MULTISTAGE DECISION SCHEMES FOR SPEAKER RECOGNITION

H.M. Dante
Dept. of Elec. Commun. Engg

V.V.S. Sarma and G.R. Dattatreya
School of Automation

Indian Institute of Science
Bangalore-560012

ABSTRACT

Speaker recognition schemes which work satisfactorily for small populations often fail when the number of classes is very large. One way of solving such problems is to go in for multistage classification schemes. The basic technique is to successively reduce the number of classes in several stages using one feature at each stage and when the number of classes is less than a predetermined value, then the final decision is made. The whole scheme is designed so that the probability of error is fixed at an acceptable level. The computational cost of such a multistage scheme depends on the features used at each stage and the cost of measurement of each feature. The features to be used at each stage are determined so as to reduce the average computational cost for making a decision. This procedure is formulated as a stochastic optimal control problem and is illustrated by designing a speaker recognition system for 60 speakers. The overall accuracy of the system is 97.2 %.

I. INTRODUCTION

In many pattern recognition problems of practical interest, such as fingerprint classification, identifying a chemical compound from its mass spectrogram and identification of a person by his voice for forensic and security applications, the number of classes is very large, several thousands or even tens of thousands[1] Most of the published literature in pattern recognition deals with a small number of classes, around ten. Methods which are available for ten classes cannot be extended directly to a very large number of classes as both recognition error and computation time increase monotonically with population size. For example, schemes which work efficiently for ten speaker recognition do not work satisfactorily for larger populations[2,3]

Multistage classification schemes give better results when the number of classes is large. An approach to this problem is as follows: At first phase, large number of classes to which the given sample cannot belong to are rejected. This process could be executed as a multistage decision process. The resulting subset of the total classes is carefully considered at the next phase for an absolute identification of the class label. The whole scheme is thus a decision tree. The features to be used at each node of the tree and the decision strategy are evolved by minimizing a cost function.

A two stage classification scheme using the approach outlined above was used for speaker identification in a population size of 30[4]. This was done on a purely huristic basis and no optimization technique was used. In section II, we give an outline of this scheme. and a systematic approach to optimize the decision tree based on stochastic optimal control.

II. OPTIMIZATION OF MULTISTAGE CLASSIFIER

The multistage scheme pursued in this study is a development over the adhoc 2-stage scheme introduced in the earlier study[4]. The following assumptions are made.

1. The number of classes, K is very large
2. When the number of classes is less than or equal to c, we have a good classification scheme using a feature victor Y. An additional feature set $X = \{x_1, \ldots, x_N\}$ is used to reduce the population size from K to c
3. The feature $x_1, \ldots, x_N$ are independent
4. Only one feature is used at each step
5. The class conditional density of each feature is normal with different means and same variance.
6. The mixture density of each feature for population sizes above c is Gaussian

In the earlier study[4], we have shown that the number of classes that has to be considered for a fixed probability of rejecting a correct class is a non-

CH1379-7/79/0000-0797 [illegible]

linear function of (a) the observed value of the feature, (b) the class conditional variance and (c) the variance of the mixture. In other words, the fraction of the total number of classes that will be picked up after using i features is given by $g_i(\alpha_1,\alpha_2,\ldots \alpha_i) = f_1(x_1= \alpha_1)f_2(x_2=\alpha_2)\ldots f_i(x_i=\alpha_i)$; and when K is large, $n_i=g_iK$. So, the number of steps necessary to reduce the given population size from K to c is a variable quantity.

