THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 4, April, 1971
Printed in U.S.A.

Statistical Techniques for Talker Identification

By P. D. BRICKER, R. GNANADESIKAN, M. V. MATHEWS, MISS S. PRUZANSKY, P. A. TUKEY, K. W. WACHTER and J. L. WARNER

(Manuscript received November 16, 1970)

This paper provides an overview of work on statistical formulations and analyses associated with the problem of identifying persons on the basis of spectral energy representations of acoustical utterances. The investigation has been largely empirical and the paper focuses on the statistical techniques and strategies that have been developed in the context of analyzing two sizeable bodies of data. The problems and procedures to be discussed include: (i) data condensation and representation; (ii) efficient and practical criteria for classification and discrimination; and (iii) strategies for automatic identification of talkers in relatively large populations.

I. INTRODUCTION

Many of us can perhaps recall the experience of identifying a caller on the telephone from a relatively short utterance such as the word "Hello." This might indicate that even short utterances contain sufficient information for identification, and it is an intriguing and interesting problem to inquire whether automatic, objective, accurate and economic methods can be developed for talker recognition. The authors of at least ten papers in the last eight years have reported experiments with (simulated) automatic talker recognizers. Using a variety of approaches to different aspects of the problem, these experimenters have met with strikingly similar success—90 percent (or more) correct recognition.

Previous studies may be classified into two groups according to whether the problem addressed was *verification* (is the speaker who he claims to be?) or *identification* (assignment of an unknown utterance to one person in a given group of speakers). While two studies of the first kind involved 34 voices or less,[1,2] the third[3] and most extensive

(118 voices) was most successful, achieving an average of 99 percent correct verifications. Four studies of the second type used quite different bases for identification in small (10–30 voices) populations, varying from spectral analyses of nasal sounds[4] or whole words[5,6] to measurements of phonological features.[7] Whereas all the above studies required the speaker to utter a prescribed text, three others have achieved from 90 to 100 percent correct identification with no constraints on what the speaker says, provided that a sufficient quantity of speech from each talker is available. Again, the procedures employed in these last three studies have differed widely: spectral analyses of whole speech[8] or of vowels only,[9] and intervals between extremal points in various frequency bands[10] were used with three different recognition schemes.

These results with small populations suggest that the speech signal contains so much information about the talker that one can be distinguished from among 30 or so by a variety of procedures, and that we cannot learn from these studies the relative merits of various ways of representing the signal and reaching a decision. The only one of these studies that used a hundred or more voices[3] required only that each unknown be assigned to one of two classes (genuine or impostor); there have been no studies of identification in large populations. This paper describes the evolution of work addressed to both the small- and large-population identification problems by a group of people, including the present authors, over the last few years. Aside from the present authors, others who have participated in different facets of this work are: Mrs. M. H. Becker, Mrs. L. P. Hughes, T. L. DeChaine, R. S. Pinkham and M. B. Wilk.

The work to be described here evolved empirically and experimentally in the context of analyzing two bodies of data. With no general theory being available to aid in designing a process for talker identification, this work relied heavily on the analysis of data not only to generate ideas and techniques of possible relevance but also to assess the performance of any scheme. Thus, the pragmatic criterion of observed proportions of correct recognition in the two bodies of data was utilized as the touchstone rather than any general theoretical optimality properties. The data analytic orientation in this problem proves to be practical and productive, and most of the successful ideas and methods are fairly obvious—especially after the fact!

The presentation of the data analysis and decision processes may be viewed in four parts: (*i*) *The data*—the two bodies of data studied will be described, the basic digital format of an acoustical utterance will

be described and displayed, and finally, some features of the data will be discussed. (*ii*) *Data condensation*—several primitive procedures will be mentioned for deriving manageably low-dimensional representations from the original data. (*iii*) *Definition of a space and metrics*—the unknown, and the various candidates for assigning it to, may be represented in the space of the summary data, and several metrics may be specified for measuring the distance between the unknown and the candidates for identification. (*iv*) *Classification schemes and strategies for identification*—i.e., procedures for assigning an unknown in a relatively small population of contending speakers, as well as statistical strategies for allocation in relatively large populations of speakers.

II. THE DATA

The two bodies of data involved in our study both deal with repeated utterances of single words. The first set of data (cf. S. Pruzansky[5] for a detailed description) is from ten talkers each of whom yielded several repetitions of ten words commonly used in telephone conversations. The actual utterances were excerpted from sentences in which the words were embedded and the talkers, in fact, read the sentences. For most talker-word combinations there were seven replications with only a few missing. (693 utterances were available instead of $700 = 10 \times 10 \times 7$.)

The second body of data, which was collected subsequent to promising results obtained with the first set of data, deals with a population of 172 speakers each of whom repeated each of five digit names (*one, two, three, four* and *nine*) five times. The words were uttered in isolation rather than being embedded in sentences. The second body of data involved many more speakers, fewer words and fewer replications relative to the first set of data.

Whereas the first set of recordings was made under carefully controlled conditions (see Ref. 5), the second set was made in an unattended booth in a busy concourse. Although a high-quality microphone was used, it was housed in a telephone handset, held a short (but variable) distance from the lips. Automatic equipment controlled a display in the booth which cued the talkers as to which digit name to say and when. In both cases, all utterances by a given talker were recorded in one session.

In the present report, the displays and examples are drawn from analyses of both bodies of data and the presentation will switch back and forth between the two sets of analyses.

The most raw form of the data is just the audio recordings. However, for purposes of analysis, the audio recordings were fed into an analog filter bank and the filter outputs were sampled at fixed, frequent intervals of time (10-millisecond intervals in the first body of data and 6-millisecond intervals in the second set). In the first set of data, the outputs from 17 frequency channels covering a range of 100 to 7000 Hz were retained; the first 16 channels were approximately equally spaced along a Koenig scale from 200 to 4000 Hz, while the 17th covered the range 4000 to 7000 Hz. In the second set of data, the outputs from 20 frequency channels spanning the range 20 to 2900 Hz were retained; the upper and lower cutoff frequencies of each of the 20 filters are shown on the abscissa of Fig. 5. Each audio utterance input thus yielded a certain number (17 in the first set of data and 20 in the second) of separate time series as outputs, with each series representing the energy in a specific frequency band as it varies across time. Together the series represent the short-time spectrum of the utterance.

Thus, the basic digital form of the data for an utterance consists of a matrix of spectral energies classified according to frequency bands in each of a sequence of time intervals. (see Pruzansky & Mathews[6] for a description of energy-frequency-time quantization.) Table I is an example of a data matrix from the second set of data. One can obtain pictorial representations of such a matrix. The classic representation is the sound spectrogram, which is unfortunately not in a form easily read by computers. Figure 1 shows a contour plot of log energy as a function of time and frequency; it was obtained as a computer printout from a data matrix. Although derived in a straightforward way from computer-readable data, this plot conveys some of the visual aspects of the sound spectrogram.

Some comments on certain aspects of the data are in order: (*i*) The total volume of data is large. (*ii*) The basic digital representation of

TABLE I—DATA MATRIX FOR AN UTTERANCE

	Time in milliseconds						
Frequency in Hz.	006	012	018	024	030	036	...
0–100	14	11	7	19	35	62	...
50–150	16	17	11	20	44	74	...
100–200	14	8	16	17	25	56	...
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	

