THE APPLICATION OF PROBABILITY DENSITY ESTIMATION TO TEXT-INDEPENDENT SPEAKER IDENTIFICATION

R. Schwartz, S. Roucos, M. Berouti

Bolt Beranek and Newman Inc.

Cambridge, MA 02238

ABSTRACT

Most text-independent speaker identification methods to date depend on the use of some distance metric for classification. In this paper we develop the use of probability density function (pdf) estimation for text-independent speaker identification.

We compare the performance of two parametric and one non-parametric pdf estimation methods to one distance classification method that uses the Mahalanobis distance. Under all conditions tested, the pdf estimation methods performed substantially better than the Mahalanobis distance method. The best method is a non-parametric pdf estimation method.

INTRODUCTION

In text-independent speaker identification, one is given an utterance from an unknown speaker and the task is to identify this unknown speaker as being one of a set of known speakers. The speakers are represented by means of labeled data, which we call the training set of each speaker. An input utterance from an unknown speaker is called the test set. The training and test sets consist of vectors which are points in a multidimensional space. These vectors, referred to as observations or feature vectors, are parametric representations of the short-time spectral envelope of the speech signal. For each speaker, the collection of vectors can be described by a multivariate probability density function (pdf) [1]. When the pdf's generated by different speakers are concentrated at different locations in space, it becomes possible to distinguish among the speakers, without regard to text, by applying standard statistical pattern recognition procedures.

In this paper, we present an approach for speaker recognition based on probability density estimation and statistical pattern recognition. We consider several variations on this approach, and present the results of a study comparing these probabilistic methods to a commonly used distance-based classification method.

SPEAKER RECOGNITION METHODS

1. Minimum Distance Classifiers

Minimum distance classifiers use a distance metric to compare the test passage to each of the training passages and classify the test to the class (speaker) with the smallest distance. One such classifier is described below.

Mahalanobis distance from test mean to training mean (MD). Given a set of observations (feature vectors denoted by $\underline{x}$) in a test passage, we compare the mean vector of the test, $\bar{\underline{x}}$, to the mean vector of the training data of each speaker, $\underline{u}_i$. The distance metric used is the Mahalanobis distance:

$$d_M(\bar{\underline{x}},\underline{u}_i) = [(\bar{\underline{x}}-\underline{u}_i)' \, C_i^{-1}(\bar{\underline{x}}-\underline{u}_i)]^{1/2}, \qquad (1)$$

where ' denotes transpose, and C_i^{-1} is the inverse of the covariance matrix of speaker i. This method has been used and tested extensively in speaker recognition [1-4].

By averaging the observations, it is hoped that the phonetic variability is reduced and what remains characterizes only the speaker. However, this method ignores the higher order statistics of the test passage and does not make full use of all the available information.

A different approach is to compute the weighted Euclidean distance from each test vector without any averaging to each of the speaker means, where the weighting matrix is C_i^{-1} [5]. Wrench [6] has shown experimentally that this approach gives a poorer performance than that of the Mahalanobis distance method described above.

2. Probabilistic Classifiers

Given the observation vectors, $\underline{x}$, from an unknown speaker, we compute the conditional joint probability of observing the feature vectors in the unknown test passage under the assumption that the unknown passage was spoken by speaker i. Successive vectors are assumed independent and the conditional probabilities of the vectors are multiplied (the log probabilities are averaged) to obtain the a posteriori probability of speaker

CH 1746-7/82/0000 - 1649 $ 00.75 © 1982 IEEE

i. The classification rule with minimum probability of error chooses the speaker with maximum a posteriori probability (Bayes classifier). It is well known that, under proper conditions, probabilistic classifiers perform better than minimum distance classifiers. Below we present three methods to estimate the pdf of each speaker, as required by the probabilistic approach.

Gaussian Probability Density Function Estimation (GPDF). In the first method we model the conditional pdf of the observation vectors as Gaussian:

$$p(\underline{x}|w_i)=\frac{1}{(2\pi)^{p/2}|C_i|^{1/2}}\exp\{-\frac{1}{2}(\underline{x}-\underline{u}_i)'C_i^{-1}(\underline{x}-\underline{u}_i)\} \quad (2)$$

where $|C_i|$ is the determinant of the covariance matrix, p is the number of dimensions, and w_i represents the ith speaker. For each input vector x_j, we compute a score given by

$$S_j = \ell n \frac{p(\underline{x}_j|w_i)}{p(\underline{x})} \quad (3)$$

and average the scores.

