Automatic Speaker Identification for a Large Population

HENRY M. DANTE AND V. V. S. SARMA

Abstract—Design of speaker identification schemes for a small number of speakers (around 10) with a high degree of accuracy in a controlled environment is a practical proposition today. When the number of speakers is large (say, above 20 or 30), many of these schemes cannot be directly utilized as both recognition error and computation time increase monotonically with population size. A multistage classification technique gives better results when the number of speakers is large. Such a scheme may be implemented as a decision tree classifier in which the final decision is made only after a predetermined number of stages. In the present paper, analysis and design of a two-stage pattern classifier is considered. At the first stage a large number of classes, to which the given pattern cannot belong, is rejected. This is to be done using a subset of the total feature set. Also, the accuracy of such a rejection process must be very high, consistent with the overall accuracy desired. This initial classification gives a subset of the total classes, which has to be carefully considered at the next stage utilizing the remaining features for an absolute identification of the class label (the speaker's identity). The procedure is illustrated by designing and testing a two-stage classifier for speaker identification in a population of 30.

I. INTRODUCTION

SPEAKER identification systems for a small number of persons (say 5 or 10) have been successfully designed over the past few years [1], [2]. But when these systems have to be designed for a large number of persons, as is the case in any practical application, the classification schemes which work satisfactorily for a small population size often fail. First, the identification error increases monotonically with the number of speakers [3], [4]. Second, the computational complexity and the time taken to make a decision also increase proportionately. This is because, in a speaker recognition scheme, features extracted from a test pattern are compared with each of the stored reference feature vectors of all the speakers and then a decision is made. If classification is done on the basis of discriminant functions, then the discriminant functions for all the speakers have to be computed and compared with each other before a decision is made. Thus the computational time increases linearly with the number of speakers. When a large number of features are used, as must necessarily be the case when the number of speakers is large, this will be formidable. For other classification schemes such as the nearest neighbor scheme, the task will be still more complicated as the number of distances that have to be computed is very large. For the nearest neighbor classification scheme, even the storage requirements increase exponentially with the number of speakers and the dimensionality of the feature vector [5]. The foregoing discussion applies to classification schemes, which may be called single-stage decision schemes in which all features are used in one step to make a decision.

We can look at the problem of speaker identification as a multiclass pattern recognition problem and make use of the recent theoretical results available for such problems. A recent approach for such problems is to go in for multistage schemes in view of the several advantages of these schemes [6], [7]. One of the major advantages of these schemes is that a smaller number of features is used at each stage, thus simplifying computations. When the number of classes is large, as in a speaker recognition scheme for large population, multistage recognition seems to be the only solution.

There are two approaches to multistage pattern recognition: 1) the clustering approach, and 2) the sequential approach. In the clustering approach it is assumed that samples from the classes form a number of nonoverlapping clusters in a particular feature space (the feature is a subset from the available feature set and can be of more than one dimension), i.e., in this feature space we can find a subset of the total classes to which the given test sample belongs. For example, let $w_1, w_2, \cdots, w_M$ be the M classes. Denote by $S = \{w_1, w_2, \cdots, w_M\}$. Let $\underline{X} = (x_1, x_2, \cdots, x_N)$ be the feature vector. Then, there exists a subset of features $\underline{Y} = (x_p, \cdots, x_q)$, in which space classes are clustered as

$$s_1 = \{w_1, w_2, \cdots, w_i\},\ s_2 = \{w_{i+1}, \cdots, w_j\},$$
$$\cdots, s_k = \{\cdots w_M\}, \text{ and } k < M.$$

$s_1, s_2, \cdots, s_k$ are nonoverlapping, i.e., $s_i \cap s_j = \phi$ if $i \neq j$. The restriction $s_i \cap s_j = \phi$ if $i \neq j$ is not necessary in general. We can easily have clusters of overlapping subsets also. This approach has been suggested by Kashyap [8] for speaker identification in large speaker populations. The clustering approach is analogous to the hierarchical classifier approach given by Kulkarni in [6].

The second approach is the sequential approach given in Fu [9]. The method essentially uses a likelihood ratio test, called the generalized sequential probability ratio test. Using one feature at a time, the generalized likelihood ratio is computed for each class, compared with a threshold (which may depend on the particular class), and each class is either

Manuscript received December 30, 1977; revised May 15, 1978, August 31, 1978, and December 19, 1978.

H. M. Dante is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India.

V. V. S. Sarma is with the School of Automation, Indian Institute of Science, Bangalore, India.

0096-3518/79/0600-0255$00.75

retained for further testing at the next stage or rejected from further consideration as unlikely. This process is continued until only a single class remains. The difference between hierarchical and sequential classifiers is that in a hierarchical classifier, a particular class is identified or rejected from further consideration only after a predetermined number of stages (determined by the class in question), whereas in a sequential scheme, any class may be identified or rejected at any stage. Computational considerations make the sequential classifier impractical when the number of classes is large.