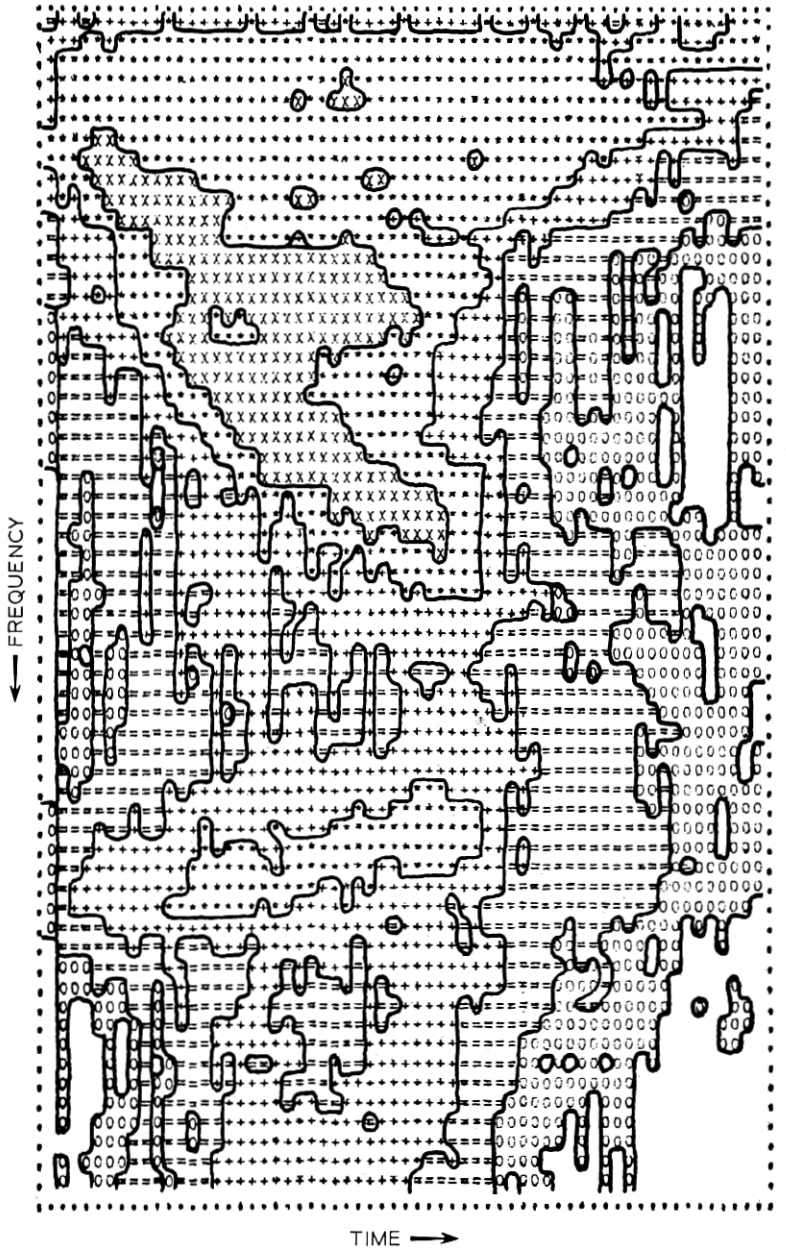

Fig. 1—Contour plot of log energy surface.

the data for an utterance is intractably high in dimensionality for performing statistical analyses. [For the first set of data the matrices were 17×50 (approx.), or 850-dimensional, per utterance; for the second they were 20×275 (approx.), or 5500-dimensional, per utterance!] (*iii*) The general level of the energies may shift from utterance to utterance of even the same speaker due to artefactual reasons. (Loudness may vary for example because of varying proximity to the microphone.) (*iv*) There is no natural time origin for the data and its specification is arbitrary in that what is labelled as time slot 1 does not depend on the actual commencement of the utterance; this implies a lack of alignment of the data for different utterances of a given word even by the same speaker.

The conjunction of the above four issues conveys certain implications for the subsequent analyses. First, it is essential, even for exploratory investigations, to pay attention to practicality and efficiency in computer procedures. Second, it is crucial to find effective lower-dimensional representations of the data using methods of summarization that will be of general utility both for different persons and for different words. Finally, adjustment must be provided for artefactual effects, such as energy-level variation and arbitrariness of the time origin. Such adjustments may be accomplished either by treating the data prior to analysis or by adopting analytical procedures which make provisions for the artefactual effects. Thus, for example, energy-level variation can be handled either by normalizing the energies so that their sum is unity for each utterance or by using classification procedures which allow for level changes amongst the replicated utterances of a speaker (cf. R. Gnanadesikan & M. B. Wilk[11]). Similarly, the arbitrariness of time origin may be handled either by pre-aligning the utterances by some criterion, such as the one used by Pruzansky[5] with the first body of data, or by using origin-invariant time information in later analyses.

III. DATA CONDENSATION

The high dimensionality of the basic quantitative representation of an utterance (viz., the matrix of spectral energies) is not only computationally untenable and conceptually difficult but also perhaps unnecessary. One would expect that the high physical and statistical correlations among the energies should imply redundancy. The limited number of replications available would, moreover, impose a mathematical constraint on usable dimensionality. For all these reasons, sum-

marization is necessary. The choices for summary statistics are legion and the consequences important. Various schemes for condensing the information in terms of manageably low-dimensional statistics were studied.

Table II shows a list of some of the types of information summaries that were investigated.

For instance, summarizing via the time margin means that one considers the energies (normalized) collapsed on to the time scale alone without any frequency breakdown. Similarly, frequency margin means that the energies (normalized) are summed over all the time intervals, thus eliminating the information about time variation of the spectrum. Looking at frequency slices implies the consideration of the energy distributions in each of the frequency channels. Within each of these ways of looking at the data, several alternate methods were investigated for summarizing the information. For instance, in studying the time margin both the energies themselves as well as characterizations of their distribution across time in terms of certain low-order moments (mean, standard deviation, etc.) were investigated. The distribution of energy within a frequency slice, however, was typified either by the deviations of its two tertiles (i.e., time values which divide the energy distribution into three equal parts) from the marginal time median or by the inter-tertile distance. These two time-dependent characterizations are origin invariant. (See Becker, et al.,[12] for more details concerning the reduction and analyses of first set of data.)

One of the important summaries, from the standpoint of performance in identification procedures, turns out to be the frequency margin normalized energies. This led to a 17-dimensional representation with the first body of data and a 20-dimensional representation in the

TABLE II—SUMMARIZATIONS OF DATA

(*i*) TIME MARGIN
- (*a*) Moments
- (*b*) Energies (normalized)

(*ii*) FREQUENCY SLICES
- (*a*) Tertile deviations from marginal median
- (*b*) Inter-tertile ranges

(*iii*) FREQUENCY MARGIN
- (*a*) Power spectral estimates derived from energies
- (*b*) Energies (normalized)

(*iv*) TIME × FREQUENCY
- (*a*) Moments
- (*b*) Variously grouped normalized energies

(*v*) VARIOUS COMBINATIONS OF INPUTS

second set. To illustrate how this summary representation may look, Figs. 2a and b each show the normalized energies in the frequency margin for all the utterances of a word by a specific talker. Figure 2a is for one speaker and Fig. 2b is for another. Qualitative and quantitative differences between the two speakers are evident as viewed against the relative cohesiveness of the different utterances within a speaker.

IV. SPATIAL REPRESENTATION AND CHOICE OF METRICS

Each scheme for summarizing the basic data leads to a set of input statistics whose values for each utterance yield a vector corresponding to that utterance. The analysis involved designating certain of the utterances from each speaker as unknown and treating the remaining utterances as the reference set of known utterances to be used for purposes of statistical estimation, etc., of the features of the reference population.

Thus, as shown in Table III, corresponding to the uth reference utterance (i.e., the talker is known) of a specific word by the ith talker, one would have a p-dimensional row vector of input statistics,

$$\mathbf{Y}'_{iu} = (y_{i1u}, y_{i2u}, \cdots, y_{ipu}); \quad i = 1, 2, \cdots, k, \quad u = 1, 2, \cdots, n_i,$$

where the jth element of the vector, y_{iju}, is the value of the jth input statistic for the uth utterance of the ith talker. There are k talkers in all and n_i known utterances from the ith talker. The n_i known utterances of the ith talker may then be used, as shown in Table III, to obtain the p-dimensional centroid, $\bar{\mathbf{Y}}'_i$, and the $p \times p$ covariance matrix, $\mathbf{S}_i$, for the ith talker.

Corresponding to an unknown utterance (i.e., the talker is unknown and is to be identified), which is known only to be an utterance of some one of the talkers in the study, one would similarly have a p-dimensional representation, shown in Table III as

$$\mathbf{Z}' = (z_1, z_2, \cdots, z_p).$$

Also shown in Table III are the overall centroid $\bar{\mathbf{Y}}'$ and two matrices $\mathbf{B}$ and $\mathbf{W}$. $\mathbf{B}$ is a measure of the dispersion of the speaker centroids in p-space and is called the between-talkers covariance matrix. $\mathbf{W}$ is a pooled measure of dispersion of the replicate known utterances around the talker centroids and is called the within-talkers covariance matrix.

