Two Statistical Feature Evaluation Techniques Applied to Speaker Identification

WILLIAM S. MOHN, JR.

Abstract—The problem of identifying people based solely upon samples of their speech is viewed as a statistical pattern classification problem, emphasizing the portion of the process in which an informative, concise set of features is extracted from the speech signal. This work takes into account both the limited amount of data available in a real application and the statistical dependence among the various proposed features. In addition, the results of feature evaluation should apply to speakers not involved in the evaluation set.

Two specific feature evaluation and dimensionality reduction methods—analysis of variance and discriminant analysis—are investigated. It is shown experimentally that evaluation of features individually, ignoring statistical dependence, is of no benefit unless the original set of features was less dependent than is normally the case. On the other hand, a linear transformation, taking dependence into account, is shown to produce efficient dimensionality reduction only for the speakers for which it was designed.

Index Terms—Analysis of variance, discriminant analysis, eigenvector solution, feature evaluation, pattern recognition, speaker identification.

Introduction

FEATURE evaluation and dimensionality reduction are certainly not new areas of investigation in pattern recognition. Surprisingly few firm results are as yet available, however, for realistic problems. In the work reported here, the author has chosen two commonly used techniques, applied them to a specific problem, and objectively compared them with each other and with chance.

The specific problem used as a vehicle for this study is speaker identification. In general, it may be defined as follows. A population of k people, $S_1, S_2, \cdots, S_k$, has made available examples of their speech, speaker S_i providing r_i utterances, $U_{i1}, U_{i2}, \cdots, U_{ir_i}$. An utterance is defined as a speech signal of any convenient length, in this case a particular phrase. The source of each utterance U_{ij} is known, that is, each utterance is labeled. The identification task is that of attempting to determine correctly the speaker (S_i) of a new utterance not among the original set $U_{ij}, j=1, \cdots, r$, $i=1, \cdots, k$.

The specific experiments reported involved only ten male speakers, all uttering the same phrase. They were digitized using a filter-bank analyzer, and the feature extraction and decision experiments were implemented in software. Naturally, a number of factors, especially the number of speakers, strongly affected the absolute accuracy, but the area of interest here is relative effectiveness of various dimensionality reduction techniques of a large set of interdependent features used in a difficult pattern recognition problem.

Manuscript received November 16, 1971; revised March 26, 1971. A preliminary version of this paper was presented at the IEEE Symposium on Feature Extraction and Selection in Pattern Recognition, Argonne, Ill., October 5–7, 1970.

The author is with the IBM System Development Division Laboratory, Research Triangle Park, N. C. 27709.

Background

Analysis of Variance

There seem to be two ways to group various dimensionality reduction techniques. They are the "subsetting methods" and the "transformation methods." If one interprets the original features as being a set of coordinates in a multidimensional space, the subsetting methods reduce dimensionality by using some but not all of the original coordinates without change. The transformation methods are more general in that they measure an object along a new set of coordinates, often a linear combination of the original ones.

Analysis of variance is a method of subsetting [1], [2]. The most straightforward subset approach is to find a method of rank-ordering the original features as to "goodness." If one then wants to use only the "best" half of the original features, he may choose those in the top half of the list. The simplest way of measuring "goodness" is to study each feature independently and assign each a "goodness values." A better way would be to investigate each for "goodness" but also retain how much or how little new information is conveyed by the feature beyond that already contained in those features ranking above it.

Either of the above techniques produces a single rank-ordering of features that may be used no matter how great a reduction in dimensionality is needed. Of course, a single ranking will not allow one to select the optimum subset of features, even assuming that one must employ the original coordinates. The optimum subset of n features out of m is not usually the optimum subset of $n-1$ features with one added.

In order to develop the rank order, a method of measuring each feature's worth in isolation was needed. From the many measures that were available, the simple signal-to-noise, or F ratio of analysis of variance was chosen. In its simplest terms, F measures a quantity proportional to the variance of the feature from class to class divided by the average variance of the feature within each class. The measure is reasonable in that it measures interclass and intraclass variability simultaneously. The exact expression for F is given later. Depending as it does only on the mean and variance of the feature, it is necessarily a complete measure of individual feature worth in only a limited number of cases, including the Gaussian case. The simplicity of its calculation and its reasonableness as a measure of feature

worth even in the absence of a Gaussian distribution are its strong points.

Another argument suggests the use of a technique with a simple model in many problems. The larger the number of parameters of a probability model, the larger the data base must be to estimate these parameters with accuracy. Whereas 10 to 20 samples for each class may suffice to estimate means and variances with sufficient accuracy, 10 times that number might be needed to estimate a probability density from a histogram of a feature having 10 possible values.

Thus, analysis of variance was chosen as the means of ranking the features because it was simple to calculate and the size of the experimental data base approximately matched the number of parameters of the Gaussian model. The method was optimum in some well-defined cases and the data could be tested to see if they met the assumptions of these cases. Finally, the decision techniques to be used involved nothing more complicated than a multivariate normal assumption that seemed to agree with this ranking technique. The technique had been used in speaker identification work before but had not been compared with other techniques [2].

Another condition was deemed important in this study. A ranking of features derived on one set of data should be comparable to a different ranking of the same features based on another set of data. In speaker identification, one would rather not redesign the feature extraction device when adding another speaker. Thus, a consistent ranking is desirable. This condition was experimentally tested.