The hierarchical classifier could be implemented as a decision tree, and the tree could be optimized according to some optimality criterion. Various forms of decision trees and their optimization procedures can be found in the literature [6], [7]. But the assumption that different classes tend to form natural clusters with reference to some features may not hold well when the number of classes is very large. Instead, a more natural assumption may be that the samples from the classes tend to distribute continuously when the number of classes is large. When this is the case, it is futile to look for clusters. Sequential classifiers do not suffer from this drawback as the feature space is not divided into different regions.

In this paper we show that a knowledge of the distributions of the feature vectors can be utilized for designing a multistage decision scheme for large number of classes. The decision scheme could be broadly split into two stages as follows. In the first stage, by using a subset of features, a small number of classes are picked up with a high degree of accuracy for further testing. In the second stage the actual identification is carried out by using the remaining features. The whole scheme is thus a two-stage decision tree. In Section II we develop a mathematical basis for the first-stage classification and derive an expression for the number of classes that have to be picked in this stage for a specified accuracy. Using this theory, we develop in Section III a two-stage speaker recognition scheme for 30 speakers.

II. Multistage Classification Methods

In this section the mathematical basis for a multistage classification scheme is developed which is subsequently used for the design of a two-stage classifier for identification of 30 speakers. The following assumptions are made.

1) The features $x_1, x_2, \ldots, x_N$ are independent.

2) The number of classes M is very large and can be assumed to approach infinity for analytical purposes.

3) Only one feature is considered at each step.

4) The class conditional density of each feature is normal as given by

$$p(x_j|w_i) \sim N(\mu_{ji}, \sigma_j^2). \tag{1}$$

For each feature x_j, the mean μ_{ji} depends on the particular class label w_i, but the variance σ_j^2 does not depend on w_i.

5) As M tends to infinity, the mean takes a continuous range of values and will be denoted by the continuous variable μ_j. Further, Ω will denote the set of classes for which the mean is in a small interval around and containing a particular value μ of μ_j.

In view of this assumption, as $M \to \infty$, (1) assumes the form

$$p(x_j|\Omega) \underline{\text{def}}\; p(x_j|w_i \epsilon \Omega) \sim N(\mu, \sigma_j^2). \tag{2}$$

Equation *(2)* implies that, for a range of classes $w_i \epsilon \Omega$, the feature x_j is distributed normally with mean μ and variance σ_j^2.

6) The mixture density of any feature x_j over all the classes is assumed to be either: a) uniform, i.e.,

$$p(x) = \frac{}{} \quad \text{if} \quad A \leqslant x \leqslant B$$

$$= 0 \quad \text{otherwise} \tag{3}$$

or b) Gaussian, i.e.,

$$p(x) \sim N(\alpha, K^2). \tag{4}$$

At this point it is appropriate to discuss the meaning of some of the above assumptions, especially assumptions *2)* and 3). When one considers a pattern recognition problem with a finite number of classes, say $w_1, \cdots, w_M$, in the Bayesian analysis one can assign *a priori* probabilities $P(w_1), \ldots, P(w_M)$ to each of these classes which could be assumed equal in the case of equally likely classes. But when $M \to \infty$, one can only talk of an *a priori* density $p(\Omega)$ where $p(\Omega)\,\Delta\mu$ gives the *a priori* probability of having a mean in the interval $(\mu - (\Delta\mu/2), \mu + (\Delta\mu/2))$. Assumptions 1), **3),** and 4) are usual and need no elaboration. Assumption 6) appears to be reasonable as it is and on the basis of experimental evidence to be presented later.

The next step is to establish the forms of the densities $p(\Omega)$ and $p(\Omega|x)$. In the subsequent development, the subscript j of the feature x_j and the corresponding variance σ_j^2 is dropped, as any feature can be considered without loss of generality.

Proposition 1: If $p(x|\Omega) \sim N(\mu, \sigma^2)$ and $p(x)$ is uniform, then $p(\Omega)$ is uniform. On the other hand, if $p(x)$ is Gaussian, then $p(\Omega)$ is Gaussian.

Proof: The proof is given in Appendix I.

Proposition 2: If $p(x|\Omega) \sim N(\mu, \sigma^2)$, $p(x)$ is uniform or Gaussian, then the *a posteriori* density $p(\Omega|x) \sim N(x, \sigma^2)$.

Proof The proof is given in Appendix 11.

Proposition *2* suggests a way to proceed sequentially for the identification task. A subset could be selected from the total set of classes after observing the feature x so that the error of rejecting a correct class is as small as we choose. Of course, a reduction in the rejection error means that a larger set of classes has to be selected and vice versa. The tradeoff can be decided by the final error rate that could be tolerated. The theorem below expresses the fraction of the total classes that has to be considered for further testing when the rejection error is fixed.

Theorem: If $p(x|\Omega) \sim N(\mu, \sigma^2)$ and we fix the probability of rejecting a correct class using feature x at a given stage as q, then the fraction of the total classes as a function of the observed feature x is a constant if $p(x)$ is uniform. On the other hand, if $p(x)$ is Gaussian, the fraction of the total classes depends on the observed feature value x in a complex nonlinear manner.