If a metric or distance measure were defined in the p-dimensional space of the input statistics, then one could calculate the distance of the unknown, viz. $\mathbf{Z}'$, from each of the centroids, viz. $\bar{\mathbf{Y}}'_i$'s, of the

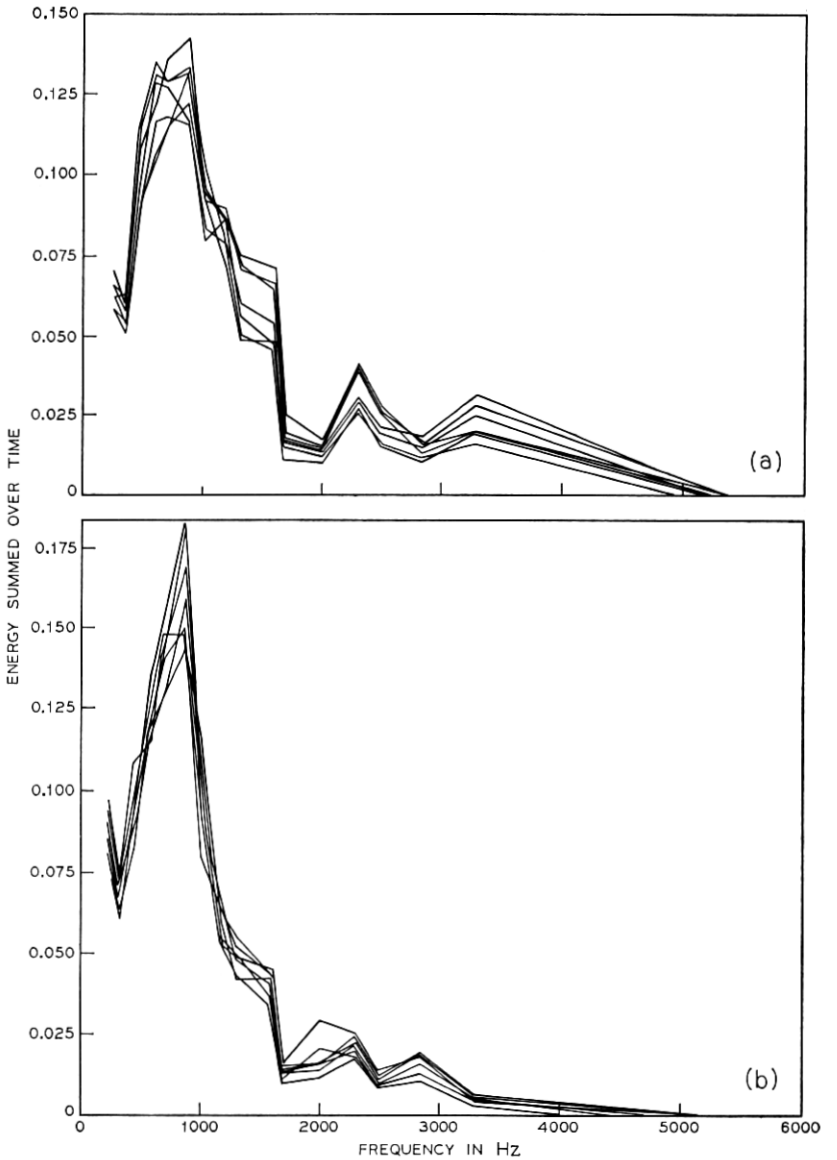

Fig. 2a—Frequency margin energy versus frequency (normalized data).
Fig. 2b—Frequency margin energy versus frequency (normalized data).

different talkers, and then use these distances to assign the unknown to one of the talkers.

All the measures of squared distance used in our work were positive semidefinite quadratic forms, a class whose typical member may be algebraically defined as shown in item (0) of Table IV. This class includes not only the familiar unweighted Euclidean squared distance ($\mathbf{M} = \mathbf{I}$) and the weighted Euclidean squared distance, which makes allowances for unequal variances of the different variables, but also measures of squared distance which allow for correlations among the variables. Figure 3, dealing with the case of two variables, shows an appropriate manner of measuring squared distance when the correlation is positive. According to such an elliptical measure of squared distance, points like A_1 and B_1 which lie on the same ellipse are considered to be the same distance away from the center C of the ellipse, whereas points like A_1, A_2 and A_3 which lie on the different ellipses numbered 1, 2 and 3 are considered to be at increasing distances away from C. The way to reflect this choice formally in the definition of squared distance is to use for $\mathbf{M}$ the inverse of an estimate of the covariance matrix of the variables.

Table IV also shows three specializations of the matrix $\mathbf{M}$ that lead to three squared distance measures D_1, D_2 and D_3 shown, respectively, as equations (1), (2) and (3).

The choice of $\mathbf{M}$ that leads to D_1 uses each talker's individual covariance matrix in measuring the distance of the unknown to that talker's

TABLE III—NOTATION AND ESTIMATES FOR REFERENCE SETS

(1) $\mathbf{Y}'_{iu} = (y_{i1u}, y_{i2u}, \cdots, y_{ipu})$; $i = 1, \cdots, k, u = 1, 2, \cdots, n_i$.

(2) $\bar{\mathbf{Y}}'_i = \frac{1}{n_i} \sum_{u=1}^{n_i} \mathbf{Y}'_{iu}$; $\mathbf{S}_i = \frac{1}{(n_i - 1)} \sum_{u=1}^{n_i} \{(\mathbf{Y}_{iu} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{iu} - \bar{\mathbf{Y}}_i)'\}$; $i = 1, \cdots, k$.

(3) $\mathbf{Z}' = (z_1, z_2, \cdots, z_p)$.

(4) $\bar{\mathbf{Y}}' = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{\mathbf{Y}}'_i$, where $n = \sum_{i=1}^{k} n_i$,

$$\mathbf{B} = \frac{1}{(k-1)} \sum_{i=1}^{k} n_i \{(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})'\},$$

$$\mathbf{W} = \frac{1}{(n-k)} \sum_{i=1}^{k} (n_i - 1)\mathbf{S}_i .$$

TABLE IV—METRICS

(0) $D(i) = (\mathbf{Z} - \bar{\mathbf{Y}}_i)'\mathbf{M}(\mathbf{Z} - \bar{\mathbf{Y}}_i); \quad i = 1, 2, \ldots, k.$
$\mathbf{M}$ is p.s.d. so that $D(i) \geqq 0.$

(1) $\mathbf{M} = \mathbf{S}_i^{-1}; \quad D_1(i) = (\mathbf{Z} - \bar{\mathbf{Y}}_i)'\mathbf{S}_i^{-1}(\mathbf{Z} - \bar{\mathbf{Y}}_i); \quad i = 1, 2, \ldots, k.$

(2) $\mathbf{M} = \mathbf{A}_r\mathbf{A}_r'$, where $\mathbf{A}_r$ is $(p \times r)$ with r eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ for columns $(r = 1, 2, \ldots, t)$;
$D_2(i) = (\mathbf{Z} - \bar{\mathbf{Y}}_i)'\mathbf{A}_r\mathbf{A}_r'(\mathbf{Z} - \bar{\mathbf{Y}}_i); \quad i = 1, 2, \ldots, k.$

(3) $\mathbf{M} = \mathbf{W}^{-1}; \quad D_3(i) = (\mathbf{Z} - \bar{\mathbf{Y}}_i)'\mathbf{W}^{-1}(\mathbf{Z} - \bar{\mathbf{Y}}_i); \quad i = 1, 2, \ldots, k.$

centroid. This choice of $\mathbf{M}$ implies that the covariance matrix for each talker be nonsingular. This in general requires that the number of known utterances for every talker is at least one more than the number of input statistics. If p were large, therefore, in order to use D_1 one would require a large number of known utterances (replications) for each talker.

Also, this choice for $\mathbf{M}$ means that $\mathbf{M}$ changes from talker to talker with a consequent increase in computational time and effort. The hope is that there will be a pay-off in terms of efficiency to be gained from using a distance measure that is sensitive not only to the location (centroid) features of a talker but also to his individual covariance pattern. The use of this distance measure is thus particularly appropriate when different speakers do not have the same covariance matrix for their replicate utterances.

A second choice for $\mathbf{M}$, leading to D_2 in equation (2) of Table IV, is provided by the so-called discriminant analysis approach of multivariate statistical analysis. Here $\mathbf{M}$ is the product of a matrix by its transpose and the columns of the matrix are eigenvectors obtained from a discriminant analysis. The discriminant analysis attempts to reduce the number of dimensions in the space in which distances are measured by selecting a subspace which in a sense contains the most important information for discrimination purposes.

