AUTOMATIC SPEAKER IDENTIFICATION FOR A LARGE NUMBER OF SPEAKERS

H.M. DANTE AND V.V.S. SARMA

Indian Institute of Science, Bangalore 560 012, INDIA

Abstract

Design of speaker identification systems for a small number of speakers (around 10) with a high degree of accuracy has evolved over the past few years. A sequential identification technique gives better results when the number of speakers is large. This scheme is implemented as a decision tree classifier in which the final decision is made only after a predetermined number of stages. The error rate could be controlled consistent with the features selected. This paper describes a 2-stage decision tree classifier implemented on a HP Fourier analyser system for identification of 30 speakers. The scheme proceeds as follows. At the first stage of the decision tree the population under consideration is reduced by a factor of 3 with high degree of accuracy, and at the second stage, the scheme available for ten speaker identification is used. The computational savings and performance achieved are compared with that obtained in a single stage classifier.

I Introduction

Speaker identification systems for a small number of speakers (say 5-10) have been successfully designed over the past few years [1]. But when these systems have to be designed for a large number of speakers, the classification schemes, which work satisfactorily for a small number of speakers, often fail. Firstly, the identification error increases monotonically with the number of speakers [2, 3]. Secondly, the computational time for the identification task increases linearly as the number of speakers. We can look at the speaker identification problem as a multiclass pattern recognition problem. A popular way of solving this is by using a multistage classification scheme. In fact, when the number of classes is large, for example, as in a recognition scheme for 100 speakers, multistage schemes seem to be the only solution. There are two approaches to multistage pattern recognition; (i) Clustering approach and (ii) 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. This approach has been suggested by Kashyap [4] for speaker identification when the number of speakers is large. The clustering approach is analogous to the hierarchical classifier approach given by Kulkarni [5].

In the sequential approach, each class is individually tested using features taken one at a time and then either accepted for further testing or rejected depending on the value of the test statistics [6]. The threshold for acceptance is fixed depending on the accuracy needed. This is continued till a single class remains in the accepted category. In a hierarchical scheme, a particular class is identified or rejected only after a predetermined number of stages (determined by the class in question) whereas in a sequential scheme, any class may be accepted or rejected at any stage. Computational considerations make sequential classifier impractical when the number of classes is large. The hierarchical classifier could be implemented as a decision tree in which the number of terminal nodes is equal to the number of classes. Here it is assumed that the classes tend to form natural clusters and this assumption is open to question. Even if it is assumed that the classes are linearly separable, when sufficient number of features are used, there may not be any tendency among them to form clusters with respect to some subset of features. Instead, a Gaussian or uniform assumption on the feature distribution seems to be better suited. So, a sequential decision scheme which does not make any such assumptions is expected to give better results.

II Sequential Methods

In this section, it will be proved that a sequential method could be used for identification with any accuracy consistent with the features selected. Let the number of classes be very large so that w_i can be taken as a continuous variable, w. Let x be the feature, which is assumed to be a scalar; $p(x|w)$, the class conditional density function of the feature x; and $p(x)$, the density function of feature x over the entire feature space. Equal a priori probabilities for each class are assumed.

CH1285-6/78/0000-0295$00.75©1978IEEE

Lemma: (a) Let (i) the class conditional distribution of the feature x be normal with different means and the same variance; i.e. $p(x|w) \sim N(\mu, \sigma^2)$; and (ii) Let the feature x over the entire feature space be uniformly distributed. Then the a posteriori probability density of class w is normally distributed, i.e. $p(w|x)$ is normal.

(b) Let (i) be the same as in case (a), and (ii) the feature when considered over the whole feature space be normally distributed; i.e. $p(x) \sim N(\alpha, K)$. Then, the a posteriori probability density of class w is normal.

Proof: By Bayes rule,

$$p(w|x) = \frac{p(x|w)\, p(w)}{p(x)} = a.\, p(x|w)\, p(w) \qquad (1)$$

since $p(x)$ is constant for a given x. Let $p(w)$ be the probability density function for the classes w; i.e. probability density of the class having mean μ. Since all the class conditional densities have the same variance, σ^2, the probability density of picking a class with mean μ is $p(x)$ itself.