With this background, we proceed to the optimization of the multistage classifier, i.e. allotment of features at the different stages of the classifier so that the overall cost of the whole scheme is a minimum. The problem can be formulated as follows:

We are given K classes. We have to choose one feature at a step from the feature set $X = \{x_1,\ldots x_N\}$ and reduce the population size from K to c. We assume that the probability of rejecting a correct class for each step is fixed. Each feature is associated with a cost of measurement $C_j, j =1,\ldots N$. After using one feature, say x_i, depending on the outcome the number of classes will be reduced from K to n_i. If $n_i \leq c$, we can proceed to the second phase, i.e. classifying using feature vector $\underline{Y}$. If $n_i > c$, we have to use some other feature $x_j, x_j \in \{X - x_i\}$ to further reduce the number of classes. This procedure is continued till the number of classes becomes less than or equal to c. The number of steps necessary to arrive at the second phase depends on the outcome of the feature measurements and hence is a random variable. Since the number of classes may vary anywhere between c to K, optimization has to be done for each integer n, $c < n \leq K$. This is an impossible task, so we divide the range from c to K into a number of intervals. That is, if $c < n \leq c_1$, denote this range by L_t. If $c_1 < n \leq c_2$, denote by L_{t-1} etc., till $c_t < n \leq K$ denote by L_i. When a feature is measured at a particular stage where the numbers of classes is in the range L_i, we can go to any L_j, $j > i$. So, if we draw this process as a decision tree classifier, the decision tree will look as in Fig. 1. To this tree features should be allotted in an optimal way so as to minimize the expected cost.

We can model the problem of optimizing the decision tree as a stochastic optimal control problem. Let us consider the stochastic optimal control problem(5),

$$z_{k+1}= F_k(z_k,u_k, \nu_k), k = 0,1,\ldots M-1 \quad (1)$$

where z_k is the state of the system at the time instant $u_k \in U_k$ is the control variable, and ν_k is the plant noise. Let the observations be given by

$$v_k = z_k \quad (2)$$

The performance index is defined by

$$J = \sum_{k=1}^{M} R_k(z_k, u_{k-1}), R_k \geq 0, \quad (3)$$

When the stopping time M is a random variable, the above formulation has a general(5),

$$J = \sum_{k=1}^{M} R_k (z_k, u_{k-1}; M) \quad (4)$$

The optimization problem is of finding $u_k \in U_k$ for minimizing the cost function J could be minimized with respect to u_k(5). In the pattern recognition context described above, we can consider the intervals $L_1,L_2,\ldots L_t$ as t states i.e. when the number of classes is in a particular range denoted by L_j, we say that the system is in state L_j. The state itself is the observable as given in eqn.(2). Since we can reach the second phase at any step, the stopping time M is a random variable. The new state(the resulting number of classes) depends not only on the feature selected but also on the value of the feature. In other words,

$$L_{k+1} = F_k (L_k, x_k, \alpha_k) \quad (5)$$

where L_k is the number of classes at kth step, $x_k \in X - (x_1,\ldots x_{k-1})$ and α_k is the observed value of the feature. The performance index is

$$J = \sum_{k=1}^{M} R_k (L_k, x_{k-1} ; M) \quad (6)$$

This performance index has to be minimized with respect to the features $x_1,x_2 \ldots x_N$.

The transition from state L_i to L_j $(j > i)$ is probabilistic since it depends on the outcome of the feature measurement. Since we have assumed full knowledge of the probability distributions, we can compute the probability of transition from state L_i to L_j. As a result, we can estimate the expected cost.

At state L_i, we can write the expected cost of using feature x_m as $C(L_i,x_m)=C_m+$ {Expected cost of reaching the final state from L_i} (7)

where C_m= cost of measurement of feature x_m and $x_m \notin \{x_1,\ldots x_{m-1}\}$

If we know the optimum costs for all states L_j $(j > i)$, we can estimate the expected cost of using feature x_m at state L_i. Let the transition probability from Li to Lj be $PL_i(L_j)$. Then the cost of using feature x_m at state L_i is

$$C(L_i,x_m)= C_m + \sum_{j=i+1}^{t} P_{L_i}(L_j)\ S(L_j) \quad (8)$$

where $S(L_j)$ is the expected cost of the optimal decision rule at state L_j. So, if we know $S(L_j)$ for all $L_j (j>i)$, then we can find the expected cost of using x_m at state L_j. Finally, choose that feature which gives the minimum of $C(L_i, x_m)$, $x_m \not\in \{x_i, .. x_{m-1}\}$. So, the optimal decision policy at any state depends on the particular state and also the path through which this state is reached.