Robust Probability Estimates - "Score Clipping" - (GPDF+C). Since the Gaussian distribution falls very quickly, the scores of occasional observations may be extremely low and can incorrectly change the overall classification. This happens when the training data contains only a subset of the possible sounds and the variance terms are underestimated. To alleviate this problem we apply a "soft clipping" function to the log-likelihood score, S_j, for the observation for each speaker, given by:

$$\hat{S}_j = b+1 - (b+1-S_j)^{1/2} \quad \text{for } S\leq b$$
$$= S_j \quad \text{for } S<b \quad (4)$$

The breakpoint, b, used in all cases was -3. We found that this technique greatly improves the performance of the GPDF method. We note here that the use of the Gaussian model with some form of score clipping has been reported in [7].

Non-Parametric PDF Estimation (NPDF). In the third method, we investigated the use of a non-parametric pdf estimation technique. We estimate the conditional pdf by a variation on the k-nearest neighbor (kNN) estimation technique. Given an unknown test observation the algorithm finds the nearest k samples from each class (speaker). Then, the conditional probability for $\underline{x}$, given speaker i, is given by:

$$\hat{p}(\underline{x}|w_i) = W\frac{k}{N_i V_i} \quad (5)$$

where N_i is the total number of training samples for speaker i, V_i is the volume of the region that encloses the kth point, and W is a weight between 0 and 1. An appropriate value of k is chosen depending on the length of the training passage. After finding the k-nearest samples, W is computed based on the weighted Euclidean distances between the unknown test vector and the k samples. The weight is given by

$$W = \sum_{i=1}^{k-1}[\cos(\pi d_i/d_k)+1]/2 \quad (6)$$

This weighting allows the method to adjust for the fine detail of the density within the kNN region.

The non-parametric pdf estimation method has the advantage that it is not dependent on a particular model for the distribution (e.g., Gaussian). However, it is generally believed that it may not be able to make use of a very large number of dimensions.

EXPERIMENT

We performed an experiment to compare the performance of the above described classifiers under a number of different conditions. The database used consists of the speech of 21 male speakers. Each speaker read some speech material for one minute. The minute of speech was arbitrarily broken up into training and test passages - including any silences (typically 20-45%). The speech was bandlimited to the range 300-3600 Hz and sampled at 7200 Hz.

For this study we restricted the scope of available features to LPC reflection coefficients, or the corresponding log-area-ratios (LAR's), or a set of real cepstral coefficients. In all cases a 20 ms analysis window was used with an overlap of 10 ms between consecutive windows. In order to test the behavior of the recognition algorithms under noise conditions, we added various levels of white noise to the input speech. The original speech had a noise level that was roughly equivalent to a 45 dB SNR (signal-to-noise ratio).

After a preliminary study, we decided to use a sub-population filter based on a simple energy threshold to determine which frames of speech are to be used. We found that, for noisy speech, recognition performance improves when we use this subpopulation filter instead of a voicing filter.

We present here the performance results of four algorithms. The algorithms and the symbols used in figures are:

Symbol

- o (1) Mahalanobis Distance of Test Mean (MD)
- □ (2) Gaussian PDF Estimation (GPDF)
- ⊞ (3) Gaussian PDF Estimation + clipping (GPDF+C)
- △ (4) Non-parametric PDF Estimation (NPDF)

The variables investigated in this study are:

- o Different spectral models.
- o Number of parameters (10 to 18).
- o Signal-to-noise ratio (SNR) (45,20,15 dB).
- o Different SNR between training and test.
- o Duration of training and test data.
- o Channel equalization.

A total of 10 seconds was always available for

testing. Therefore, in the case of 2-second tests, there were 5 test passages for each speaker or a total of 105 tests for all 21 speakers. It must be noted here again that training and test durations are measured in total elapsed time (and include silences). The subpopulation filter was the same for all methods and was allowed to vary only as a function of SNR.

RESULTS

First, we considered the three different spectral models. We found that for clean speech, all three models were within 1% in performance, with the LAR parameters consistently slightly better. However, for noisy speech, the cepstral parameters were significantly (10%) worse. Therefore, except for the different channels case, all results given are for LAR parameters.

Figure 1 shows the performance for the reference test condition: Clean speech (SNR=45 dB), LAR parameters to represent the spectrum, 10-second training duration and 2-second test duration. As can be seen in Fig. 1, the first three methods improve significantly with more spectral features. Mahalanobis distance (MD) is the worst method. Gaussian PDF estimation (GPDF) ranges from 80% to 90%. The addition of clipping (GPDF+C) improves its performance further, particularly with 18 parameters. For 10 coefficients, non-parametric PDF estimation (NPDF) performs much better than the other methods (95%). All of these trends, including the relative ranking of the algorithms, as a function of the number of parameters, hold true for most of the test conditions.