Discriminant Analysis

Of all the possible *transformations* of the original coordinates one could choose, the author selected "discriminant analysis" [3]. In many ways this is multivariate extension of analysis of variance. A mathematical description will follow, but in words, discriminant analysis determines a rank-ordered set of vectors z_i. The first of these, z_1, is the direction in the original feature space which, when data are projected onto it, produces the best possible single feature as measured by analysis of variance. If the $m-1$ dimensional subspace orthogonal to this first vector (m=original number of features) is then considered and the best possible direction within it is chosen, this direction will correspond to z_2. Finally, z_m is the line perpendicular to all the higher ranking z_i's and F is smallest when data are projected into it. For the case of an equal number of samples from each class this problem is solved analytically as follows.

Define the following quantities.

$\mathbf{X}_{ij}$ m-dimension feature vector corresponding to the jth sample from the ith class.

$\bar{\mathbf{X}}_i = \frac{1}{r}\sum_{j=1}^{r} \mathbf{X}_{ij}$. Mean vector of the ith class.

$\bar{\mathbf{X}} = \frac{1}{k}\sum_{i=1}^{k} \bar{\mathbf{X}}_i$. Overall mean vector.

r Number of samples per class.

k Number of classes.

$B = \frac{1}{k}\sum_{i=1}^{k} (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^T$. Between-class covariance matrix.

$W_i = \frac{1}{r}\sum_{j=1}^{r} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T$. Particular within-class covariance matrix.

$W = \frac{1}{k}\sum_{i=1}^{k} W_i$. Pooled within-class covariance matrix.

λ, z Eigenvalues and eigenvectors to be determined.

In terms of these definitions, the equation relating to discriminant analysis is

$$(B - \lambda W)z = 0. \tag{1}$$

This is the generalized eigenvalue problem. Corresponding to various λ_i, there are z_i. The z's are ranked according to λ since it has been shown that if one projects the data onto one of the z_i's, the F ratio of these values is proportional to λ_i [3].

The method of solution of (1) depends upon the singularity of W. If W is nonsingular, then (1) may be transformed into an ordinary eigenvalue problem by premultiplying by W^{-1}. If W is singular, two approaches seem to exist. In [4] it is suggested that a transformation of coordinates be performed such that in the new space W is invertible. This suffers the disadvantage that potentially good directions (i.e., those with small within-class variance) are eliminated. The author suggests a different approach. In the "modified" discriminant analyses described here, an initial transformation was performed to eliminate singularities in B. The lower dimensionality space remaining, therefore, contained no direction with zero between-class variance. Any dimensionality thus lost was worthless in an analysis of variance sense. Since B in the new space was invertible, the original generalized problem was transformed into the ordinary eigenvalue problem involving the matrix $B'^{-1}W'$, where the prime indicates a matrix in the transformed space. Vectors corresponding to small eigenvalues in this case are the desirable directions.

In order for discriminant analysis to be applied to a problem without violating any assumptions, the following three conditions must be met.

Condition 1: The data must be distributed normally within each class.

Condition 2: The covariance matrices within each class must be equal.

Condition 3: A statistically accurate estimate of the mean vectors and covariance matrices must be available.

The degree to which these assumptions are met with the data of the experiments will be described.

This technique is clearly not usually optimum. An optimum result for a multiclass ($k>2$) problem with normal data having different covariance matrices is unknown, even though many have worked on the problem [5]–[7]. Discriminant analysis was chosen by the author for this work because there are conditions under which it is known to be optimum and it requires a minimum number of parameter

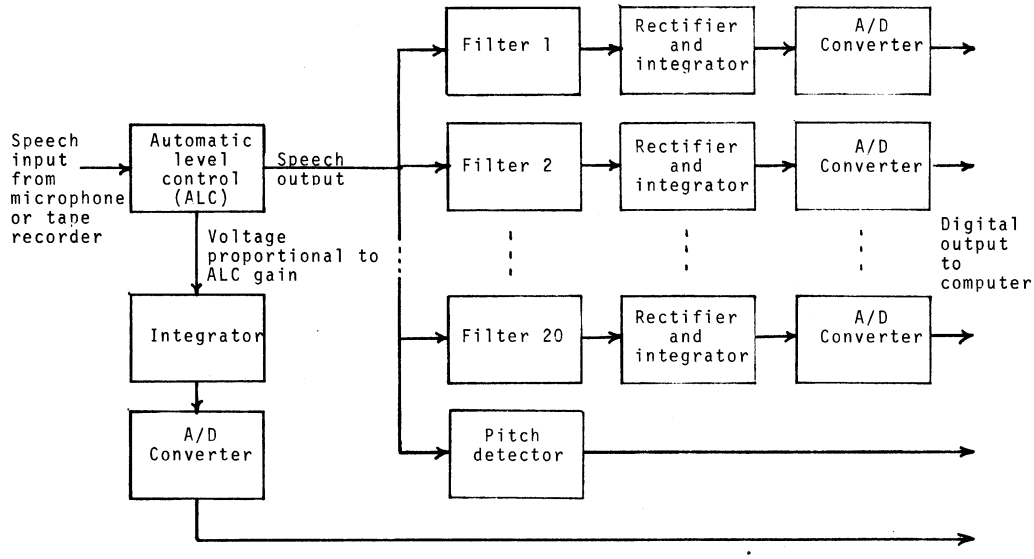

Fig. 1. Functions of analyzing hardware.

estimations. Also, it is more pertinent to the pattern recognition problem than principal components analysis which maximizes total-class variance without minimizing within-class variance [8]. One could say that it makes exactly the same assumptions as analysis of variance except that the covariance terms no longer need to be zero, i.e., features no longer need to be independent.