Broadly speaking, statistical discriminant analysis is concerned, in part, with finding a representation of the data from several prespecifiable groups (talkers) in terms of coordinates which separate the group centroids maximally relative to the variation within groups. Specifically, as shown in Table V, if $y_1, \cdots, y_p$ denote the variables in the initial p-dimensional representation of an utterance, then at the first stage one considers a linear combination, x, of the original coordinates. A one-way analysis of variance for this derived variable would lead to the F-ratio shown, where $\mathbf{B}$ and $\mathbf{W}$ were defined earlier. One can now

specify maximization of this F as a criterion for choosing the coefficients $(a_1, \cdots, a_p)$ in the linear combination. The required solution is to choose, for $\mathbf{a}$, the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. Having chosen one linear combination, a second different from the first may be sought so that its F-ratio will be maximized and so on. This method of seeking a linear transformation involves the eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$. There will be t positive eigenvalues, $c_1, c_2, \cdots, c_t$, in general, where t is the smaller of p and $(k-1)$. This is a consequence of the fact that if k (the number of talkers) is less than p (the dimensionality of the input data), then the k talker centroids are contained in a $(k-1)$-dimensional hyperplane. At any rate, one can use each eigenvector that corresponds with one of the nonzero eigenvalues to obtain the new coordinates $x_1, x_2, \cdots$. The space of x's may be called the *discriminant space* and the coordinates, $x_1, x_2, \cdots$ called *discriminant coordinates*, or CRIMCOORDS.

A geometrical interpretation of the discriminant analysis for the case of two variables is shown in Fig. 4. Centroids of the known talkers are shown in Fig. 4a surrounded by an ellipse indicating the distance measure appropriate to $\mathbf{W}$, the pooled within-talkers covariance matrix. The discriminant measure of distance is equivalent to: (*i*) transforming the space of Fig. 4a to one in which the ellipses have become circles by suitably compressing or expanding and reorienting the various coordinates—this space, with axes y_1^* and y_2^*, is shown in Fig. 4b; (*ii*) rotating the coordinates y_1^* and y_2^* in Fig. 4b so that the speaker centroids have maximum mean square separation in the direction of the first coordinate (x_1), next smaller separation in the direction of the

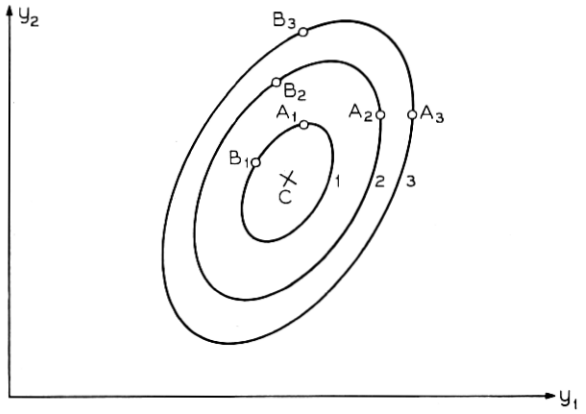

Fig. 3—Elliptical measure of squared distance.

TABLE V—STATISTICAL DISCRIMINANT ANALYSIS

(1) $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$.

(2) $x = a_1y_1 + a_2y_2 + \ldots + a_py_p = \mathbf{a}'\mathbf{y}$.

One-way Analysis of Variance

	D.F.	M.S.
Between Talkers	$k - 1$	$\mathbf{a}'\mathbf{B}\mathbf{a}$,
Within Talkers	$n - k$	$\mathbf{a}'\mathbf{W}\mathbf{a}$,

$F_{\mathbf{a}} = \mathbf{a}'\mathbf{B}\mathbf{a}/\mathbf{a}'\mathbf{W}\mathbf{a}$.

(3) Choose $\mathbf{a}$ so as to maximize $F_{\mathbf{a}}$,
Solution: $\mathbf{a} = \mathbf{a}_1$ the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to its largest eigenvalue.

(4) $c_1 \geqq c_2 \geqq \ldots \geqq c_t > 0, \quad t = \min(p, k - 1)$.
$\updownarrow \quad \updownarrow \qquad \updownarrow$
$\mathbf{a}_1 \quad \mathbf{a}_2 \quad \ldots \quad \mathbf{a}_t$

(5) $x_1 = \mathbf{a}_1'\mathbf{y}, \quad x_2 = \mathbf{a}_2'\mathbf{y}, \quad \ldots$.

second coordinate (x_2), etc; and (*iii*) measuring simple Euclidean distances in the space thus derived. (Note: With more than two variables, one may decide to use only the subspace formed from the first r coordinates of the discriminant space.) Discriminant analysis makes more intuitive sense if the individual talkers all have similar covariance matrices for their repeated utterances (so that each is similar to the pooled covariance matrix) than if they have widely differing covariance matrices.

The measure of squared distance, D_2, is just that Euclidean squared distance measure in the space of the first $r(\leqq t)$ CRIMCOORDS. While $\mathbf{M}$, chosen thus, does not change across talkers, yet it does depend on r, the number of eigenvectors to be used in the discriminant analysis approach. This use of an increasing number of the eigenvectors implies diminishing returns and may not necessarily improve the identification. By trial and error, a satisfactory value of r when the frequency margin energies were used as the initial variables was found to be 5 in the first body of data and 10 in the second set.

A third choice for $\mathbf{M}$, leading to the squared distance measure D_3 of equation (3) in Table IV is obtained by taking $\mathbf{M}$ equal to the inverse of the pooled within-talkers covariance matrix $\mathbf{W}$, defined earlier. This choice of $\mathbf{M}$, which requires $\mathbf{W}$ to be nonsingular, is in general possible whenever the number of input statistics, p, does not exceed the total number of known utterances of all talkers minus the number of talkers. This constraint on p (or, equivalently, on the number of known utterances) is far less restricting than the constraint on p imposed by the choice of $\mathbf{M}$ that leads to D_1.

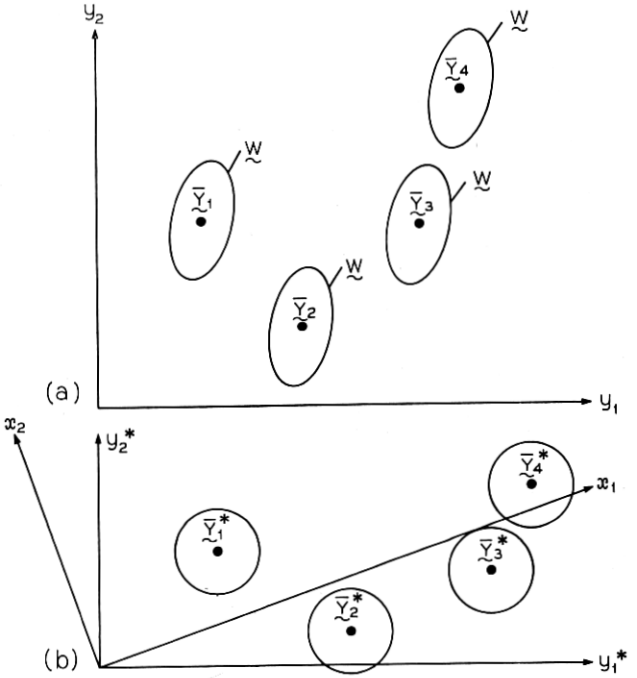

Fig. 4—Sketch to indicate geometrical interpretation of discriminant analysis.

There are certain relationships and equivalences amongst these three measures of squared distance. D_1 and D_3 are similar ellipsoidal measures (in the sense of Fig. 3) of squared distance and are identical if the talkers all have the same covariance matrix for their repeated utterances. Using $r = t$, the "maximum" number of eigenvectors from the eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$ would make D_2 entirely equivalent to D_3. Furthermore, D_3 is entirely equivalent to a discriminant analysis approach with a pairwise comparison of the distances of the unknown utterance from the centroids of the talkers considered in all possible pairs.

Other metrics, which were approximations to D_1, D_2 and D_3 in varying degrees of appropriateness and simplicity, were also investigated, but the results to be presented here are confined to these three measures.