So, $p(w)$ is uniform when $p(x)$ is uniform and $p(w)$ is gaussian when $p(x)$ is gaussian. As a result, in case (a),

$$p(w|x) = p(x|w).p(w).a = p(x|w) \; .$$
$$\text{i.e.} \quad p(w|x) \sim N(x, \sigma^2) \; .$$

For case (b), the proof is similar to the one given in Duda and Hart [7] in a different context.

$$p(w|x) = a.\, p(x|w).\, p(w) \quad \text{where}$$

$p(w) = (1/\sqrt{2\pi}K) \exp\left\{-1/2[(\mu-\alpha)/K]^2\right\}$; Therefore,

$$p(w|x) = a(1/\sqrt{2\pi}\;) \exp\left\{-[(x-\mu)/\sigma]^2/2\right\}.(1/\sqrt{2\pi}K).$$
$$\exp\left\{-[(\mu-\alpha)/K]^2/2\right\}$$
$$= a'.\exp\left\{-[(\mu-x)/\sigma]^2/2-[(\mu-\alpha)/K]^2/2\right\}$$
$$= a''.\exp\left\{\left(-[1/\sigma^2+1/K^2]\mu^2+2[x/\sigma^2+\alpha/K^2]\mu\right)/2\right\} \qquad (2)$$

since functions of α and x are constants, (2) is a normal density function. So we put

$$p(x) = [1/\sqrt{2\pi}\beta] \exp\left\{-[(\mu-\mu_o)/\beta]^2/2\right\} \qquad (3)$$

Combining eqn.(3) and eqn.(2) and solving for μ_o and β, we get $\mu_o = [K^2x+\sigma^2\alpha]/[K^2+\sigma^2]$ (4)
and $\beta^2 = \sigma^2K^2/(\sigma^2+K^2)$ (5)
and $p(w|x) \sim N(\mu_o, \beta)$.

If $K^2 \gg \sigma^2$, (which is valid since K^2 is the variance of x for all classes which is much larger compared to the variance σ^2), we get $\mu_o \approx x$ and $\beta^2 \approx \sigma^2$, and

$$p(w|x) \approx (1/\sqrt{2\pi}\;) \exp\left\{-[(\mu-x)/\sigma]^2/2\right\}$$

The above lemma suggests a way to proceed sequentially for identification. A sub class could be selected from the total number of classes after observing the feature x so that the error of rejecting a correct class is as small as we choose. The theorem below expresses the fraction of classes that have to be considered for further testing when the rejection error is fixed.

Theorem: We assume (i) of lemma to be true and (ii) to be either of the two possibilities (a) and (b). We further assume that the rejection probability is a constant, q. Then the fraction of classes that have to be considered is a constant, when considered as a function of the observed feature, x, if (ii) is as in (a) of lemma and is a nonlinear function of x, if (ii) is as in (b) of lemma.

Proof: From lemma, $p(w|x) \sim N(x, \sigma^2)$. The probability of rejecting a correct class is q. We want the fraction of classes that has to be considered as a function of the feature x so that the rejection error is q.

$$\int_{-\infty}^{\infty} \left[\int_A p(w|x)\, d\mu \right] p(x)\, dx = q \; , \qquad (6)$$

where A is the region over which $p(w|x)$ is sufficiently small so that eqn.(6) is true. Since $p(w|x)$ has a constant variance, σ^2, the region A is nothing but the interval $(-\infty, x-\lambda)U(x+\lambda, \infty)$, where λ is determined by q by the following relation.

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

Substituting the value of $p(w|x)$, the above expres-

$$\text{is} \int_{-\infty}^{\infty} \left[\int_{x-\lambda}^{x+\lambda} (1/\sqrt{2\pi}\sigma) \exp\left\{-[(\mu-x)/\;]^2/2\right\} d\mu \right] p(x)\, dx$$

put $(\mu-x)/\sigma = y$. Then the above expression be-

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

since the terms inside the square brackets are independent of x, we get

$$(1/\sqrt{2\pi}) \int_{-\lambda/\sigma}^{\lambda/\sigma} \exp(-y^2/2)\, dy = 1-q \qquad (7)$$

From eqn,(7), we can conclude that λ is determined by the rejection rate q alone (as σ is a constant).

Let C(x) denote the number of classes picked up when the observed value of feature is x. i.e. C(x) is the set of all those w for which $p(w|x) \geq \lambda$;

or $C(x) = \{(w \mid p(w \mid x) \geq \lambda_0\}$

So, the fraction of classes falling in $C(x) \hat{=} f(x)$

$$= \int_{[p(w \mid x) \geq \lambda_0]} p(w)\, d\mu = \int_{A^c} p(w)\, d\mu. \text{ i.e. } f(x) = \int_{x-\lambda}^{x+\lambda} p(w) d\mu \qquad (8)$$