Experiments

Data Base

The data used in all of the experiments were derived from a set of recordings of the single phrase "Check Available Terminals" by a group of 50 adult male speakers. All of the recorded utterances were converted to digital form by the analyzer shown in Fig. 1 and Table I and stored for use by later programs. Many programmed preprocessing steps were needed before feature evaluation methods could be compared. The first step, segmentation, was implemented in a set of programs designed to find key phonetic events in each utterance.

Next, a preliminary set of 405 features was selected. This set consisted of many types of functions calculated at many points throughout the utterance. Some were averages of the power in certain filters for certain time samples (measured relative to particular segmentation points). Others were time differences, estimated formant amplitudes and frequencies, and hardware-calculated pitch–frequency estimates. In order to take into account the fact that an individual may speak at different rates at different times, the utterances were also "time-normalized." This operation consisted of linearly interpolating between the values obtained from each filter through time such that a constant number of "new" time samples was produced between chosen segmentation points. Formants and averages were extracted from time-normalized data as well as the original data. This set of features was chosen after considering previous experiments by the author and his colleagues, as well as experiments elsewhere in the literature [2], [10], [12].

TABLE I
Filter Bank Specifications

Filter Number	Center Frequency (Hz)	Bandwidth (6 dB) (Hz)
1	188	250
2	459	250
3	715	250
4	969	250
5	1220	250
6	1472	250
7	1725	250
8	1975	250
9	2225	250
10	2475	250
11	2725	250
12	2991	290
13	3300	330
14	3659	390
15	4083	460
16	4586	550
17	5194	670
18	5954	860
19	6932	1110
20	8203	1450

More detail on the segmentation rules and a complete list of features is available elsewhere [9], [11].

Analysis of Variance

The approach of this research has been to answer three questions about each evaluation technique.

1) How well does the data used meet the optimality assumptions?

2) How consistent is the evaluation from data base to data base?

3) How effectively does it reduce dimensionality compared to a random reduction?

The optimality assumption implicit in the model appropriate to analysis of variance is a population of normal distributions, all with the same variance but possibly different means. A test for homogeneity of variance was applied to one feature with data from a set of ten speakers, 100 utterances from each. The results in Table II show that there is

TABLE II
TEST OF HOMOGENEITY OF VARIANCE

Data base characteristics: 10 speakers, 100 utterances (approximately) per speaker, feature number 1 from composite ordering

Speaker	s^2	Speaker	s^2
1	12.97	6	24.00
2	11.37	7	16.94
3	30.42	8	9.12
4	20.97	9	41.70
5	22.91	10	16.85

Estimated chi-squared = 95.7. Tabulated chi-squared = 23.6 at 0.995 level of confidence and 9 deg freedom.

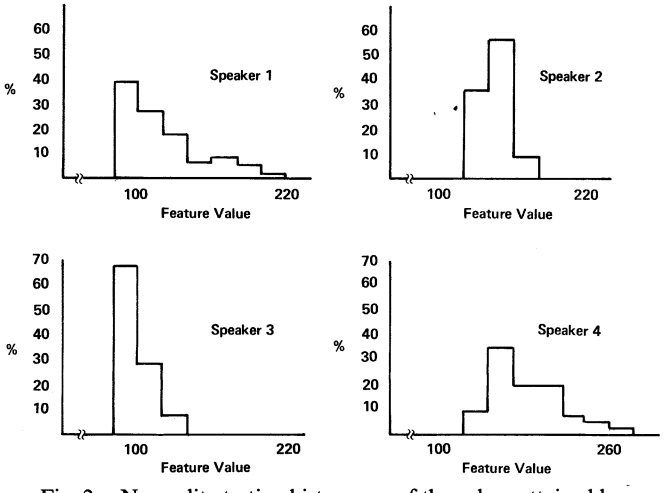

Fig. 2. Normality testing histograms of the values attained by feature 43 for four speakers.

sufficient statistical evidence to reject the hypothesis that the variances of the feature for the various speakers are equal. Next consider the assumption requiring normality. Fig. 2 shows histograms of another feature plotted for four of the ten speakers. The apparent skewed distribution of three of the four is sufficient cause to reject this assumption as well.

Having shown that the assumptions are not strictly met by the data, one is not forced to abandon the technique. There is no optimum method to turn to instead, with the limited amount of data. Other methods may exist that are better for evaluating features with these statistics, assuming these estimates to be accurate enough, but neither those methods nor this will be optimum. Analysis of variance will still evaluate a *reasonable* measure of goodness which is, however, not directly related to the eventual probability of misclassification.

How consistent is the ranking of features based on analysis of variance from data base to data base? In order to answer this question, a set of 20 utterances from 50 speakers was divided into four groups. First, the set of speakers was arbitrarily divided into two groups of 25 speakers each. The utterances of these speakers were divided into two groups, ten from each speaker. Thus, four groups labeled $S1U1$, $S1U2$, $S2U1$, and $S2U2$ were formed where the digit following S refers to the speaker group and that following U the utterance group. The F ratio was then calculated for each of the 405 features for each of these data bases according to the following formula:

$$x_{i.} = \sum_{j=1}^{r_i} x_{ij}; \; x_{..} = \sum_{i=1}^{k} x_{i.}, \tag{2}$$

where r_i is the number of utterances spoken by the ith speaker,

$$\text{total } SS = \sum_{i=1}^{k} \sum_{j=1}^{r_i} x_{ij}^2 - x_{..}^2 / \sum_{i=1}^{k} r_i$$