Proof We want the fraction of the total classes that have to be considered as a function of the observed feature x so

that the probability of rejecting a correct class is q, i.e.,

$$\int_{-\infty}^{\infty} \left[\int_{A} p(\Omega|x)\, d\mu \right] p(x)\, dx = q \tag{5}$$

where A is the region over which $p(\Omega|x)$ is sufficiently small so that (5) is true. Since $p(\Omega|x)$ has mean x and a constant variance σ^2 (as can be seen from A8), the region A is nothing but the interval $(-\infty, x-\lambda) \cup (x+\lambda, \infty)$ where λ is determined from q using the following relation:

$$\int_{-\infty}^{\infty} \left[\int_{A^c} p(\Omega|x)\, d\mu \right] p(x)\, dx = 1 - q$$

where A^c is the complement of A and is equal to the interval $[x-\lambda, x+\lambda]$, i.e.,

$$\int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}\,\sigma} \int_{x-\lambda}^{x+\lambda} \exp\left\{ -\frac{1}{2}\left(\frac{\mu - x}{\sigma} \right)^2 \right\} d\mu \right] \cdot p(x)\, dx = 1 - q.$$

Set $(\mu - x)/\sigma = y$. Then the above equation becomes

$$\int_{-\infty}^{\infty} \int_{-\lambda/\sigma}^{\lambda/\sigma} \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2} \right) dy \right] p(x)\, dx.$$

The term inside the square brackets is independent of x, and hence, the above equation reduces to

$$\int_{-\infty}^{\infty} [\operatorname{erf}(\lambda/\sigma) - \operatorname{erf}(-\lambda/\sigma)]\, p(x)\, dx = \operatorname{erf}(\lambda/\sigma) - \operatorname{erf}(-\lambda/\sigma) = 1 - q \tag{6}$$

where

$$\operatorname{erf}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{z^2}{2} \right) dz$$

is the error function.

From (6) we can conclude that when σ is a constant, λ is determined by the rejection rate q alone. For a fixed rejection rate, when we observe a particular value x of the feature, we can reject all the classes whose mean lies outside the interval $[x-\lambda, x+\lambda]$. Since $p(\Omega|x)$ has mean x and variance σ^2, which is independent of x, this implies that we have to reject all the classes for which $p(\Omega|x) < \epsilon$ where

$$\epsilon = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\lambda^2}{2\sigma^2} \right). \tag{7}$$

Let $f(x)$ denote the fraction of the total number of classes that have to be considered when the observed value of the feature is x. Then

$$f(x) = \int_{\left[\substack{\text{region where} \\ p(\Omega|x) \geq \epsilon}\right]} p(\Omega)\, d\mu = \int_{x-\lambda}^{x+\lambda} p(\Omega)\, d\mu \tag{8}$$

If $p(\Omega)$ is a uniform density, then $f(x)$ is a constant as seen from (8). For any other $p(\Omega)$, $f(x)$ is a function of the observed value of the feature x. In particular, when $p(x)$, and hence $p(\Omega)$, is normally distributed as given in (A4),

$$f(x) = \int_{x-A}^{x+A} \frac{1}{\sqrt{2\pi}\,K} \exp\left\{ -\frac{1}{2}\left(\frac{\mu - \alpha}{K} \right)^2 \right\} d\mu.$$

Set $(\mu - \alpha)/K = y$; then

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{y_1}^{y_2} \exp\left(-\frac{y^2}{2} \right) dy = \operatorname{erf}(y_2) - \operatorname{erf}(y_1) \tag{9}$$

where

$$y_1 = \frac{x - \lambda - \alpha}{K}, \quad Y2 = \frac{x + \lambda - \alpha}{K}.$$

We can observe from (9) that the fraction of the total number of classes to be considered is a nonlinear function of the observed value of the feature x when the feature is normally distributed. In addition, in both the above cases it depends upon the rejection probability q. For a given probability of rejecting a correct class, λ depends upon the variance σ^2 as given in (6), and from (9) we can conclude that $f(x)$, for a given probability of rejecting a correct class, depends on the ratio between K^2 and σ^2 (in a nonlinear manner), which is intuitively quite satisfying.

In a practical system the accepted classes are again tested using one more feature to further reduce the classes to be subsequently considered. This procedure is continued until only a small number of classes are left in the accepted category. The final identification from these classes is done by using the remaining features in the second stage.

In the first stage, after successively testing l features, the fraction of the total number of classes left for further testing is given by $f(x_1) f(x_2) \cdots f(x_l)$ if we assume that the features are independent. The probability of rejecting a correct class after using l features is given by

$$P_{el} = 1 - (1 - q_1)(1 - q_2) \cdots (1 - q_l)$$

where q_j is the prescribed rejection probability at the jth level of the first stage.

In the above formulation we have assumed that the number of classes M is infinite for analytical purposes. But in all practical pattern recognition systems the number of classes is always finite, although it may be quite large. But as M is very large, we commit little error in assuming that $p(x)$ is not multimodal and can be approximated by a uniform or Gaussian density function, and $f(x_i) = (n_i/M)$ where n_i is the number of classes that have to be considered for further testing when feature x_i is used as the ith feature. We can use this procedure until a moderate or small number of classes are left and the identification task can be done using the remaining features.

