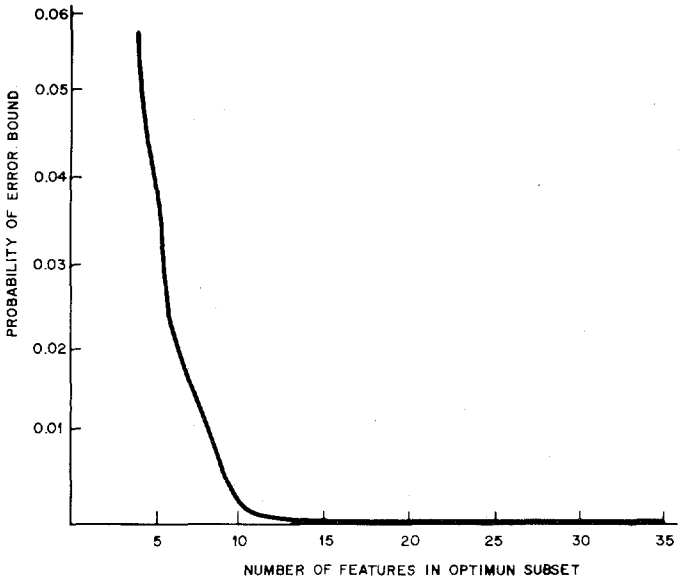

Fig. 8. Theoretical error bound for optimum feature subsets.

3 percent. To test how accurately these error bounds reflect the performance of a set of features, an actual speaker recognition experiment was conducted.

The recognition system examined in the experimentation consisted of only the five "best" features listed in Table I. The multisession data were divided into a design and test set. Each of the data points was used in turn as the test set, while the remaining data points were used as the design set. A total of 320 separate speaker identification experiments were made.

The classification algorithm was the optimum linear classifier discussed in Section II. The classification is performed by calculating the "distances" from each speaker's mean feature vector to the test point. The talker whose mean reference data are "closest" to the test point is classified as the unknown speaker. An error is made when the classified talker is not the speaker who produced the test points.

In the classification algorithm distances between the test point, t, and the kth speaker are given by

$$d_{kt}^2 = (\bar{m}_k - t)^T W^{-1} (\bar{m}_k - t)$$

where $\bar{m}_k$ represents the mean reference vector for the kth speaker and W is the pooled covariance matrix (see Section II). The determination of each $\bar{m}_k$ and the covariance matrix W is made using only the specified design data. The test point had no role in the reference data. There was only 1 error made in the 320 speaker identification experiment. The error rate of 0.00312 is well within the predicted bound of 0.03. This result gives us some confidence that the model on which the error bound calculation is based has some validity.

VI. DISCUSSION

The purpose of this investigation was the isolation of features that would be effective for the task of speaker identification. The core of features initially selected as potentially important speaker characterizing measures was guided by theoretical considerations and past experimental work. Many of the features examined in this report have been investigated for the first time because of the quick and efficient measurement ability of the method of linear prediction analysis. However, many important speaker identification measures can not be extracted using this analysis technique. These parameters include the zero locations in the production of nasals and strident consonants. In addition, the measurement of glottal source "poles" appears to be an unsatisfactory representation of the glottal source and a more elaborate inverse filtering scheme is probably necessary.

The relative effectiveness of the features was evaluated by means of a probability of error criterion. The criterion besides affording an ordered list of feature effectiveness showed that it is important that the test data be collected during many sessions so that the true statistical nature of the measurement set will be reflected. In light of the results of the recognition experiment, the error criterion proved to be quite useful in guiding the researcher in selecting the features that best characterize the speaker.

ACKNOWLEDGMENT

The author wishes to thank Prof. K. N. Stevens for his advice and criticism during the course of this study. The author also acknowledges the guidance of Dr. L. Rabiner, Dr. A. Rosenberg, and Dr. J. Flanagan in the preparation of this paper.

REFERENCES

[1] M. R. Sambur, "Speaker recognition and verification using linear prediction analysis," Ph.D. dissertation, Dep. Elec. Eng., Mass. Inst. Technol., Cambridge, Mass., Sept. 1972.
[2] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044–2056, 1972.
[3] O. Tosi *et al.*, "Experiment on voice identification," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2030–2043, 1972.
[4] K. N. Stevens, "Sources of inter- and intra-speaker variability in acoustic properties of speech sounds," in *Proc. 7th Int. Congr. Phonetic Sciences*, Montreal, Canada, Aug. 21–28, 1971 (Mouton and Company, The Hague, Netherlands, in press).
[5] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, p. 637, 1971.
[6] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York: Macmillan, 1962.
[7] G. S. Sebestyen and A. K. Hartley, "Study program of pattern recognition research," Litton Syst., Inc., Waltham, Mass., Rep. AFSRL 6265, Dec. 31, 1961, AD 273235.
[8] W. S. Mohns, "Statistical feature evaluation in speaker identification," Ph.D. dissertation, Dep. Elec. Eng., North Carolina State Univ., Raleigh, N. C., 1969.
[9] P. D. Bricker *et al.*, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, 1971.
[10] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.
[11] O. Fujimura, "Analysis of nasal consonants," *J. Acoust. Soc. Amer.*, vol. 49, p. 541, 1962.
[12] G. W. Snedecor and W. G. Cochran, *Statistical Methods*. Ames, Iowa: Iowa State Univ. Press, 1967.