$$\text{between } SS = \sum_{i=1}^{k} (x_{i.}^2/r_i) - x_{..}^2 / \sum_{i=1}^{k} r_i$$

$$\text{within } SS = (\text{total } SS) - (\text{between } SS)$$

$$\text{between } MS = (\text{between } SS)/(k-1)$$

$$\text{within } MS = (\text{within } SS)/\left(\sum_{i=1}^{k} r_i - k\right)$$

$$F = (\text{between } MS)/(\text{within } MS).$$

In this formula, x_{ij} is the value of the feature in question for the jth utterance of the ith speaker. This relatively complex form of the F ratio is needed when the number of samples per class is unequal, as was the case here because of segmentation errors. The terms "between" and "within" refer to "between speaker" and "within speaker," respectively. SS and MS refer to "sum of squares" and "mean square," respectively.

The four sets of F ratios obtained were used to generate four rank-ordered lists of the 405 features. The desire for "consistency" defined above leads us to ask whether these lists agree. Unfortunately, the theory of analysis of variance is not developed along the line of answering this question. A way of providing a partial answer to the question is provided in rank correlation coefficients. Two that were used here are that by Kendall (t) and by Speakman (r). Both are defined in detail elsewhere [13]. They have the following properties.

Property 1: If the two lists are in perfect agreement, the coefficients are equal to unity and are never greater than unity.

Property 2: If the two lists are in perfect disagreement, that is, the two rank orders are opposite, the coefficient is -1, and never less.

Property 3: If the two lists are purely random and independent of one another, the expected value of the coefficient is zero.

Property 4: The distribution of coefficient values for the random, independent case tends quickly toward the normal distribution with a known variance.

Both coefficients were used to compare the four lists above, but in order to do this efficiently, only a subset of 100 of the features were compared. Table III lists the results. It is to be expected that the Spearman values are somewhat larger than those of Kendall. The least significant correlation obtained is $t = 0.49$ between $S2U1$ and $S1U2$. The probability of a value arising that is this large or larger is

TABLE III
TESTING RANK CORRELATION BETWEEN PAIRS OF ANALYSIS OF VARIANCE RANK ORDERS

Kendall (t)

	$S1U1$	$S2U2$	$S1U2$	$S2U2$
$S1U1$	1.00	0.54	0.69	0.54
$S2U1$	0.54	1.00	0.49	0.71
$S1U2$	0.69	0.49	1.00	0.52
$S2U2$	0.54	0.71	0.52	1.00

Spearman (r)

	$S1U1$	$S2U1$	$S1U2$	$S2U2$
$S1U1$	1.00	0.73	0.86	0.73
$S2U1$	0.73	1.00	0.66	0.88
$S1U2$	0.86	0.66	1.00	0.70
$S2U2$	0.73	0.88	0.70	1.00

1.00 = perfect correlation; 0.00 = random and independent.

equal to the probability that a normal variable is as far as 7.2 standard deviations from its mean, a negligible probability. Thus, even though the apparent correlation of the lists is not high relative to perfect agreement, one can safely reject the hypothesis that the lists are random and independent of one another. It was concluded that the four rank orderings were significantly correlated, and a single *composite rank order list* was formed using the method suggested by Kendall [13].

The answer to the remaining question concerning the effectiveness of dimensionality reduction will be answered after the discussion of discriminant analysis in the next section. Before that, however, it is necessary to describe some other rank orders employed.

First, as described earlier, the analysis of variance rank order does not treat feature dependence. Subsets chosen by taking the top n features in the list most likely contain redundant features and hence the number of features is unnecessarily large. In this case, as in many others, the experimenter knows by their definition that some are very closely related, perhaps involving adjacent time samples or frequencies. By choosing subsets from the analysis of variance rank order while keeping these functional dependencies in mind, one can reduce statistical dependence as well.

A modified analysis of variance rank ordering was formed in just this way. Beginning with the composite rank ordering, features were deleted that duplicated any higher ranking feature significantly. Such an imprecisely defined algorithm will certainly not produce a unique ranking, but it could be expected to at least work better than a random ranking.

To test all of the rankings, a control set of three rank-order lists of features was formed. There were not "ranked" at all, but instead were three random rearrangements of the top 200 features from the composite list. Thus, in the decision experiments described later, there were five orderings, a composite analysis of variance, a modified analysis of variance, and three random orderings.

MODIFIED DISCRIMINANT ANALYSIS

The same three questions asked about the data used for analysis of variance are relevant to discriminant analysis. Are the assumptions met? No, since the assumptions for discriminant analysis are the same as those for analysis of variance except that independence is not required. The same arguments that no other optimum method is available and that this is still a very reasonable technique persuaded the author to continue the evaluation. How effectively does discriminant analysis reduce dimensionality? The answer to this must wait until the report of decision experiment results. The question of consistency remains.

If one ignored consistency, it would in one sense seem reasonable to take a group of 25 speakers, 10 utterances per person, and submit the entire 405 features to discriminant analysis, choosing as new features the linear combinations described by the top ranking eigenvectors. First of all, this is impractical because it would necessitate operating with 405-dimensional matrices in (1). Second, no matter how many features were used, only 24 feature vectors, one less than the number of speakers, would be produced. These 24 vectors determine the hyperplane of the means of the 25 speakers exactly and are not the best general set of vectors for other speakers. They simply form a complete basis set for retaining the entire variation for the particular 25 people involved. The 24 features would, however, be nearly statistically independent for those speakers. As will be shown, the tradeoff between generality of resulting features and feature independence is basic to the method. The compromise decided upon here was the following.