III. Speaker Identification System Design

In this section the two-stage method developed in Section II is used for designing an identification scheme for 30 speakers.

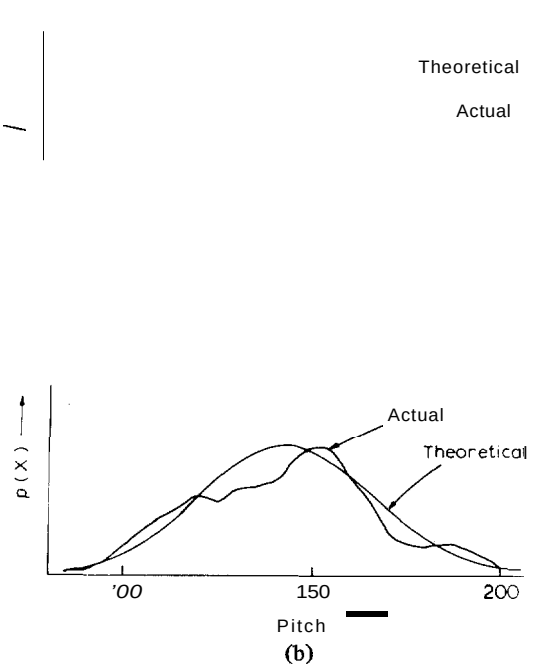

Fig. 1. (a) Distribution of average pitch for a single speaker. (b) Distribution of average pitch for over 30 speakers.

The recognition scheme is based on a single preselected code word **"MUM."** Average pitch over the vowel portion of the utterance is used as a scalar feature at the first stage for picking a subset of the speaker population. At the second stage the autocorrelation function extracted over the utterance is used as a feature. This latter feature and a minimum Euclidean distance classifier have recently been shown to be quite effective in speaker recognition in small (say, 10) population [10], [11].

Pitch contours over utterances have been employed as speaker identifying features for small speaker populations [12]. But a recent study by Atkinson [13] suggests that the intraspeaker variation in pitch is as much as the interspeaker variation. This rules out using pitch alone as a speaker identifying feature. However, in the present study, we show that it can be very conveniently used at the first stage in a two-stage classifier to obtain a subset of speakers with a high degree of accuracy by applying the method of Section **11.** To do this, we need the knowledge of density functions $p(x|w_i)$, $p(x)$, and $p(\Omega)$ defined earlier. In view of this, the statistical properties of the feature, average pitch, for each speaker and for the entire population of thirty speakers for the vowel portion of "MUM" have been studied.

The experimental studies of this paper are conducted on a dedicated interactive signal processing facility. The system is a Hewlett-Packard **5451B** Fourier analyzer built around an **HP2100S** microprogrammable minicomputer with a 16K memory of 16 bit words. Short segments of speech could be entered directly by speaking into a Schure microphone connected to the system's A/D converter. A sampling rate of **10** kHz is used. The beginning of the word is detected by using a triggering level for the A/D converter input. The duration is of length **2048** samples. This duration is chosen, as the operation of the system is in block mode, of block length in multiples of 64, and it has been observed that this duration is adequate. The data set consists of 25 utterances of the code word recorded in a recording room in a single sitting for **30** speakers. Average pitch over the vowel /a/ of the word **"MUM"** is extracted by cepstral technique.

The density functions $p(x|w_i)$, $p(x)$, and $p(\Omega)$ for the average pitch are estimated using the Parzen window technique. The window used is a rectangular one which is equivalent to a modified histogram approach [14, p. 103]. In Fig. 1(a) and (b) the distributions of the average pitch for a typical speaker and also over all the **30** speakers are given. For comparison, we have also given the theoretical distributions, assuming that the form is Gaussian, using the estimated mean and variance. It may be observed that the theoretical and actual distributions in this case match quite closely. The variance of pitch is not the same for **all** the speakers but we have considered the maximum value of the variance to **fix** the threshold λ. The speaker subset densities $p(\Omega)$ (both theoretical and experimental) as defined in Section II are plotted in Fig. 2. In this case, also, the actual and theoretical values match quite closely.

In the design of the system we fix $\lambda = 3\sigma$ **so** that the probability of rejecting a correct speaker [as computed from (6)] in the first stage is **0.3** percent. In Fig. **3** we plot the number of speakers that have to be considered as a function of the observed pitch x. From **(9)** we have computed the function $f(x)$ for $x = \alpha$, i.e., for the mean of the pitch where $f(x)$ is maximum, and also for $x = \alpha \pm 2K$. The values of $f(x)$ are **0.38** and 0.06, respectively. This means that the approximate number of speakers that have to be considered at $x = \alpha$ is $0.38 \times 30 = 11.5$, and at $x = \alpha \pm 2K$ is $0.06 \times 30 = 1.8$. From the plotted curve for the actual case we get the numbers corresponding to $x = \alpha$ as 12, for $x = \alpha + 2K$ as 2, and $x = \alpha - 2K$ as only 1.