Figure 2 shows the effect of increasing the training duration when all algorithms use only 10 parameters. The results are shown for both clean speech (45 dB SNR) and noisy speech (15 dB SNR). It is interesting to note that the performance of the NPDF method with 5-second training, on noisy speech, is better than the performance of the MD method with 20-second training on clean speech.

Another interesting observation to be made from these results is that the pdf of parameter vectors for a speaker is not modeled well by a Gaussian pdf. This fact explains why the NPDF method performs significantly better than the methods based on a Gaussian model, even for the cases of 5-second training or noisy data. It also explains why clipping improves the GPDF method significantly.

In Fig. 3 the number of parameters used is the number that yields the optimum performance for each method. For the MD, GPDF and GPDF+C methods, this optimum number is typically 18 (see Fig. 1). These parametric methods benefit from the additional information contained in the high order coefficients. In contrast, the nonparametric approach requires substantially more training data as the number of dimensions is increased. Even with 20-second training, the NPDF method cannot make effective use of more than 10 dimensions.

The effect of increasing the test duration (figure not shown) from 2 s to 4 s is such that, in our experiments, the NPDF method with 10 LAR's and the GPDF+C method with 18 LAR's achieve 100% accuracy on clean speech and 97% on noisy speech (SNR=15 dB).

The performance for several different noise conditions is presented in Fig. 4. On the left of the figure we compare the results for three different SNR's - with the SNR being the same in training and test. As expected, performance drops somewhat with the addition of noise.

In the middle section of this figure we show two conditions of different SNR between training and test. In the first set of results the training has a SNR of 15 dB, and the test has a SNR of 20 dB. The second set of results illustrates the reverse case. In both cases, the MD method does very poorly. This is because, with different noise levels, the training and test for the same speaker have significantly different means (compared to the distance between means of different speakers). As can be seen, if the test has more noise than the training, performance degrades much more than in the reverse case. These results indicate that it is essential to equalize the noise level in the training and test passages, either by noise reduction or by other methods.

A real world problem is that of having different communications channels for different speech segments, which degrades the performance of the recognition algorithms. We implemented a solution to this problem similar to that given in [1]. Assuming the frequency response of the channel to be a stationary multiplicative function in the spectral domain, it can be cancelled by subtracting the average long-term cepstrum of each training and test passage from all the cepstrum vectors of that passage.

The performance of the recognition algorithms under this condition is shown in the rightmost part of Fig. 4. The MD method cannot be used here since it uses the training and test means, which are now all zero. The results indicate that this condition is about as severe as having different noise levels in the training and test sets.

Finally, the above algorithms were also tried in a speaker verification paradigm. The ranking of the algorithms remained the same. In general, the verification error rates were half the error rates given above for the identification tasks.

SUMMARY

In this paper, we presented a probabilistic approach to text-independent speaker recognition based on pdf estimation. We have shown that the methods that use pdf estimation for classification yield recognizers that perform substantially better, under virtually all conditions, than the Mahalanobis distance method. In addition, nonparametric pdf estimation achieves better performance than the parametric methods when the

methods are constrained to use only 10 parameters. When allowed to use up to 18 parameters, the performance of the modified Gaussian pdf estimation method improves dramatically to equal that of the nonparametric method.

ACKNOWLEDGMENTS

The authors wish to thank Dr. John Makhoul for his support throughout this research.

REFERENCES

(1) B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Am, Vol. 55, pp. 1304-1312, June 1974.

(2) J.D. Markel, B.T. Oshika, and A.H. Gray, Jr., "Long Term Feature Averaging for Speaker Recognition," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-25, pp. 330-337, 1977.

(3) M.R. Sambur, "Speaker Recognition using Orthogonal Linear Prediction," IEEE Trans. on ASSP, Vol. ASSP-24, pp. 283-289, 1976.

(4) R.E. Wohlford, E.H. Wrench, Jr., and B.P. Landell, Automatic Speaker Recognition Comparison Study, Final Report, Contract 78-E27505-00, ITT Defense Communications Division, San Diego, CA, 1979.

(5) L.L. Pfeifer, "New Techniques for Text-Independent Speaker Identification," IEEE International Conf. Acoustics, Speech and Signal Processing, pp. 283-286, 1978.

(6) E.H. Wrench, Jr., Speaker Authentication Operational Test and Evaluation, Rome Air Development Center, Griffiss AFB, NY, Final Technical Report RADC-TR-80-64, 1980.