The top ranking 200 features from the composite analysis of variance list were divided into 8 groups, the top 25 in the first, the 26th through the 50th in the second, and so on. Next, a modified discriminant analysis was performed on each of these eight groups using the data earlier called $S1U1$ (speaker group one, utterance group one). Each discriminant analysis resulted in 24 eigenvectors corresponding to 24 nonzero eigenvalues. Since the same data were used by all discriminant analyses, and since the eigenvalue is proportional to the F ratio along the eigenvector, the relative importance of the eigenvectors from the various sets were quantitatively comparable, and a ranking of all 192 vectors was obtained (8 groups, 24 vectors each). These vectors were then applied to the data to produce new feature vectors of 192 elements each.

It was hoped that some generality would be obtained by not having eliminated all but 24 of the dimensions. It

was realized, however, that potential generality was gained knowing that the new features would be statistically dependent.

The test for generality involved two additional discriminant analyses. The data chosen were $S1U2$ and $S2U1$, representing more data of the same speakers and data from different speakers, respectively. Only the first 25 features were employed in these tests. Thus, there resulted three sets of 24 rank-ordered eigenvectors relating to 25 features and based on three different sets of data.

The earlier problem of comparing rank orders of scalars was difficult, but the comparison of rank orders of different vectors is much more difficult. One must somehow decide if the vectors of approximately equal rank correspond to approximately the same direction in a 25-dimensional space.

The author developed the following tool to help answer this question. First, to answer the question of whether or not two vectors agree in direction, one must consider the geometry of a multidimensional space. Referring to Fig. 3, one assumes a uniform angular distribution of unit vectors about the origin in an n-dimensional space. One asks the question: What is the probability that two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, chosen at random, will be at least as close together (in angle) as θ? This is equal to the area of the shaded "cap" $A(\theta, n)$ divided by the total surface area $A(n)$. Expressions for these areas have been given as

$$A(n) = (2\pi^{n/2})/\Gamma(n/2) \tag{3}$$

$$A(\theta, n) = k(n)\int_0^{\theta} \sin^{n-2} x \, dx$$

$$k(n) = 2\pi^{(n-1)/2}\Gamma((n-1)/2) \tag{4}$$

where Γ is the Gamma function [14].

Applying these functions to the case where $n=24$, one finds that the probability that $\theta \leq 1.1$ rad is 0.02 and the probability that $\theta \geq 2.0$ rad is also 0.02. These then are reasonable confidence limits that allow one to say that any two vectors closer together than 1.1 rad or farther apart than 2.0 rad are too parallel to have occurred by chance.

Using these confidence levels, the angles between all vectors derived for set $S1U1$ and those from $S1U2$ were tested for significance. Fig. 4 shows the result. A "+" indicates a pair of vectors significantly in agreement in direction, while a "−" indicates vectors significantly parallel but opposite in sign. (The sign difference is unimportant since any scale factor, including minus one, may be applied to an eigenvector without changing its properties.) Fig. 4 therefore shows the degree of agreement between eigenvectors chosen for different sets of utterances from the same set of people. The tendency of the + and − symbols to cluster on the diagonal indicates the agreement in order and direction of the eigenvectors. For comparison, the 24 vectors derived for $S1U1$ were tested with those from $S2U1$, a different group of speakers. The result is shown in Fig. 5. In this case there are not only fewer significantly parallel vectors, but there is also less clustering about the diagonal, particularly for high-ranking vectors.

This type of analysis apparently cannot be made more quantitative, and one must interpret diagrams such as Figs. 4 and 5 qualitatively to evaluate consistency. A different consistency test was applied as well. The input data $S1U1$ and $S2U1$ were presented to the 24 eigenvectors derived for $S1U1$, a transformation of coordinates was made, and 24-dimension feature vectors were produced. Two sets of analyses of variance were performed on these 24 features, one for the data base $S1U1$ and the other on $S2U1$. Remembering that F ratios calculated for data projected onto eigenvectors were ranked according to eigenvalue, the F ratios should have been monotonic for the data for which the discriminant analysis was performed. The degree to

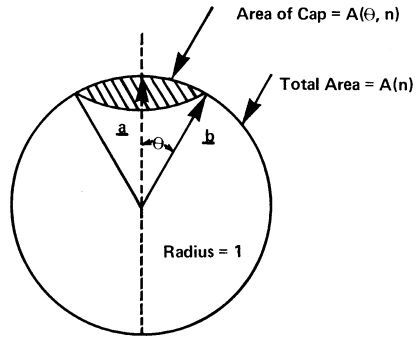

Fig. 3. Geometry for hypersphere cap area calculations in three dimensions.

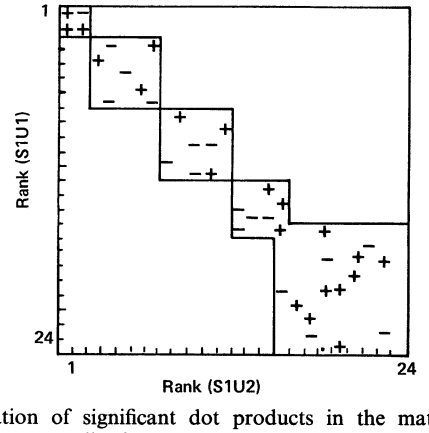

Fig. 4. Location of significant dot products in the matrix of cross products between discriminant analysis eigenvectors, same speakers, different utterances.