The two-stage scheme described so far can be interpreted as a decision tree, but the decision tree cannot be rigid as is

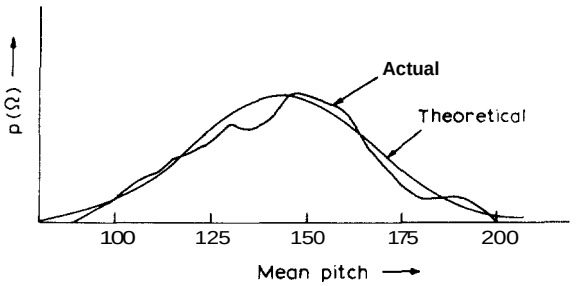

Fig. 2. Distribution of the speakers as a function of mean pitch.

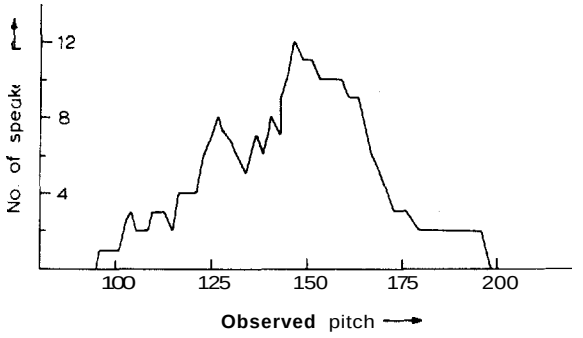

Fig. 3. Number of speakers to be picked up as a function of the observed pitch x.

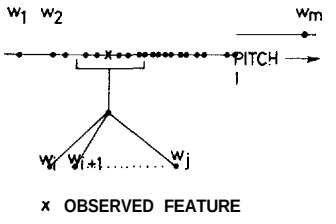

Fig. 4. Decision tree for the two-stage classifier.

the case with the conventional decision tree classifiers given in [6], [7]. The number of classes that have to be considered at the second stage after observing the average pitch varies with the observed value. The structure of the tree is as shown in Fig. 4.

A rigid decision tree could easily be obtained if overlapping subsets are allowed, i.e., a class label can appear more than once at the terminal modes. This results from a slightly modified form of the tree structure of Fig. 4. This tree has the advantage of being rigid though with the disadvantage of increased computational costs. Such a modified tree is shown in Fig. 5.

The reference patterns for pitch and autocorrelation for each speaker are obtained by taking the average pitch and autocorrelation of 10 utterances. The remaining 15 utterances are used for testing the system. In the first stage of the classifier the average pitch of the test utterance is taken, and using the decision tree given in Fig. 4, a subset of speakers is selected from the 30 speakers. In the second stage the normalized 32-point autocorrelation of the test utterance as defined in [10] is computed. The Euclidean distance from this test pattern to the different stored reference patterns is computed.

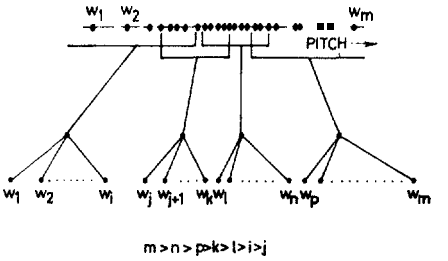

Fig. 5. A possible rigid decision tree structure.

The speaker corresponding to the minimum Euclidean distance from the test utterance is identified.

IV. Results and Discussions

The results obtained by the two-stage recognition scheme described in Section III are given in the form of the confusion matrix in row a) of Table I. For comparison, we have given the results obtained for the one-stage minimum Euclidean distance classifiers using pitch and autocorrelation together in row b), and using only autocorrelation in row c). The zero entries in the confusion matrix are omitted as a **30** X **30** confusion matrix could not be accommodated. While using pitch and autocorrelation together in a one-stage classifier, the pitch was normalized, with respect to the first autocorrelation coefficient by normalizing the maximum deviation of the pitch [14, p. 10]. The overall performance of the two-stage classifier is **87** percent, whereas those of the two single-stage classifiers using both pitch and autocorrelation and autocorrelation alone are **69** and **68** percent, respectively. It may be observed that there is no significant improvement in performance if pitch and autocorrelation are used together in a single-stage classifier. In some cases, addition of pitch as one more feature in the single stage classifier even reduces the performance as can be seen from Table I. The last row of Table I gives the total performance, i.e., for case a), out of **450** test patterns, **391** are recognized correctly with a overall accuracy of 87 percent, whereas for cases b) and c), the recognition score is 312 and **307** (out of **450)** with accuracies of **69** and *68* percent, respectively.

Compared to the single-stage classifier, two-stage classifiers are computationally less expensive. For the speaker recognition experiment discussed so far, if a single-stage classifier is used, using both pitch and autocorrelation in one stage, **30** Euclidean distances have to be computed and then compared to get the minimum among these. In the two-stage classifier scheme, after picking a subset of speakers in the first stage, Euclidean distances have to be computed and compared for speakers in this subset. The number of speakers in this subset is variable as discussed in Section **111,** i.e., the maximum is 12 and the minimum is only 1. ***So,*** on the average, we need to compute only a few distances and then pick up the minimum among these. If only one speaker is picked up in stage 1, then this speaker is identified.

