S5.5

A WEIGHTED DISTANCE MEASURE BASED ON THE FINE STRUCTURE OF FEATURE SPACE : APPLICATION TO SPEAKER RECOGNITION

Wang Ren-hua, He Lin-shen and Hiroya Fujisaki

Department of Radio Electronics
University of Science & Technology of China
Hefei, Anhui, China

ABSTRACT

In this paper a novel weighted cepstral distance measure is proposed and is tested in a speaker recognition system using speaker-based VQ approach. Based on the fine structure of the feature vector space we define a statistically optimized distance measure with weights equal to the partition-normalized inverse variance of cepstral coefficients. The weights can be adjusted individually for each partition and each component of the feature vector across all codebooks (speakers). Experiments on a 50 speaker database showed that the suggested weighted cepstral distance measure works substantially better than the Euclidean cepstral distance or the inverse variance weighted cepstral distance. An accuracy of about 99% was achieved using a 16 level codebook in speaker verification.

1. INTRODUCTION

Automatic speaker recognition has long been an interesting and challenging problem for speech researchers. Depending on the nature of the final task, the problem can be classified into two different ones: automatic speaker verification (ASV) and automatic speaker identification (ASI). The input speech material to be used for ASV/I can be either text-dependent or text-independent.

Effective feature extraction is most important in ASV/I. Among all varieties of LPC parameters (e.g., reflection coefficients, log-area ratios, cepstrum, etc), the cepstral representation has been suggested as best suited for ASV/I [1]. Acceptable results have been demonstrated through template matching of 18-dimensional cepstral vectors [2]. Furthermore, the performance improvement can be obtained if the cepstral coefficients are weighted appropriately [3]. A statistically weighted distance measure, with weights equal to the inverse variance of the cepstral coefficients, has been proposed both for speech recognition and for ASV/I [2][4].

A conventional approach to ASV/I is vector quantization (VQ). VQ is a source coding technique by which the source vector is coded as one of a prestored set of codewords in a codebook. The primary advantage of VQ for ASV/I lies in that, the similarity between utterances can be determined by codebook searches. For each speaker a VQ codebook is constructed from a training sequence composed of individual feature vectors through a clustering algorithm. Each test utterance is identified with the speaker whose codebook yields the lowest distortion [5].

The VQ approach involves the partitioning of a feature vector space. We observed that the statistical distributions of cepstral coefficients were different for each partition of the feature vector space. Therefore if we introduce more statistical characteristics of the feature space into the VQ codebook, the achieved ASV/I accuracy would be higher. It is necessary to define a statistically optimized distance measure for the speaker-based VQ codebook generation and for the minimum distance classification.

2. PROPOSAL FOR A STATISTICALLY OPTIMIZED DISTANCE MEASURE

2.1 The Partition-Normalized Distance Measure(PNDM)

A distance measure which has been widely used in VQ-based ASV/I is the Mahalanobis distance, d_M, defined as follows:

$$d_M(X,C) = (X-C)^T V^{-1}(X-C) \qquad (1)$$

where $X = (x_1, x_2, \ldots x_K)$ is a K-dimensional feature vector which is composed of the cepstral coefficients obtained from a test utterance, $C = (c_1, c_2, \ldots c_K)$ is one of the codewords of the codebook, and V is the covariance matrix of the feature vector. The measure d_M is a covariance weighted distance measure which can be used for clustering feature vectors as well as for recognition.

CH2847-2/90/0000-0273 $1.00 © 1990 IEEE

Since the off-diagonal elements of the covariance matrix V are relatively small as compared with the diagonal ones, we shall use the following weighted cepstral distance measure d_w:

$$d_w(X,C) = (X-C)^T W(X-C) = \sum_{k=1}^{K} w_k(x_k-c_k)^2 \quad (2)$$

where $W=\mathrm{diag}\{w_1,w_2,\ldots,w_K\}$ is a K-dimensional diagonal matrix and w_k is the reciprocal of the k-th diagonal element v_{kk} of the covariance matrix V.