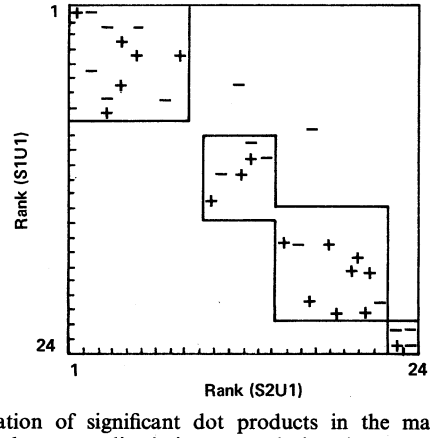

Fig. 5. Location of significant dot products in the matrix of cross products between discriminant analysis eigenvectors, different speakers.

TABLE IV
COMPARISON OF F RATIOS ALONG 24 DIMENSION EIGENVECTORS

Eigenvector Rank Determined on $S1U1$	Analysis of Variance on $S1U1$		Analysis of Variance on $S2U1$	
	F	Rank	F	Rank
1	86.4	1	67.3	1
2	73.9	2	54.7	2
3	53.9	3	24.9	14
4	37.5	4	27.0	12
5	33.7	5	41.6	4
6	26.0	6	51.8	3
7	20.6	7	29.9	9
8	17.0	8	32.4	8
9	9.87	9	15.9	20
10	8.18	11	18.9	18
11	8.48	10	23.8	15
12	5.19	13	40.4	5
13	5.27	12	25.0	13
14	4.20	14	29.7	10
15	2.53	15	28.2	11
16	1.77	17	22.9	16
17	2.18	16	20.6	17
18	0.753	18	17.0	19
19	0.572	19	5.33	23
20	0.353	20	34.1	7
21	0.187	21	39.0	6
22	0.143	22	9.58	22
23	0.064	23	13.4	21
24	0.051	24	4.04	24

Spearman rank correlation coefficient $(r)=0.605$. Corresponding $t=3.56$, which (for 24 df) is significant at 0.99 but not 0.999 level of confidence.

which the F ratios for new data were also monotonic would be indicative of consistency. Table IV shows the results. Note that the $S1U1$ data ranked almost perfectly, as expected. Slight errors were due to the fact that discriminant analysis did not take into account different numbers of utterances per person. The ranking for $S2U1$ data is in considerable disagreement. These rank orders of eigenvectors based on F ratios were given Spearman's test and were found significantly in agreement at a 0.99 level of confidence but not a 0.999 level of confidence. In other words, the F ratios along eigenvectors were ranked considerably less consistently than F ratios of the original features.

Based upon both the hypersphere probability argument and the F-ratio argument, one must conclude that the degree of consistency of eigenvector goodness across different speaker groups is marginal. Part of this may be attributed to insufficient data used to estimate the various covariance matrices. The author feels that the more fundamental reason for disagreement is the fact that the number of classes is equal to or smaller than the original number of features. This must result in eigenvectors in the plane of the means of the speakers used in the analysis, and cannot be particularly appropriate for other speakers whose means almost certainly will not lie in this plane.

DECISION EXPERIMENTS

The author began by asking three questions: 1) How well are assumptions met? 2) How consistent is the method when applied to different data bases? 3) How effectively does the method reduce dimensionality? Fairly definitive answers were found for the first and second questions. The third

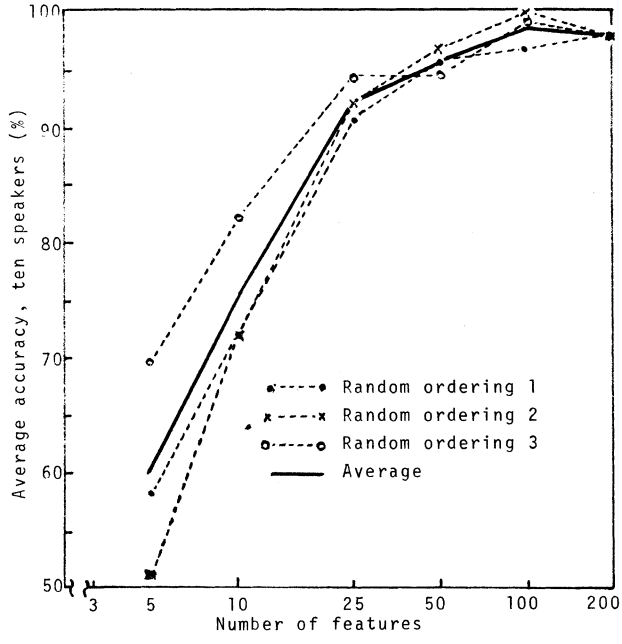

Fig. 6. Effect of reduced dimensionality on classification accuracy (linear decision) for the random feature orderings.

question is best answered by applying the reduced feature sets to decision techniques and comparing recognition accuracy. The choice of what decision method to use is important since accuracy relates both to feature worth and the ability to use feature information effectively.

Two decision methods were used here. The first and possibly simplest of all used only mean vectors for the various classes and classified a new vector as the class whose mean was closest in a Euclidean sense. The second and more complex method assumed multivariate normal distributions of the features and calculated not only mean vectors but also covariance matrices for each class. The conditional probability that a new point came from each class was calculated and the point was assigned to the most probable class.