V. IDENTIFICATION SCHEMES

5.1 *Classification Procedures for Small Populations of Talkers*

The distance measures are used for assigning an unknown utterance

to one of the speakers. For relatively small populations of speakers, one can compute the distance of an unknown from each of the speaker centroids, using any measure of distance, and then assign the unknown to the speaker whose centroid is closest. The empirical criterion used for evaluating the operating characteristics of any combination of input data and distance measure was the percent of unknowns that were correctly identified.

Table VI, based on the findings of the analysis of the first body of data, shows a summary table of percent correctly identified, for some of the different data summaries, when used with the three squared

TABLE VI—SUMMARY OF AVERAGE PERCENT CORRECT FOR VARIOUS TALKER IDENTIFICATION TECHNIQUES USED WITH FIRST BODY OF DATA

Input Statistics	Distances		
	D_1	D_2	D_3
Time Margin			
Moments	34†	55	62*
Energies	—	30	34
Frequency Slices			
Deviation of Tertiles from Median (TER)	—	82	—
Inter-Tertile Range (ITR)	—	67	—
Frequency Margin			
Power Spectral Estimates	—	73	—
Energies	—	91	97*
Time × Frequency Groupings			
μ_f, σ_f^2, σ_t^2, σ_{ft}, N_{50}	30†	—	—
2 × 17	—	86	100†
2 × 7	—	83	90†
2 × 3	—	57	70†
3 × 2	—	47	50†
16 × 2	—	40	40†
Combinations of Inputs			
Frequency Energies + Time Moments	—	91	97*
Frequency Energies + Time Energies	—	83	90
Frequency Energies + ITR	—	88	—
Frequency Energies + TER EIG	—	—	93
Frequency Energies + ITR EIG	—	—	94
Combinations of Eigenvector Transforms (EIG)			
Frequency Energies EIG + TER EIG	—	87	—
Frequency Energies EIG + ITR EIG	—	94	93
Combinations of Words	—	98	—

* All utterances of each word used as unknowns.
† Only used word 1.

distance measures D_1, D_2 and D_3. The dimensionality of several of the inputs was too high, relative to the number of replicated utterances available per talker, so that D_1 could not be used with these inputs. Dashes in the table denote such cases and others, wherein the particular combinations of input and distance measure were not studied.

As far as the distance measures are concerned, D_3 appears to perform best. However, D_2 because of the reduced dimensionality associated with it, and D_1 because of its sensitivity to variations in the dispersion characteristics of the talkers, may be more appropriate and efficient for some uses and should not necessarily be discarded.

The general conclusion to be drawn from the various attempts to summarize the original data in terms of input statistics appears to be that, by and large, frequency information is more important than time information. This is evident from the low percentages of correct identification (30 percent and 34 percent) for the time margin energies, at one extreme, and the high ones of 91 percent and 97 percent for the frequency margin energies, at the other extreme. The results for the various time-by-frequency groupings also suggest the same conclusion. As the time structure increases, the percent correctly identified decreases (cf. also Pruzansky & Mathews[6]). Certain schemes for using the time information as an adjunct to frequency information, however, do seem promising. Thus, using D_2, one achieves 91 percent correct identification on the basis of frequency margin energies alone, whereas an increase to 94 percent is possible by augmenting the frequency margin information with certain kinds of time information from the frequency slices.

A general indication of the results shown is that significant improvement may be achieved by using appropriate statistical methods for the choice of both the input statistics and the distance measures. Thus, for example, using the normalized energies in the frequency margin as a summary of the data, one could go from 91 percent correct identification to 97 percent by using D_3 instead of D_2, thus achieving a reduction in error rate by a factor of 3.

5.2 *Strategies for Large Populations*

Encouraged by the results of the first analysis, we undertook the collection of the second body of data. In order to simulate more practical situations, the recording conditions were not as strictly controlled this time and we also decided to increase the number of speakers and to prune the number of replications per speaker. The increase in the size of the speaker population introduces an immediate challenge for

the analysis. Even with the aid of modern high-speed computers, the effort required for comparing the distances of the unknown from every talker centroid would become prohibitive with a very large number of talkers. For this case of a large number of talkers, which is of great practical interest, one has to develop a method for limiting the number of contenders for assignment of an unknown.

The approach to be described next in broad outline is simple and seems to be effective for accomplishing the task with the 172 speakers involved in the second body of data. For the present discussion, only the normalized frequency margins (viz., the 20-dimensional input) will be used as the representation of any utterance.

The basic idea is to use the first few CRIMCOORDS for restricting the set of speakers with whom an unknown is to be compared. Before describing the essential nature of the approach, it is perhaps in order to comment on some properties of CRIMCOORDS as related to the present problem.

Firstly, there is the question of interpretability. Figure 5 shows a pictorial representation of the five eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ that correspond to the first five CRIMCOORDS. The lengths of the bars correspond to the magnitudes of the elements of a specific eigenvector and the orientations correspond to their signs. The first CRIMCOORD, which seems to be largely a difference between the energies in the two lowest frequency bands, appears to be reflecting a difference between male and female glottal fundamental frequencies for the early vowel part of the word *one*, which was the word that gave rise to the set of eigenvectors in Fig. 5. The first CRIMCOORD does indeed efficiently separate male and female speakers in the study. Unfortunately, the second and later CRIMCOORDS do not seem to have as easy an interpretation, perhaps due to the mathematical constraints imposed on the eigenvectors at the later stages.

The linear transformation to CRIMCOORDS is dependent on only the reference set of known utterances. It is perhaps interesting to inquire about the validity of the pooling of the separate within-talker dispersion matrices to obtain the pooled estimate $\mathbf{W}$ of variation among the replicate utterances. The pooling also underlies the justification for using unweighted Euclidean squared distance (viz., D_2) in the CRIMCOORDS space. An internal comparisons statistical technique was developed (cf. R. Gnanadesikan and E. T. Lee[13]) for assessing the comparability of the individual talker covariance matrices in terms of certain measures of their sizes. This method of assessment, when used with the frequency margin energies, suggested

that it is not unreasonable to pool the speaker dispersions to obtain **W**.

A third issue concerning CRIMCOORDS is their stability as they depend on the words and on the speakers. Based on an empirical investigation using the second body of data, they do seem to be word dependent but fairly stable with respect to the speakers, in that they do not seem to change substantially once they are based on the known utterances from about 80 speakers.

Returning now to the question of using the first few CRIMCOORDS for limiting the contenders for an unknown, one can look at a representation of the knowns in the space of say the first two CRIMCOORDS. As shown in Fig. 6, with only the ten speakers in the first study, some talkers are clearly separated (e.g., talkers 2, 4, 7 and 10) while others are clustered (e.g., talker 8 and 9) even in this two-

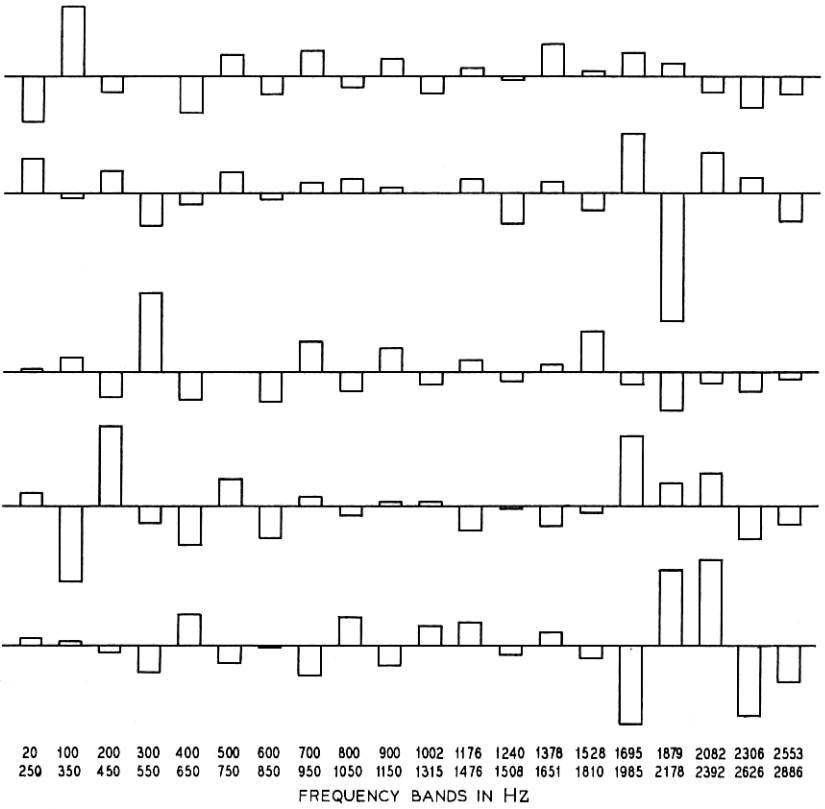

Fig. 5—First five eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$.