While the computational advantages of the hierarchical classifier are intuitively obvious, the recognition performance improvement as seen from Table I needs some discussion. The performance evaluation of any practical pattern recog-

TABLE I
CONFUSION MATRIX FOR RECOGNITION OF 30 SPEAKERS

TRUE SPEAKEK	RECOGNIZED AS	PERCENTAGE ACCURACY
1	a) 1-17, 6-1, 7-1, 9-1	80.5
	b) 1-9, 6-2, 9-1, 25-3	60.5
	c) 1-11, 7-2, 6-1	80.0
2	a) 2-15	100.5
	b) 2-13, 3-2	86.7
	c) 2-11, 3-4	73.3
3	a) 3-15	100.0
	b) 3-7, 8-8	46.7
	c) 3-6, 8-8, 1-1	40.0
4	a) 4-14, 1-1	93.3
	b) 4-11, 29-4	73.3
	c) 4-10, 29-3, 1-1, 25-1	64.7
5	a) 5-13, 3-2	86.7
	b) 5-10, 3-5	66.7
	c) 5-10, 3-5	66.7
6	a) 6-13, 9-1, 25-1	86.7
	b) 6-10, 9-1, 25-4	66.7
	c) 6-11, 25-3, 9-1	73.3
7	a) 7-11, 6-3, 9-1	73.3
	b) 7-6, 2-1, 6-3, 9-5	43.5
	c) 7-3, 6-5, 9-5, 2-1, 15-1	20.0
8	a) 8-15	100.0
	b) 8-11, 1-3, 7-1	73.3
	c) 8-11, 1-3, 7-1	73.3
9	a) 9-13, 4-1, 7-1	86.7
	b) 9-a, 7-2, 24-1, 25-4	53.3
	c) 9-9, 25-4, 7-1, 24-1	63.5
10	a) 13-15	130.5
	b) 10-13, 14-2	83.3
	c) 10-12, 14-2, 26-1	80.0
11	a) 11-15	100.0
	b) 11-15	100.0
	c) 11-15	100.0
12	a) 12-15	100.0
	b) 12-15	100.0
	c) 12-14, 17-1	93.3
13	a) 13-15	103.0
	b) 13-15	100.0
	c) 14-14, 9-1	93.3
14	a) 14-15	100.0
	b) 14-15	137.5
	c) 14-15	100.3
15	a) 15-10, 2-5	66.7
	b) 15-9, 2-6	60.0
	c) 15-8, 2-7	53.3
16	a) 16-11, 12-2, 13-1, 17-1	73.3
	b) 16-8, 13-4, 17-2, 15-1	53.3
	c) 16-10, 12-2, 13-2, 17-1	66.7
17	a) 17-9, 25-6	60.0
	b) 17-3, 20-8, 23-2, 9-1, 27-1	20.0
	c) 17-7, 27-8	46.7
18	a) 18-15	100.0
	b) 18-15	100.0
	c) 18-15	100.0
19	a) 19-15	100.0
	b) 19-15	100.0
	c) 19-14, 1-1	93.3
20	a) 20-13, 15-1, 27-1	86.7
	b) 20-3, 9-9, 15-2, 7-1	20.0
	c) 20-3, 8-9, 15-2, 28-1	20.0
21	a) 21-11, 2-2, 9-2	73.3
	b) 21-5, 25-5, 9-2, 29-2, 6-1	33.3
	c) 21-8, 25-3, 29-3, 2-1	53.3
22	a) 22-7, 15-7, 2-1	46.7
	b) 22-1, 15-11, 2-2, 17-1	6.7
	c) 22-3, 15-7, 27-3, 2-1, 20-1	20.0
23	a) 23-7, 27-5, 17-2, 18-1	46.7
	b) 23-1, 28-7, 25-3, 4-2, 27-1	6.7
	c) 23-2, 27-5, 28-5, 17-2, 181	13.3
24	a) 24-15	100.0
	b) 24-11, 25-4	73.3
	c) 24-9, 25-5	60.0
25	a) 25-11, 19-2, 29-2	73.3
	b) 25-14, 4-1	93.3
	c) 25-9, 19-3, 29-3	60.0
26	a) 26-15	100.0
	b) 26-13, 4-2	86.7
	c) 26-14, 14-1	93.3
27	a) 27-11, 17-3, 23-1	73.3
	b) 27-13, 23-2	86.7
	c) 27-10, 17-3, 23-2	66.7
28	a) 28-15	100.0
	b) 28-15	100.0
	c) 28-15	100.0
29	a) 23-15	100.0
	b) 29-13, 4-2	86.7
	c) 29-12, 9-2, 19-1	80.0
30	a) 30-15	100.0
	b) 33-15	100.0
	c) 30-15	100.0
TOTAL PERFORMANCE	a) 391/450	87.0
	b) 312/450	69.0
	c) 307/450	68.0

Note: **The confusion matrix is to be read in the following way: in case a) for the two-stage classifier, when the true speaker is 1, he is correctly recognized as speaker 1 twelve times, once as speaker 6, once as speaker 7, and once as speaker 9. Hence, the percentage of accuracy is (12/15) 100 = 80 percent.**

nition scheme is complicated as many factors enter into picture, such as: 1) the finite sample sizes of design and test data, 2) the independence (or otherwise) of features used, 3) the distance criterion chosen, and 4) the number of alternative decisions that are made at a node.