Since the first was a much simpler method to implement, it was tested on large numbers of features. First, the control group of three random rank orders of 200 features was evaluated. This and all other decision experiments involved just ten people, five from $S1$ and five from $S2$. Mean vectors were calculated using 50 vectors from each of the 10 speakers. Accuracies were derived using an additional 50 vectors from each speaker. Fig. 6 shows the results of 16 experiments with different feature groups. This shows that the random rankings perform essentially equally.

In Fig. 7, the average of three curves in Fig. 6 has been reproduced, along with the results of similar experiments for the composite analysis of variance list features and the discriminant analysis features (based on $S1U1$). All of the experiments were based on the same ten speakers, but in the discriminant analysis case the accuracies of the ten were broken into two groups before averaging: those who happen to have been in speaker group $S1$ and those from $S2$. Generally speaking, the only method of dimensionality reduction that seemed to perform better than random selection was discriminant analysis, and even this only worked

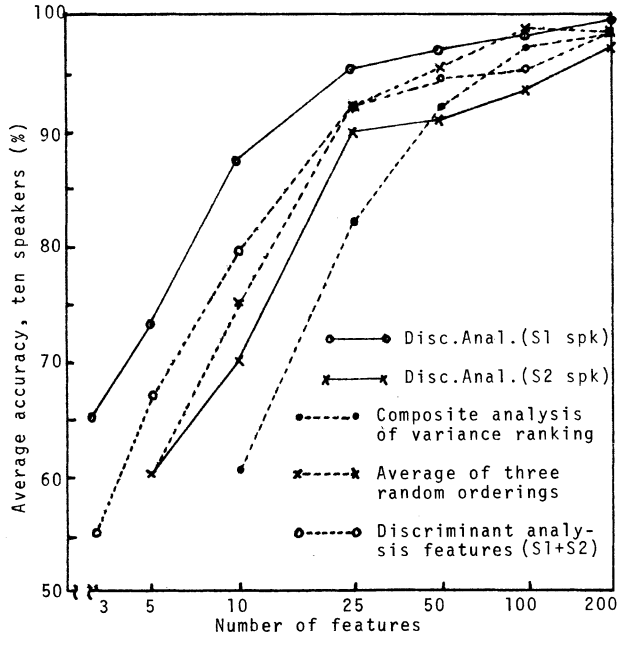

Fig. 7. Effect of reduced dimensionality on classification accuracy (linear decision).

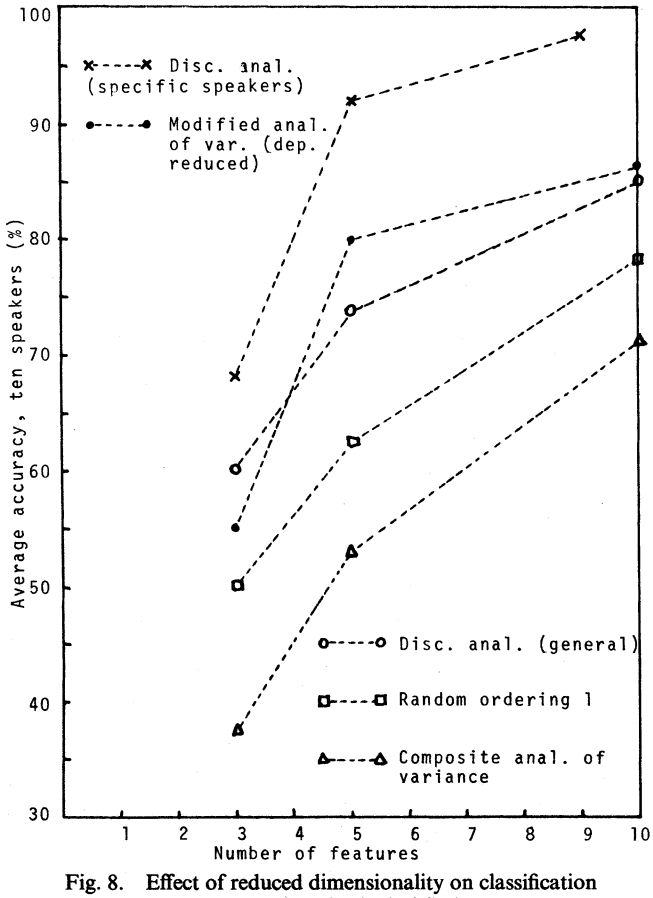

Fig. 8. Effect of reduced dimensionality on classification accuracy (quadratic decision).

better for the people who were in the group for which it was designed.

The results of the quadratic decision method that should be optimum if the features were multivariate normal are shown in Fig. 8. In addition to the feature subsetting method mentioned above, two others were applied to this decision technique. First, it is suggested by all of the above results that a discriminant analysis performed just on the people to be involved in the recognition test should result in minimum error rate. To test this, the top 100 features from the composite analysis of variance list were reduced to just 9 eigenvector features for the 10 decision experiment speakers. The discriminant analysis was performed using mean vectors and covariance matrices estimated on the 50 training utterances of the decision experiment. The curve for these results is labeled "Disc. anal." (specific speakers). The curve labeled "Modified anal. of var." in Fig. 8 is that for the modified analysis of variance list described earlier in which the author attempted to manually eliminate from the composite analysis of variance list those features that were dependent upon higher ranking features.

It is clear from the results shown in Fig. 8 that a specific discriminant analysis performed solely on the people involved in the decision experiment will most effectively reduce dimensionality without excessively reducing accuracy. The modified analysis of variance and the more general discriminant analysis described earlier are somewhat less effective but still better than random selection. A composite analysis of variance without dependence reduction still scores below a random selection.