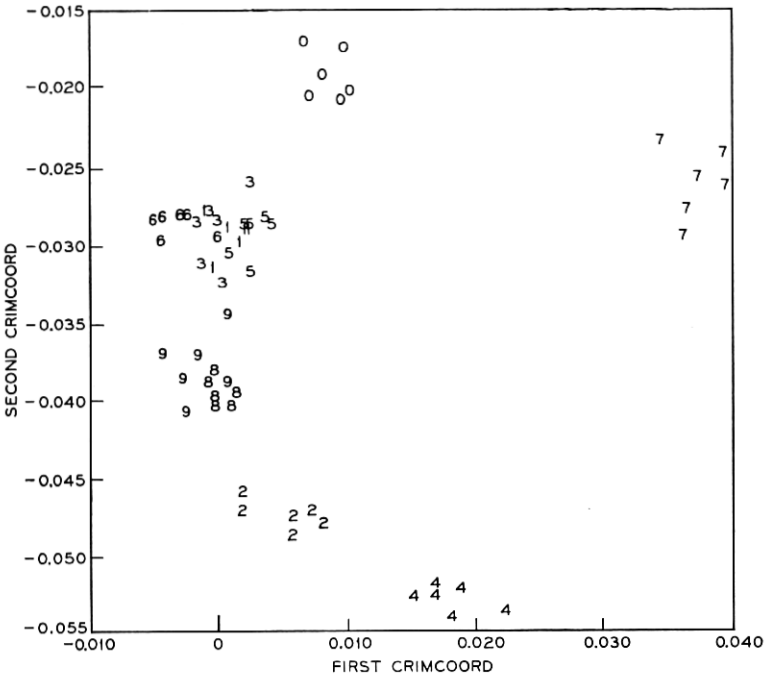

Fig. 6—Representation of utterances of ten speakers in space of first two CRIMCOORDS.

dimensional representation. However, with the 172 centroids of the talkers in the second set of data, one gets the configuration in Fig. 7a. There appear to be no obvious clusters here in the space of the first two CRIMCOORDS.

In this case, one approach is to divide the two-dimensional CRIMCOORDS space arbitrarily up into boxes as a first step. Figure 7b shows a division of the space into forty boxes which was accomplished by arbitrarily specifying nine quantiles (or percentage points) of the distribution of centroids along the first CRIMCOORD and three quantiles of the distribution along the second CRIMCOORD. Next, one determines in which of these boxes an unknown under consideration for assignment falls (cf. Fig. 7c) and then one can compare the unknown with all the speakers who fall in the same or a few nearby boxes (cf. Fig. 7d), discarding the speakers who are far removed. In the particular example used for Figs. 7a-d, while the 0 denotes the unknown, the x (cf. Fig. 7d) corresponds to the centroid

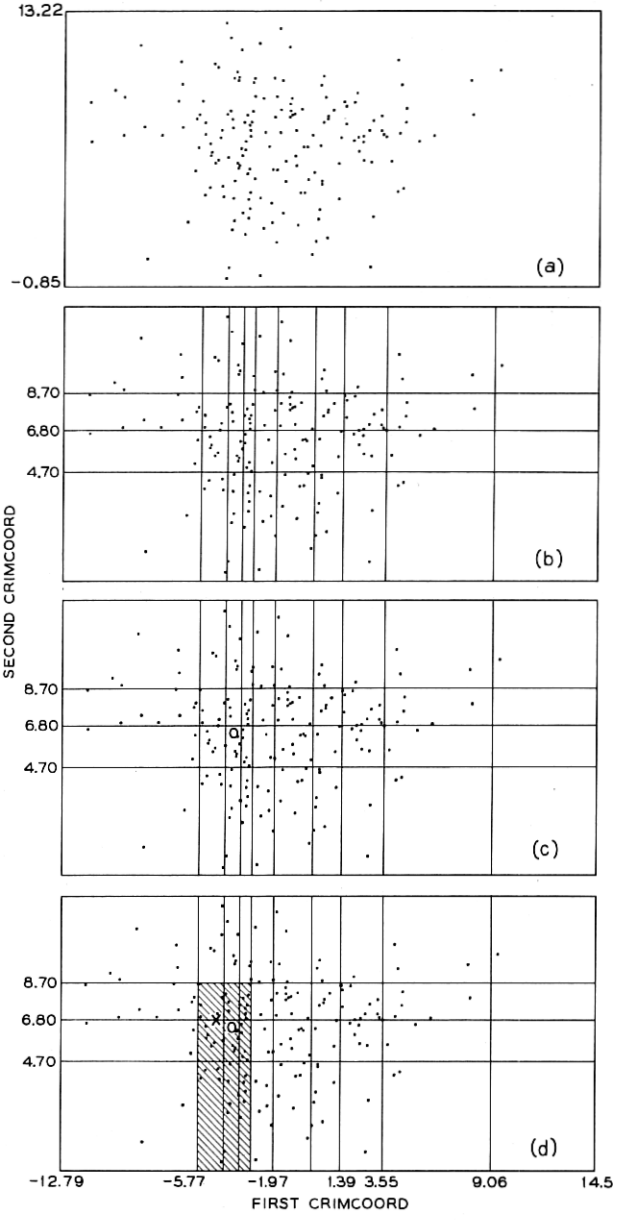

Fig. 7a—Centroids of 172 talkers in CRIMCOORDS space.
Fig. 7b—Division of CRIMCOORDS space into boxes.
Fig. 7c—Positioning of unknown (0).
Fig. 7d—Boxes searched initially (unknown, 0; corresponding talker centroid, X).

of the speaker from whom the unknown arose. While the x is not in the same box as the 0, it is in a neighboring box which is included for identification purposes.

The comparison of the unknown with the speakers in the neighboring boxes is made by calculating distances not just in the space of the first two CRIMCOORDS but by including additional CRIMCOORDS (e.g., using five or ten CRIMCOORDS in all). Based on the magnitude of the distance to the closest speaker and on the ratio of the second smallest distance to the smallest, a decision is made whether identifying the unknown with the closest speaker is safe or suspect. Statistical benchmarks for comparing observed values of quantities, such as the smallest distance or the ratio of the second smallest to the smallest distance, are obtained from "null" distributions (i.e., distributions of these quantities when a correct identification is made) generated from the data on hand. Since we are dealing with a situation in which there are sufficient data under "null" conditions (i.e., successful identification) one can obtain adequate estimates of the statistical distributions to enable reasonable assessments of the magnitudes of observed distances (or ratios) and decide whether they are small or large.

At any rate, either if an identification is suspect or if an insufficient number of comparisons have been made, the process enlarges the population of contenders by considering the speakers in additional boxes nearby. As soon as a safe identification is made, no further loops are made to add more contenders. After all the speakers have been exhausted, if the identification in terms of the closest speaker is still suspect, then the process terminates by identifying the speaker as the closest one, despite the weakness of the evidence. For the illustrative example in Fig. 7, this method led to a safe and correct identification.