To find the optimal decision at any state, the optimal decisions of all the subsequent states is to be known. From this it follows that the optimum decision policy at the starting node can be found by averaging out and folding back the decision tree. To start with, the expected costs of the state L_t for all different paths leading to L_t are computed. Next, the expected costs of the state L_{t-1} for all paths leading to L_{t-1} are computed. By going back in this way, we can arrive at the first node L_1. The optimization cannot be done by the usual dynamic programming procedure because the features used once cannot be used again.

III. SPEAKER RECOGNITION EXPERIMENT

The multistage classification scheme outlined in section II is used to design a speaker recognition scheme for 60 male speakers, based on two preselected code words 'MUM' and 'NUN'. The features that are examined for use at the first stage of the classifier are the formants of the nasal sound /n/ of the word NUN and the formants of the vowel /a/ in MUM. At the second stage, 32 point autocorrelation function over MUM is used for the final classification. The value of c is fixed as 5 so that when the number of classes picked up at the first stage is 5 or less, we go to the second stage to make the final classification.

A tenth order linear prediction analysis was used for extracting the formants. The formants are extracted from the differentiated LP phase spectrum(6). The speech was sampled at 10 KHz and 256 samples of the particular sound (/n/ and /a/) were used in the LP analysis. The data set was of 25 utterances of 'MUM' and 25 utterances of 'NUN' recorded in a single sitting in an Anechoic Chamber. Ten of the utterances were used for designing the system and the remaining 15 for testing the system.

We assume that both the class conditional and mixture densities of the formants are normal. It has been shown in an earlier study(4), that the fraction of the total classes that has to be considered after using a particular feature for a given probability of rejecting a correct class depends on the ratio between the variances of the class conditional distribution and the mixture distribution when the feature is normally distributed. So, the discriminating power of the feature is higher when this ratio is smaller. Since all the formants have the same measurement cost, the formant with the highest discriminating power is used first, followed by the formant having the next highest discriminating power, etc. so that the average cost of the scheme is minimum. The formants considered at the first phase are (1)fourth formant of /n/, (2) second formant of /n/, (3) fourth formant of /a/, and (4) first formant of /a/ in the order of discriminating power. These four formants were sufficient to reduce the population from 60 to 5 in all the cases. However, in the majority of the cases, reduction to 5 was achieved with only 3 formants and less. When the population falls to 5 or less, 32 point autocorrelation function is used as feature with a minimum Euclidian distance classifier for the final identification(4).

The overall accuracy of the whole system was 97.2 %. There were only 25 errors out of 900 test samples used.

IV. AN ALTERNATE APPROACH

In multistage speaker recognition schemes, the computational complexity obviously increases with the number of stages due to increasing number of features measured. In order to reduce this complexity, a system which rejects a constant fraction of speakers at every stage is also considered as an alternative the scheme of section II. Optimal design of such a classifier and an illustrative experiment with three features are described in this section.

The problem is to specify the first feature to be measured and the second feature as a function of the first feature outcome in such a way that the probability of error is minimized. Rejection of speakers at every stage is on the basis of updated class probabilities. This is solved by a backward recursive computation of the conditional expected costs over a graph of all possible feature sequences. Fig. 2 shows such a graph with three features f_1,f_2,f_3.

Given an ordered outcome $x_i x_j x_k$ of features $f_i\ f_j\ f_k$ $(i,j,k = 1,2,3;$ $i \neq j \neq k)$, let $R_{ijk}(x_i\ x_j\ x_k)$ be the conditional risk. This can be computed by updating the class probabilities and rejecting the specified fraction of the less probable classes.

$R_{ij}(x_i)= \int\int R_{ijk}(x_i x_j x_k)p(x_j x_k/x_i)dx_j\ dx_k$

is the conditional expected risk given that x_i has been observed as an outcome of fi and a decision to measure f_j has been taken. $R_i^*(x_i)= \underset{j,k}{\mathrm{Min}}\ \left(R_{ij}(x_i),R_{ik}(x_i)\right)$ is the minimum conditional expected risk given x_i. Similarly $R_i = \int R_i^*(x_i)p(x_i)dx_i$ and $R^* = \mathrm{Min}\ (R_1,R_2,R_3)$is the minimum risk. The first feature and the second feature as a function of the first feature outcome are obtained by backtracking the above procedure.