1) The minimum achievable error is the optimal or Bayes probability of error when the class conditional densities are known exactly or can be estimated with an infinite number of labeled samples. Also, as the number of features is increased, the achievable accuracy increases with an infinite sample size. On the other hand, it is now well known that given a fixed training data size, there is tradeoff between the information added by one more dimension and the loss in accuracies of the estimates of joint conditional densities due to the added "parameters" [15], [16]. This clearly explains the fact that the results of case b) in Table I are not much better than case c).

2) The problem of reduction in recognition accuracy as the number of classes increases is demonstrated by Kain [17]. In the context of speaker identification, this is noted by Doddington as reported in Rosenberg's review paper [4]. On the other hand, a decision tree classifier can be constructed from a relatively small number of samples of each class without running into the "curse of dimensionality" [7]. The fact that the performance of case a) is superior to b) and c) (in Table I) is directly a result of the finite sample size used for design on the one hand and the reduction in the number of classes at both stages because of the tree structure.

3) While it would help considerably if the features used at both stages are independent as assumed in the theoretical development of Secion II, the pitch and autocorrelation (as used here) are not that much correlated as only the first 32 samples of autocorrelation function are used as the feature in this study. The pitch peak occurs between 50 and 100 samples in the autocorrelation function (corresponding to pitch periods of 5–10 ms). If the features at the two stages are independent, the recognition accuracy is almost identical to the accuracy of second stage as the first-stage accuracy is prescribed to be very high (99.7 percent). It has been observed in an earlier study that the autocorrelation as a single feature in a 10 speaker recognition scheme gives accuracies of the order of 95 percent [10], while the present two-stage scheme has given about 87 percent. This may be attributed to the small degree of correlation between the features.

4) We observed that for an all male population, the pitch distribution over the entire population is Gaussian, and in the first stage of the classifier, using pitch, a reduction of the speaker population by a factor of almost **3** in the worst case (12 out of **30)** was obtained with 99.7 percent accuracy. For a mixed population we expect that pitch distribution will be nearer to uniform, and a reduction by a factor of 5 or 6 may be expected. Thus, the theoretical results of Section II for the uniform distribution case will be applicable in this case.

5) We have used only 10 samples per class to design the system and the remaining 15 samples to test the system. Since the design set was small and the test set was independent, the results obtained are very pessimistic. There are other more elaborate methods for efficiently making use of the available data [18], which may give still better results. As our objective here is to demonstrate the use of multistage classification technique, this aspect is not pursued in this study.

V. Concluding Remarks

This paper suggests a method of designing computationally attractive speaker recognition schemes with high accuracy for large populations based on the hierarchical classification techniques of pattern recognition. While much of pattern recognition theory deals with partitioning or clustering type of classifiers, for practical realization of recognition schemes interactively designed hierarchical schemes appear to be the most promising [19]. The problems of optimal allocation of features at various nodes of a decision tree and the design of optimum classifiers at each node are being presently pursued by the authors.

Appendix I
Proof of Proposition 1

We are given that $p(x|\Omega)$ is Gaussian with mean μ and variance σ^2, and $p(x)$ is either uniform or Gaussian with mean α and variance K^2. We have to prove that $p(\Omega)$ is uniform when $p(x)$ is uniform and $p(\Omega)$ is Gaussian when $p(x)$ is Gaussian.

In the finite class case $p(x)$, the mixture density, is given by

$$p(x) = \sum_{i=1}^{M} p(x|w_i)\, P(w_i).$$

When the number of classes $M \to \infty$, the summation has to be replaced by integration and $P(w_i)$ by the density $p(\Omega)$. So

$$p(x) = \int_{-\infty}^{\infty} p(x|\Omega)\, p(\Omega)\, d\mu.$$

This is an integral equation and we solve it by the method of substitution.

1) If $p(x)$ is uniform, then

$$p(x) = \frac{1}{B-A} \quad \text{for} \quad A \leqslant x \leqslant B$$

$$= 0 \quad \text{otherwise.}$$

Then

$$\frac{1}{B-A} = \int_{-\infty}^{\infty} p(x|\Omega)\, p(\Omega)\, d\mu \quad \text{for} \quad A \leqslant x \leqslant B,$$

i.e.,

$$\frac{1}{B-A} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \cdot p(\Omega)\, d\mu, \quad \text{for} \quad A \leqslant x \leqslant B. \tag{A1}$$

From the above equation, it is clear that $p(\Omega)$ is zero for μ outside the range $[A,B]$. Suppose $p(\Omega)$ is also uniform. Then

$p(\Omega) = 1/(B - A)$. This solution is quite satisfactory when x is not near the endpoints B or A and $(B - A) \gg \sigma$, which is quite a valid assumption since $(B - A)$ is the whole range of the feature x. For

$$\int_A^B \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \cdot \frac{1}{B-A}\, d\mu = \int_A^B \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} d\mu \cdot \left(\frac{1}{B-A}\right),$$

the integral is approximately equal to 1 when x is not near the boundary points B or A and when $(B - A) \gg \sigma$, so that (A1) is satisfied.