Figure 8 shows a simple flow-chart of the steps involved in the above process for identifying an unknown by a preliminary limiting of the number of contenders. On the left, in Fig. 8, are shown the steps in the initial processing of the reference utterances leading to (*i*) a determination of the CRIMCOORDS, (*ii*) a representation of the speaker centroids in CRIMCOORDS space, and (*iii*) a specification of the boxes or cells in the space of the first few (e.g., two) CRIMCOORDS. On the right, in Fig. 8, is shown the identification process for an unknown utterance. From the representation of the unknown in CRIMCOORDS space, one finds which box the unknown falls into and retains for comparison all speakers whose centroids fall in a cer-

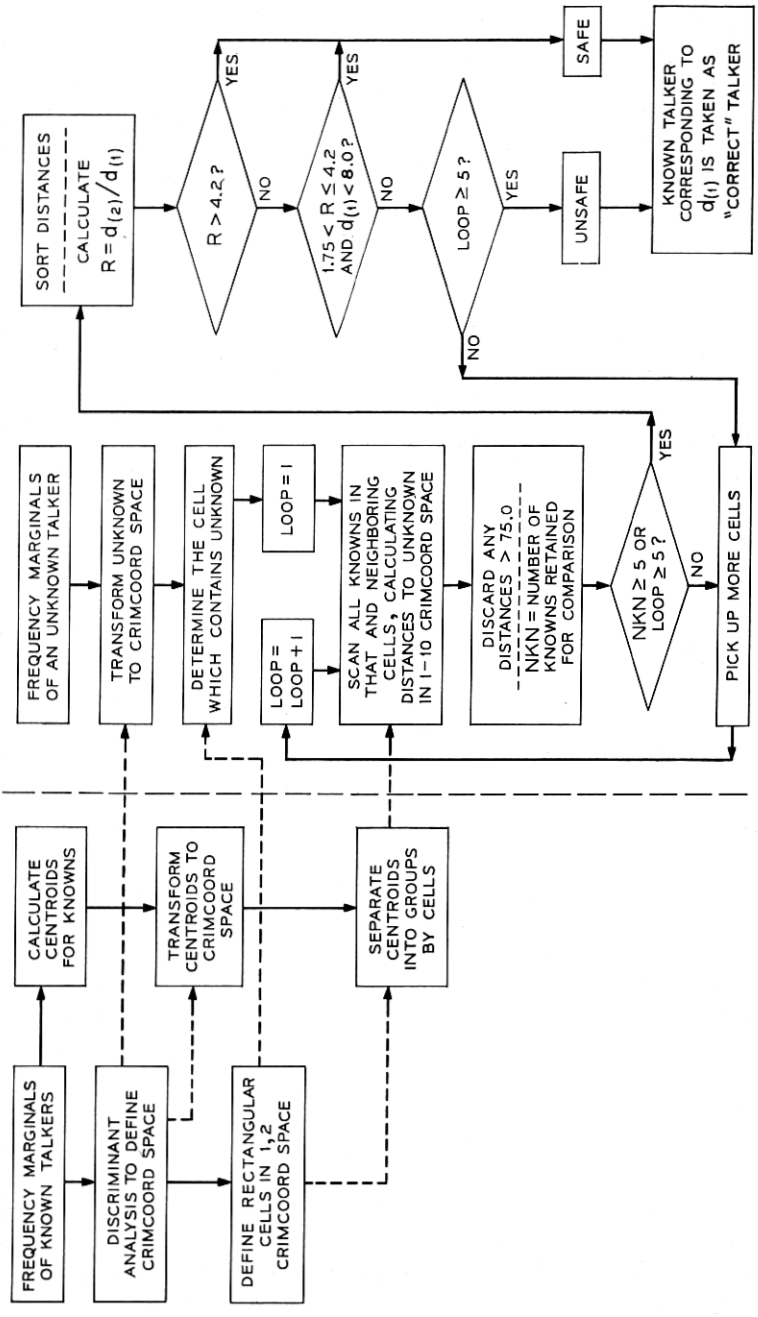

Fig. 8—Flowchart for strategic identification.

tain number of nearby boxes, while discarding all speakers who are farther away than a cut-off distance. If an adequate number of speakers have been retained for comparing the unknown against, then one computes the distance of the unknown from each of the retained speakers in an enlarged (5- to 10-dimensional) CRIMCOORDS space. If the ratio of the second smallest to the smallest is large enough (> 4.2), then the identification of the unknown with the closest speaker is assumed to be safe. Also, if this ratio is moderately large ($\leqq 4.2$ but > 1.75) and the smallest distance is small enough (< 8.0), the identification is deemed safe. If not, it is deemed suspect and more speakers in additional boxes are included for comparison with the unknown. Furthermore, if the number of speakers compared with an unknown is not large enough, then also one adds more speakers by considering additional adjacent boxes.

An efficient set of computer programs implementing this process is in use (cf. K. W. Wachter[14]). The programs provide for flexible specification of many of the parameters involved (e.g., number and size of the boxes, cut-off values for comparing the smallest distance or the ratio of the second smallest to the smallest, etc.).

Using a single word for identifying talkers in the second set of data, the above strategy yielded 81 percent correct identifications, i.e., for 81 percent of the unknowns the first most likely match was correct. In fact, if one counted the percent of times that the correct speaker was either the closest or second closest then one obtains 90 percent. The comparable percentages to these figures of 81 percent and 90 percent are 84 percent and 93 percent when one performs an exhaustive check of the unknown against every speaker. The computational cost for the exhaustive comparisons in this example involving 172 talkers, is about 70 to 90 percent more than that involved in the strategy based on preliminary limitation of the number of contenders for an unknown. The difference in computational cost will, of course, vary as one changes the number and size of the boxes chosen and the other specifications involved in the method. Also, with much larger populations of speakers, the cost differential between the two approaches would be expected to increase substantially. In the present example, the computer cost per identification is approximately 1.4 cents for the scheme which initially limits the number of contenders for an unknown.

The percentages of correct identification appear to be improvable by utilizing the identification information in additional words. Thus, instead of using the single word, *one*, when one combines the information from the separate identification results for the two words, *one* and *two*,

the percent correct identification moves up to 94 percent from the earlier stated 81 percent. Hence, the preferable direction for large populations appears to be to stay with the type of strategy described here but to combine information from more than one word.

Not all methods for combining the information from two or more words will be equally successful; however, a scheme has been developed which appears to be promising and, in fact, led to the improvement from 81 percent to 94 percent in percent correct identifications. Figure 9 shows a flowchart of the procedure. If the identification on the basis of the first word is deemed safe, then no use is made of the second word. If, on the other hand, the identification using the first word is suspect, then one looks at the identification results for the second word. If the same speaker is the closest one to the unknown utterances in both words, then this is taken as a safe indication that he is in fact

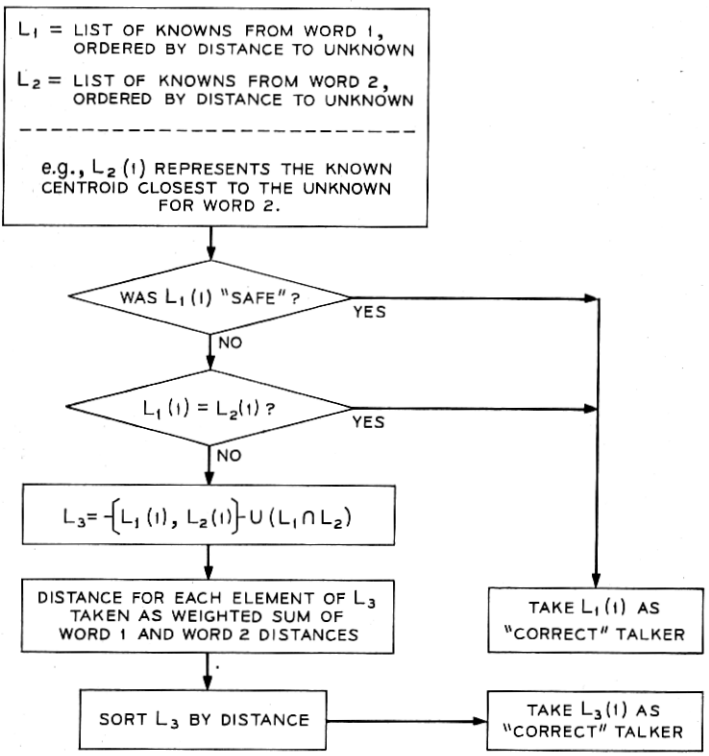

Fig. 9—Combination of word 1 and word 2 information into a single identification.

the speaker despite the weakness of the evidence for this in considering the first word by itself. If two different speakers turn out to be closest for the two words, then one considers these two speakers along with all speakers who appear concurrently on the lists of a certain number (e.g., 10) of closest speakers for both the words. For each of these speakers, one can compute a new squared distance by calculating a weighted sum of the squared distances of the speaker centroid from the unknown as obtained in the separate analyses of the two words. The weights for combining the squared distances could reflect the discrimination abilities of the two words. (Note: Implicit to this scheme is an assumption that the first word is more useful for discriminating talkers than the second.) The new squared distances may then be sorted and the identification would be made as the closest speaker in terms of this pooled measure of distance.