If p(w) is a constant, i.e. a uniform density, then f(x) is also a constant. If p(w) is any other density, f(x) is a function of the observed feature, x. When p(w) and hence $p(w \mid x)$ is normally distributed,

$f(x) = \int_{x-\lambda}^{x+\lambda} (1/\sqrt{2\pi} K) \exp\{-[(\mu-\alpha)/K]^2\}$. Put $(x-\alpha)/K = y$, then, $f(x) = (1/\sqrt{2\pi}) \int_{y_1}^{y_2} \exp(-y^2/2) dy = \text{erf}(y_2) - \text{erf}(y_1)$, where $y_1 = (x-\lambda-\alpha)/K$,

$$y_2 = (x+\lambda-\alpha)/K \text{ and } \text{erf}(y) = (1/\sqrt{2\pi}) \int_{-\infty}^{y} \exp(-z^2/2) dz \qquad (9)$$

We can observe that the fraction of classes to be considered is a function of λ in both the above cases; i.e., it depends upon the rejection probability that could be tolerated. For a given rejection probability, λ depends upon the variance σ^2 and from eqn.(9), we can conclude that f(x) for a given rejection probability depends on the ratio between K^2 and σ^2; which is heuristically quite satisfying.

In a practical system, the subclass obtained after one stage is still reduced at the second stage by considering one more feature. This procedure is continued till only one class is left. If the features are independent, then the probability of correct recognition after N such stages is $P_{cN} = (1-q_1).(1-q_2)..(1-q_N)$ where q_j, $1 \leq j \leq N$ is the probability of rejection error at the jth stage.

III Speaker Identification System

The sequential method of section II is used for designing and testing an identification scheme for 30 adult male speakers. A 2-stage classifier is used. At the first stage, average pitch over a single short utterance has been used as a feature for identifying a subset of speakers from the 30 speakers. Pitch has been studied extensively as a feature for recognition[9,10]. The present study shows that it can be conveniently used to get a subset of speakers with a high accuracy. As given by eqn.(8), the number of classes that have to be considered is a function of the observed pitch x; being maximum near the mean α of the pitch (over the entire population) and falling off as the observed pitch is farther from the mean. At the second stage the feature used is the long term autocorrelation coefficients over a single code word. In an earlier study[8] the suitability of this feature for small population has been demonstrated. This scheme could be represented as a decision tree, but the tree structure cannot be rigid as with the conventional decision tree classifiers[5]. The classes that have to be considered after observing the pitch vary with the observed value. The structure of the tree is as depicted in fig.1. But the separation of the subclass is very easy as the speakers are selected for further testing if their mean pitch falls on a particular interval of the pitch axis. After the subclass $\{w_i, w_{i+1}, \ldots w_j\}$ is obtained, the scheme in[8] may be used to identify the particular speaker.

The code word selected was "MUM" as the speaker discriminating capacity for the word "MUM" using autocorrelation as the feature vector is well established[8]. The reference pattern for each vector is obtained by collecting 10 repetitions of the word "MUM" computing the normalized autocorrelation and the average pitch for each utterance and taking the average. The data set was of 25 utterances for each speaker and the remaining 15 utterances were 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.1. a subclass of speakers is selected from the 30 speakers. The number in this subclass was variable, 12 being the maximum (at the mean α) and only one speaker at the lower extreme (i.e. in the pitch range of 95-100). These subclass of speakers are used in the second stage in which a final decision is made using a minimum Euclidean distance classifier of[8].

IV Results and Discussions

The results obtained by the above recognition scheme are given in the form of the confusion matrix in the table 1. The results for the 1-stage recognition scheme using autocorrelation alone are also given for comparison directly below the 2-stage classifier result. The zero entries in the confusion matrix are omitted since a 30 x 30 confusion matrix could not be accommodated. The overall performance of the two stage classifier is 87% whereas that of the scheme using MUM alone is only 68% for a population of 30.

A few comments regarding the above sequential procedure are in order. (1) In practice, any system has to be designed and tested over a finite data set, and hence N/L ratio becomes an important parameter in designing the system[11]. (N= number of design samples per class and L is the dimensionality of the feature vector). In a sequential scheme, since a subset of features is used per stage, N/L ratio is quite high at each stage.
(2) We have used only 10 samples per class for designing the system and the remaining 15 samples per class for testing 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 [12], which may give still better results.