(7) H. Matsumoto and T. Nimura, "Text-Independent Speaker Identification Based on Piecewise Canonical Discriminant Analysis," IEEE International Conf. Acoustics, Speech and Signal Processing, pp. 291-294, 1978.

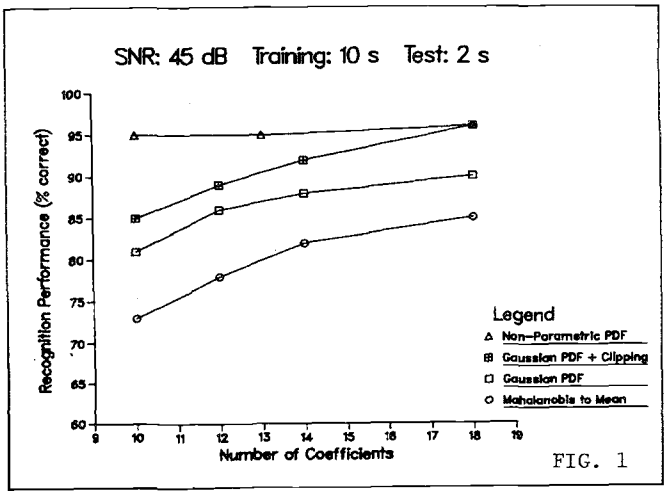

FIG. 1

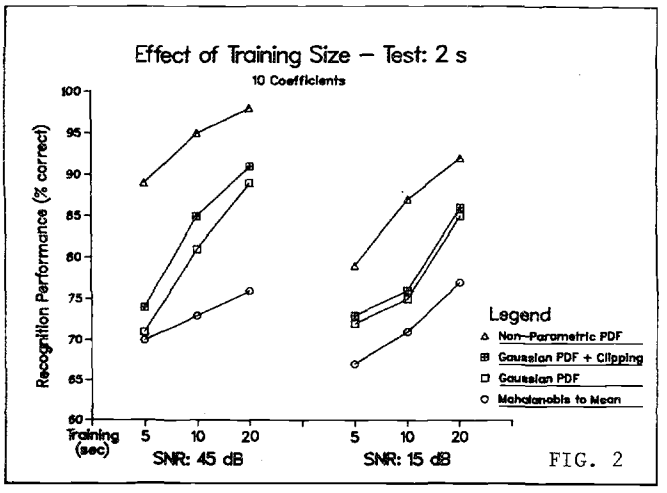

FIG. 2

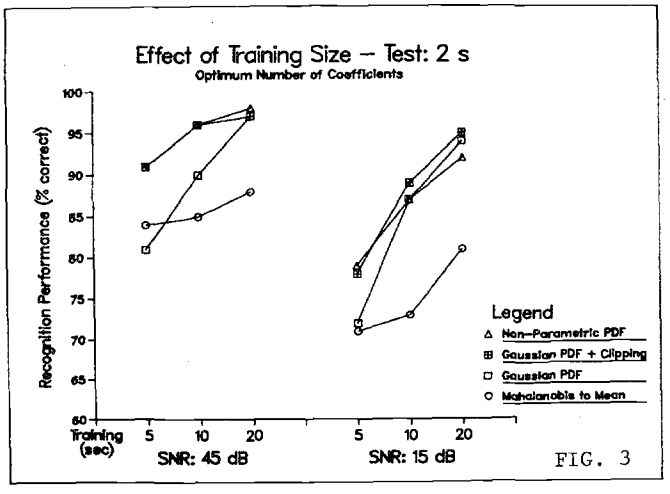

FIG. 3

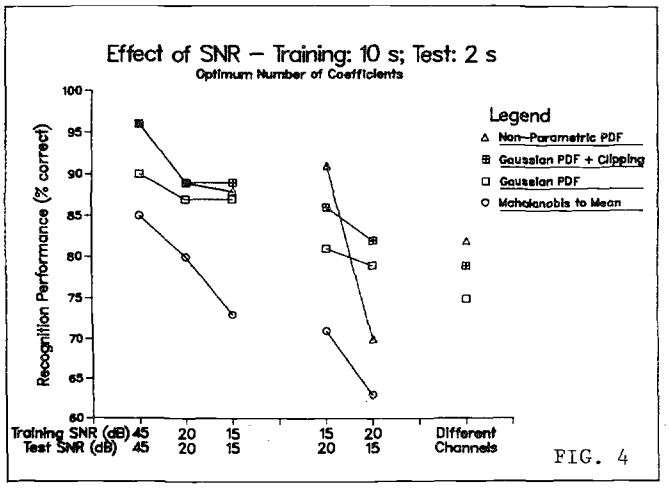

FIG. 4