VI. DISCUSSION

The emphasis in the present report has been on the statistical facets of talker identification rather than on the acoustic significance of the results. The gratifying identification successes have been achieved, in fact, with relatively unsophisticated representations of the data (viz., energy-(time)-frequency analyses of whole words). We have already mentioned that interpretation of CRIMCOORDS beyond the first is elusive. At this juncture, we approach the relation between identification techniques and acoustic factors from the other end, asking what speech production theory and related experimental results have to say about augmenting or modifying our representations for future work.

In the course of replicating our second body of data, we encountered two factors which will be dealt with systematically in forthcoming studies. The first of these is inter-session variation within talkers. W. A. Hargreaves and J. A. Starkweather[8] as well as J. E. Luck[2] have reported that this effect is so strong as to render identification or verification significantly poorer when reference and test recordings are made at widely separated times. We are now collecting new recordings from talkers who return at scheduled times on different days, having found evidence of a sessions effect for those few talkers who returned to our unattended booth voluntarily. The other factor was "circuit variability." The collection of data from the unattended booth involved a second set of 172 speakers in addition to the set of 172 used in the earlier discussion in this paper. The samples from the two sets of speakers were, however, processed separately at two different times,

and the two sets of data could not be combined because of unrecoverable variations in the "circuit" (the process involved in going from the physical utterance to its digital representation, especially the behavior of the analog filter bank). Our approach to this problem is two-pronged: (*i*) We are using elaborate control and calibration procedures in making the new recordings, and obtaining our spectral energy representations directly from Fast Fourier Transforms performed on a digital computer. (*ii*) We shall deliberately vary the circuit (by making some telephone recordings simultaneously with the high-quality recordings and by passing them through switched connections) to learn how to make identification robust in the face of circuit variations.

Another major concern is with the use of information from different words and the number of replications required from each speaker for constituting the reference set of utterances. Both Pruzansky[5] and J. W. Glenn and N. Kleiner[4] found improvement when they combined utterances in the simplest way, at the level of spectral analysis. It is worth noting that the work of S. K. Das and W. S. Mohn,[3] the most successful of the verification studies, used separate analyses of ten different segments (of a fluently uttered phrase), while the other two[1,2] used only one or two. In connection with the question of number of replications, no statement about the ultimate resolving power of automatic voice identification schemes can be made until the relationship between performance and number-of-samples-known-to-be-from-one-talker is better understood.

Perhaps the gravest issue is whether measures thought to have acoustic or speech-theoretic significance should supplant or supplement raw spectral energy representations of the utterances. Examples of such measures, used with success by J. J. Wolf,[7] are glottal fundamental frequency, nasal resonance frequency, vowel formant frequencies, and voice-onset time in voiced stop-consonants. The basic argument for this approach is that a talker's uniqueness lies perhaps in the shape of his vocal apparatus, and we should use measures sensitive to that shape. Glenn and Kleiner,[4] taking an extreme position, maintain that nasal sounds are ideal because the apparatus is stationary during the time the oral passage is occluded and radiation is chiefly from the nose. Das and Mohn[3] worked with features chosen to be relevant to their acoustic segmentation, but no report on the relative merits of their many (405 in all) features is available.

One might well have reservations about using features which are significant in speech synthesis to perform talker recognition, because what is signal in the former problem may be noise in the latter. How-

ever, it is worth noting that some recent analysis-synthesis schemes for speech transmission (e.g., R. W. Schafer and L. R. Rabiner[15]) are reported to be capable of producing speech that sounds strikingly like the original, though based on only a few parameters per time sample. Although such resynthesized speech may not be capable of reflecting subtle variations of the shapes of intra-cranial cavities, it typically quite accurately reflects two important characteristics of the individual: his average "pitch" (glottal fundamental frequency) and the temporal pattern of changing glottal and formant frequencies.

"Pitch" is clearly a vital talker cue. The present work evolved a CRIMCOORD devoted to this characteristic even though our input data did not include an explicit measure of it. Wolf[7] found pitch to be the most important of his features, and B. S. Atal[16] based an entire small-population recognition scheme on it. We are investigating economical means of including such a measure in our future work.

As for the temporal aspect of utterances, it should be recalled that certain representations of time information did augment frequency information to advantage in the analysis of our first set of data. In fact, Gnanadesikan and Wilk[17] found that transforming the energies logarithmically (the common "decibel" transformation) improved the "additivity" of the frequency and time effects. Finally, it is worth learning how to use both time and "pitch" information for one very important reason: neither is unduly affected by common variations among speech transmission circuits, whereas the raw spectrum is very vulnerable.

The search for efficient representations of the speech signal is now at a choice-point. In one direction lies the extraction of features based on speech production theory, with its current high cost but the promise of robustness in the face of circuit variation and perhaps other sources of interference. In the other direction is the development of procedures for correcting obtained spectra to compensate for distortions due to the circuit, with a non-negligible cost and unknown ultimate technical feasibility. In the last analysis, technical and economic considerations will determine which of these types of representation will play the major role in practical talker identification techniques.

REFERENCES

1. Li, K. P., Dammann, J. E., and Chapman, W. D., "Experimental Studies in Speaker Verification, Using an Adaptive System," J. Acoust. Soc. Amer., *40*, No. 5 (November 1966), pp. 966–978.
2. Luck, J. E., "Automatic Speaker Verification Using Cepstral Measurements," J. Acoust. Soc. Amer., *46*, No. 4 (October 1969), pp. 1026–1032.

3. Das, S. K., and Mohn, W. S., "Pattern Recognition in Speaker Verification," AFIPS Conf. Proc., Fall Joint Computer Conference, *35* (1969), pp. 721–732.
4. Glenn, J. W., and Kleiner, N., "Speaker Identification Based on Nasal Phonation," J. Acoust. Soc. Amer., *43,* No. 2 (February 1968), pp. 368–372.
5. Pruzansky, S., "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Amer., *35,* No. 3 (March 1963), pp. 354–358.
6. Pruzansky, S., and Mathews, M. V., "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Amer., *36,* No. 11 (November 1964), pp. 2041–2047.
7. Wolf, J. J., "Simulation of the Measurement Phase of an Automatic Speaker Recognition System," J. Acoust. Soc. Amer., *47,* No. 1 (January 1970), p. 83 (A).
8. Hargreaves, W. A., and Starkweather, J. A., "Recognition of Speaker Identity," Lang. & Speech, *6,* No. 2 (April–June 1963), pp. 63–67.
9. Smith, J. E. K., "Decision-Theoretic Speaker Recognizer," J. Acoust. Soc. Amer., *34,* No. 12 (December 1962), p. 1988 (A).
10. Ramishvili, G. S., "Automatic Voice Recognition," Engineering Cybernetics, (September–October 1966), pp. 84–90.
11. Gnanadesikan, R., and Wilk, M. B., "Data Analytic Methods in Multivariate Statistical Analysis," *Multivariate Analysis II,* (Editor, P. R. Krishnaiah), New York: Academic Press, pp. 593–638. (especially Section 4.)
12. Becker, M. H., Gnanadesikan, R., Mathews, M. V., Pinkham, R. S., Pruzansky, S., and Wilk, M. B., "Comparisons of Some Statistical Distance Measures for Talker Identification," Unpublished memo. [See also abstract in J. Acoust. Soc. Amer., *36,* No. 10 (October 1964), p. 1988.]
13. Gnanadesikan, R., and Lee, E. T., "Graphical Techniques for Internal Comparisons Amongst Equal Degree of Freedom Groupings in Multiresponse Experiments," Biometrika, *57,* No. 2 (August 1970), pp. 229–237.
14. Wachter, K. W., "Talker Recognition on Large Populations," Talk presented at the 78th Acoustical Soc. meetings in San Diego, November 1969. [See also abstract in J. Acoust. Soc. Amer., *47,* No. 1 (January 1970), p. 66.]
15. Schafer, R. W., and Rabiner, L. R., "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer., *47,* No. 2 (February 1970), pp. 634–648.
16. Atal, B. S., "Automatic Speaker Recognition Based on Pitch Contours," Polytechnic Institute of Brooklyn Doctoral Dissertation, June 1968.
17. Gnanadesikan, R., and Wilk, M. B., "Statistical Techniques for Effective Condensation of Talker Identification Data," Talk given at the European Regional Meetings of the Inst. of Math. Statist., September 1966. [See also abstract in Ann. Math. Stat., *37,* No. 5 (October 1966), p. 1415.]