References

1. Sarma, V.V.S., and Yegnanarayana, B., "A critical survey of automatic speaker recognition systems", Journal of Comp. Soc. of India, Vol.6, No.1, Dec. 1975, pp. 9-19.
2. Venugopal, D., and Sarma, V.V.S., "Performance evaluation of automatic speaker recognition schemes", IEEE 1977 Conf. on ASSP, 1977, pp. 780-783.
3. Rosenberg, A.E., "Automatic speaker verification: a review", Proceedings IEEE, Vol.64, April 1976, pp. 475-487.
4. Kashyap, R.L., "Speaker recognition from an unknown utterance and speaker-speech interaction", IEEE Trans. Acoust. Speech and Sig. Proc., Vol. ASSP-24, No.6, Dec.1976, pp.481-488.
5. Kulkarni, A.V., Optimal and heuristic synthesis of hierarchical classifier, Ph.D. thesis, University of Maryland, Maryland, 1976.
6. Fu, K.S., Sequential Methods in Pattern Recognition, Academic Press, 1968.
7. Duda, R.O., and Hart, P.E., Pattern Classification and Scene Analysis, John Wiley, 1973, p. 52.
8. Yegnanarayana, B., Sarma, V.V.S., and Venugopal, D., "Studies on speaker recognition using a Fourier Analyser System", IEEE 1977 Conf. ASSP, 1977, pp. 776-779.
9. Atal, B.S., Automatic speaker recognition based on Pitch contours, Ph.D. thesis, Polytechnic Inst. of Brooklyn, 1968.
10. Atkinson, J.E., "Inter- and intraspeaker variability in fundamental voice frequency", J. Accoust. Soc. Am., Vol. 60, No.2, Aug. 1976, pp. 440-445.
11. Sarma, V.V.S. and Venugopal, D., "Performance evaluation of automatic speaker verification systems", IEEE Tr. ASSP, Vol. ASSP-25, No. 3, June 1977, pp. 264-266.
12. Toussaint, G.T., "Bibliography on estimation of misclassification", IEEE Tr. Inf. Th., Vol. IT-20, July 1974, pp. 472-479.

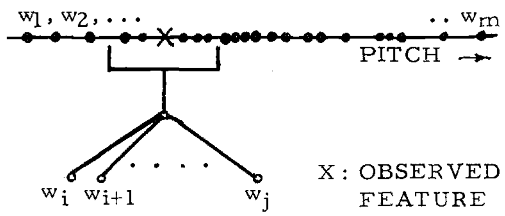

Fig. 1. Decision tree for sequential classifier

TABLE 1: Confusion matrices

True speakers	No. of times recognised as	Percentage accuracy
1	12-1, 1-6, 1-7, 1-9	80
	12-1, 2-7, 1-6	80
2	15-2,	100
	11-2, 4-3	73
3	15-3	100
	6-3, 8-8, 1-1	40
4	14-4, 1-1	93
	10-4, 3-29, 1-1, 1-25	67
5	13-5, 2-3	87
	10-5, 5-3	67
6	13-6, 1-9, 1-25	87
	11-6, 3-25, 1-9	73
7	11-7, 3-6, 1-9	73
	3-7, 5-6, 5-9, 1-2, 1-15	20
8	15-8	100
	11-8, 3-1, 1-7	73
9	13-9, 1-4, 1-7	87
	9-9, 4-25, 1-7, 1-24	60
10	15-15	100
	12-15, 2-14, 1-26	80
11	15-15	100
	15-15	100
12	15-12	100
	14-12, 1-17	93
13	15-13	100
	14-13, 1-9	93
14	15-14	100
	15-14	100
15	10-15, 5-2	67
	8-15, 7-2	53
16	11-16, 2-12, 1-13, 1-17	73
	10-16, 2-12, 2-13, 1-17	67
17	9-17, 6-25	60
	7-17, 8-27	47
18	15-18	100
	15-18	100
19	15-19	100
	14-19, 1-1	93
20	13-20, 1-15, 1-27	87
	3-20, 8-9, 2-15, 1-28	20
21	11-15, 2-2, 2-9	73
	8-15, 3-25, 3-29, 1-2	53
22	7-22, 7-15, 1-2	47
	3-22, 7-15, 3-27, 1-2, 1-20	40
23	7-23, 5-27, 2-17, 1-18	47
	2-23, 5-26, 5-28, 2-17, 1-18	13
24	15-24	100
	9-24, 6-25	60
25	11-25, 2-19, 2-29	73
	9-25, 3-19, 3-29	60
26	15-26	100
	14-26, 1-14	93
27	11-27, 3-17, 1-23	73
	10-27, 3-17, 2-23	67
28	15-28	100
	15-28	100
29	15-29	100
	12-29, 2-9, 1-19	80
30	15-30	100
	15-30	100
Total performance	391/450	87
	307/450	68