2) If $p(x)$ is Gaussian, then substituting the values of $p(x)$ and $p(x|\Omega)$ gives us

$$\frac{1}{\sqrt{2\pi}\,K} \exp\left\{-\frac{1}{2}\left(\frac{x-\alpha}{K}\right)^2\right\} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} p(\Omega)\, d\mu. \quad \text{(A2)}$$

We assume that $p(\Omega)$ is normal with mean μ_0 and variance ρ^2. Substituting for $p(\Omega)$ in the above equation (A2),

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}\,K} \exp\left\{-\frac{1}{2}\left(\frac{x-\alpha}{K}\right)^2\right\} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \frac{1}{\sqrt{2\pi}\,\rho} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\rho}\right)^2\right\} d\mu \\
&= \frac{1}{2\pi\sigma\rho} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\rho}\right)^2\right]\right\} d\mu \\
&= \frac{1}{2\pi\sigma\rho} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_0)^2}{\sigma^2+\rho^2}\right]\right\} \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\,\frac{\sigma^2+\rho^2}{\sigma^2\rho^2}\left(\mu - \frac{\rho^2 x + \sigma^2\mu_0}{\sigma^2+\rho^2}\right)^2\right\} d\mu \\
&= \frac{1}{\sqrt{2\pi}\,\sigma\rho} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_0}{\sqrt{\sigma^2+\rho^2}}\right)^2\right\} \cdot \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}.
\end{aligned} \quad \text{(A3)}$$

(A3) represents a normal density, and from (A2) and (A3),

$$\mu_0 = \alpha, \text{ and } \sigma^2 + \rho^2 = K^2$$

or

$$\rho^2 = K^2 - \sigma^2.$$

But $\sigma^2 \ll K^2$ since σ^2 is the class conditional variance of a feature, whereas K^2 is the variance of the mixture density. So $\rho^2 \approx K^2$, and hence

$$p(\Omega) \sim N(\alpha, K^2). \quad \text{(A4)}$$

Appendix II
Proof of Proposition 2

We have to prove that $p(\Omega|x) \sim N(x, \sigma^2)$. By Bayes rule,

$$p(\Omega|x) = \frac{p(x|\Omega)\, p(\Omega)}{p(x)} = a \cdot p(x|\Omega)\, p(\Omega) \quad \text{(A5)}$$

since $p(x)$ is a constant for a given x.

1) When $p(x)$, and hence $p(\Omega)$ is uniform, we get

$$p(\Omega|x) \sim N(x, \sigma^2).$$

2) When $p(x)$ is Gaussian, $p(\Omega)$ is also Gaussian as given in (A4). *So*

$$\begin{aligned}
p(\Omega|x) &= a \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right\} \cdot \frac{1}{\sqrt{2\pi}\,K} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\mu-\alpha}{K}\right)^2\right\} \\
&= a' \cdot \exp\left\{-\frac{1}{2}\left[\left(\frac{\mu-x}{\sigma}\right)^2 + \left(\frac{\mu-\alpha}{K}\right)^2\right]\right\} \\
&= a' \cdot \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{K^2}\right) \cdot \mu^2 - 2\left(\frac{x}{\sigma^2} + \frac{\alpha}{K^2}\right)\mu\right\}.
\end{aligned} \quad \text{(A6)}$$

This is a normal density, and we represent it by $N(\mu_0, \beta^2)$ where

$$\mu_0 = \frac{K^2 x + \sigma^2 \alpha}{K^2 + \sigma^2} \quad \text{(A7)}$$

and

$$\beta^2 = \frac{\sigma^2 K^2}{\sigma^2 + K^2}$$

and $p(\Omega|x) \sim N(\mu_0, \beta)$. Since $K^2 \gg \sigma^2$, $\mu_0 \approx x$, and $\beta^2 \approx \sigma^2$, and hence,

$$p(\Omega|x) \sim N(x, \sigma^2). \quad \text{(A8)}$$

Acknowledgment

The authors wish to thank the reviewers for many suggestions which helped considerably in improving the presentation of this paper.

References

[1] V. V. S. Sarma and B. Yegnanarayana, "A critical survey of automatic speaker recognition systems," *J. Comput. Soc. India*, vol. 6, p p 9–19, Dec. 1975.

[2] B. S. Atal, "Automatic recognition of speakers from their voices," Proc. *IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[3] D. Venugopal and V. V. S. Sarma, "Performance evaluation of automatic speaker recognition schemes," in *Conf. Rec.*, 1977 *IEEE Conf.* Acoust., *Speech, Signal Processing*, 1977, pp. 780–783.

[4] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475–487, Apr. 1976.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[6] A. V. Kulkarni, "Optimal and heuristic synthesis of hierarchical classifiers," Ph.D. dissertation, Univ. Maryland, College Park, 1976.