One may question some of these conclusions based upon the fact that this may not have been the best decision method to use for these data and features. To test this, artificial data were generated that were truly multivariate normal. One set was made to agree with the mean vectors and covariance matrices of the speakers based upon their training data. Another was made to agree with similar parameters estimated from the recognition data. A close agreement in accuracy between real data and artificial data was found. That is, if real training data were tested with means and covariances estimated on the training data, rather high accuracy (97.9 percent) was obtained. Artificial data with the same parameters scored 98.6 percent. Real test data scored 85.1 percent while artificial test data scored 86.4 percent. This similarity certainly suggests that the features produced by a transformation of coordinates based on eigenvectors might in fact be normal even though the original features were not.

Conclusions and Recommendations

This research has compared two methods of feature evaluation and dimensionality reduction for speaker identification. Analysis of variance and discriminant analysis were selected from among the rather large set of possible reduction methods. These methods possessed advantages regarding computational feasibility, known optimality conditions, appropriateness to likely decision techniques, and compatability with the amount of data normally available. Even though the necessary optimality assumptions were not met by the data, each method could be expected to maximize a *reasonable* measure of feature effectiveness.

It may be concluded, based upon the experimental results, that a reasonably consistent ranking of features may be obtained with normally available amounts of data through analysis of variance. These rankings are consistent

in that two rankings based upon independent sets of speakers agree. Such a ranking *must* be modified to remove dependence, however, before subsets of features are selected. On the other hand, discriminant analysis produced much less consistent results for different speakers because the number of speakers involved in discriminant analysis was less than the original number of dimensions of the feature vector. If consistency is not required, discriminant analysis may be expected to reduce dimensionality very efficiently. Furthermore, there is reason to believe that the features produced by the reduction may tend to have a Gaussian distribution. Thus, the optimum decision method is known for these features, and accuracy may be as high in the reduced dimensional space as in the original high-dimensional space.

While these conclusions may appear self-evident, the author contends that they are not. Based upon the literature, it is clear that most investigators know that dependence reduction within a rank ordering is desirable. It is *not* obvious that unless dependence is removed, performance will be *worse* than that obtained with random feature selection. Similarly, it is reasonable that discriminant analysis should be more efficient than rank ordering. It is *not* obvious that the property of consistency depends strongly upon the relationship of the initial number of features and classes. The author suggests that future investigations more carefully consider the intrinsic dependence among the initial features and whether consistency, as defined here, is desirable.

These results suggest a number of areas in which fruitful work may be performed. One is concerned with the determination of performance of other dimensionality reduction methods when faced with a smaller number of classes or samples than the dimensionality of the feature vector. This has implications not only in computational terms such as noninvertible matrices, but also in the generality of the solution obtained. Alternately, one may profitably choose to refine the features extracted so that each feature carries a greater "amount" of independent information and thus a smaller dimensional feature vector may be used. It is possible that in order to obtain features of general usefulness for a large population, it may be necessary to develop entirely new features.

Better features will probably require even better performance of the segmentation or speech recognition step of feature extraction. The more accurately one may compare like sounds from different utterances, the more likely the features derived from them are to be consistent within a given speaker and different between speakers.

The accuracy of current speaker identification efforts is limited not so much by unsophisticated decision methods as by ineffective (and dependent) features. The search for better features may be effectively guided by both of the dimensionality reduction techniques employed in this study.

References

[1] R. G. D. Steele and J. H. Torrie, *Principles and Procedures of Statistics*. New York: McGraw-Hill, 1960.

[2] S. Pruzansky, "Talker-recognition procedures based on analysis of variance," *J. Acoust. Soc. Amer.*, vol. 36, Nov. 1964, pp. 2041–2047.

[3] S. S. Wilks, *Mathematical Statistics*. New York: Wiley, 1962.

[4] D. N. Streeter and J. Raviv, "Research on advanced computer methods for biological data processing," Clearinghouse for Fed. and Sci. Tech. Inform., Springfield, Va., Rep. AMRL-TR-66-24, 1966.

[5] P. J. Min, "On feature selection in multiclass pattern recognition," Ph.D. dissertation, Purdue Univ., Lafayette, Ind., 1969.

[6] J. E. Luck, "A study of spectral speech data—An examination of the segmentation and recognition problem," Ph.D. dissertation, Yale Univ., New Haven, Conn., 1967.

[7] R. G. Casey, "Linear reduction of dimensionality in pattern recognition," Ph.D. dissertation, Columbia Univ., New York, N. Y., 1965.

[8] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

[9] W. S. Mohn, "Statistical feature evaluation in speaker identification," Ph.D. dissertation, North Carolina State Univ., Raleigh, N. C., 1969.

[10] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Academic, 1965.

[11] S. K. Das and W. S. Mohn, "Pattern recognition in speaker verification," in *1969 Fall Joint Comput. Conf., AFIPS Conf. Proc.*, vol. 35. Montvale, N. J.: AFIPS Press, 1969, pp. 721–732.

[12] J. W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Amer.*, vol. 43, Feb. 1968, pp. 368–372.

[13] M. G. Kendall, *Rank Correlation Methods*. New York: Hafner, 1962.

[14] F. H. Glanz, "Statistical extrapolation in certain adaptive pattern-classification systems," Ph.D. dissertation, Stanford Univ., Stanford, Calif., 1965.