Three scalar features were used in the experiment – third formants of vowels in 'MEAN' (f_1) and 'LAME' (f_2) and the pitch of the vowel in 'MEAN' (f_3) for sixty four speakers. Three fourths of the speakers were rejected at successive stages. Features were assumed to be independent and normally distributed.

A computer design gave the following results: Average pitch (f_3) is the first stage feature with its range divided into eleven intervals with different second stage features for each interval. The theoretical probability of error is 0.295.

The classifier was tested with the design set and an independent test set. The design set error rate was 0.286 and the test set, error rate was 0.361. These results should be viewed in comparison with the design set Bayes error of 0.26. The Bayes classifier requires considerably higher computation.

V. CONCLUSION

Speaker recognition problem when the number of speakers is very large can be solved by a multistage classifier which can be implemented as a decision tree. The optimization of this decision tree to minimize the expected cost is formulated as stochastic optimal control problem. This scheme was used for designing and testing a speaker recognition scheme for 60 speakers. The overall accuracy of the system was as high as 97.2 %.. The alternate scheme of section IV using formants and average pitch has not given accuracies comparable to formants and autocorrelation function.

REFERENCES

1. H. Chernoff, 'Some applications of a method of identifying an element of a large multidimensional population' in P.R. Krishniah (Ed), Multivariate Analysis – IV, North Holland Publishing Company, 1977.

2. V.V.S. Sarma, 'Automatic speaker recognition systems', presented at the First All India Interdisciplinary Symposium Speech and Hearing, Bangalore, 1978.

3. D. Venugopal and V.V.S. Sarma, 'Performance evaluation of automatic speaker recognition systems', IEEE 1977 Conf. ASSP, 1977, pp.780–783.

4. H.M. Dante and V.V.S. Sarma, 'Automatic Speaker Identification for a large population', to appear in IEEE Jr.Acoust.Speech, Sig. Proc., 1979.

5. M. Aoki, Optimization of Stochastic Systems, New York, Academic Press, 1977.

6. Yegnanarayana, B., 'Formant extraction from linear prediction phase spectra', J.Acoust.Soc. Am., Vol. 63, No.5, May 1978, pp. 638–640.

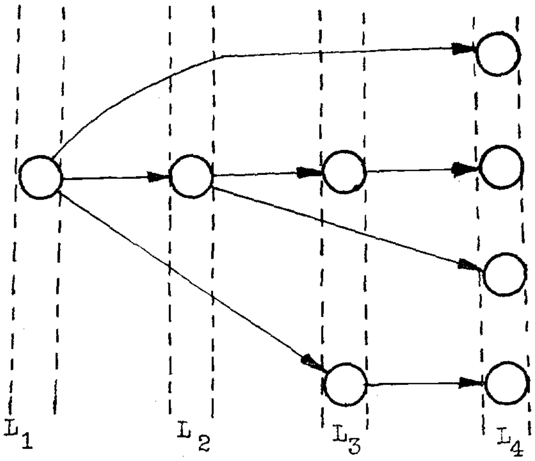

Fig.1: Decision tree for a four stage system.

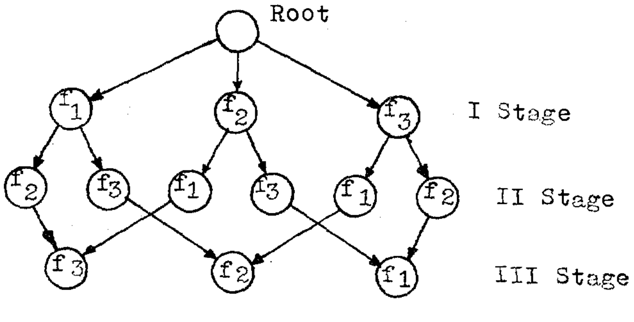

Fig. 2: Feature Sequence Graph