Conventionally, the weighting matrix W is assumed to be identical for all partitions of the feature space across all speakers. However, since the statistical distributions of cepstral coefficients are different among partitions of the feature vector space, we propose that the weighting matrix W should be adjusted individually for each partition as well as for each feature component across all speakers. The weighted distance measure d_w for each partition of the feature vector space should be defined as follows:

$$d_P(X,C_{ij}) = (X-C_{ij})^T W_{ij}(X-C_{ij}) = \sum_{k=1}^{K} w_{ijk}(x_k-c_{ijk})^2$$
$$i=1,2,\ldots I;\ j=1,2,\ldots J \quad (3)$$

where $C_{ij}=(c_{ij1},c_{ij2},\ldots,c_{ijK})$ is the centroid of the j-th partition P_{ij} of the i-th speaker's feature space, is called the j-th codeword of the i-th codebook. $W_{ij}=\mathrm{diag}\{w_{ij1},w_{ij2},\ldots,w_{ijK}\}$ is a diagonal matrix which is regarded as the weighting matrix for the partition P_{ij}. The weights w_{ijk} is selected to optimize the distance measure for each individual partition of the feature space. We shall call the distance d_P as the partition-normalized distance measure (PNDM).

2.2 Theoretical Framework

In this section we will determine the optimized weighting matrix W_{ij}. Let $X_{ij}=(x_{ij1},x_{ij2},\ldots,x_{ijK})$ represent the feature vector in the partition P_{ij}. The component of codewords c_{ijk} and the variance of coefficients v_{ijk} in each partition are expressed as

$$c_{ijk} = E[x_{ijk}]$$
$$v_{ijk} = E[(x_{ijk}-c_{ijk})^2]$$
$$i=1,2,\ldots I;\ j=1,2,\ldots J;\ k=1,2,\ldots K \quad (4)$$

where E[ ] indicates the expectation over the partition P_{ij}. Any test feature vector X can be attributed to one of the partitions (actually it is coded as one of the codewords in some codebook) by finding the codeword with the minimum distance from the test feature vector. We define the average distortion D over all test feature vectors by the following equation:

$$D = (1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J} E[d_P(X_{ij},C_{ij})]$$
$$= (1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J} E[\sum_{k=1}^{K} w_{ijk}(x_{ijk}-c_{ijk})^2]$$
$$= (1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w_{ijk}E[(x_{ijk}-c_{ijk})^2]$$
$$= (1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w_{ijk}v_{ijk} \quad (5)$$

The optimized matrix W_{ij} can be obtained by minimizing the average distortion D.

From the definition (5) the following four points are quite evident :

(1) The weight w_{ijk} should be in proportion to the inverse variance of coefficients.

(2) If the variances of coefficients in the partition P_{ij} are identical for all k, the weighting matrix W_{ij} for the partition becomes a unit matrix ($w_{ijk}=1$; k=1,2,...K).

(3) Since the weight w_{ijk} varies depending on individual speakers, some process of normalization is necessary in order to obtain meaningful distance measure.

(4) A simple relation between the weight and the variance is preferable from the viewpoint of computational costs.

Considering all of these, we shall introduce the following constrains on the w_{ijk} (k=1,2,...K):

$$\prod_{k=1}^{K} w_{ijk} = 1;\quad i=1,2,\ldots I;\ j=1,2,\ldots J \quad (6)$$

i.e. the geometric mean of the weights for each partition is equal to 1. Equations (5) and (6) form a linear programming problem, which can be solved using Lagrange's multiplier method in the following way.

Assume

$$D=(1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J}\{\sum_{k=1}^{K} w_{ijk}v_{ijk}-\lambda_{ij}(\prod_{k=1}^{K} w_{ijk}-1)\},$$

letting (7)

$$\partial D/\partial w_{ijk} = 0, \quad \text{and} \quad \partial D/\partial \lambda_{ij} = 0,$$
$$i=1,2,\ldots I;\ j=1,2,\ldots J;\ k=1,2,\ldots K. \quad (8)$$

From Equation (7) we have

$$v_{ijk} - \lambda_{ij}\prod_{\substack{m=1\\ m\neq k}}^{K} w_{ijm} = 0,$$

and $$\prod_{m=1}^{K} w_{ijm} - 1 = 0,$$
$$i=1,2,\ldots I;\ j=1,2,\ldots J;\ k=1,2,\ldots K. \quad (9)$$

Finally we obtain

$$w_{ijk} = \lambda_{ij}/v_{ijk},$$

and $$\lambda_{ij} = \prod_{m=1}^{K} v_{ijm},$$
$$i=1,2,\ldots I;\ j=1,2,\ldots J;\ k=1,2,\ldots K. \quad (10)$$

This calculation indicates that the PNDM d_P is essentially a weighted Euclidean distance measure where the weights for each partition are equal to the inverse variance of the cepstral coefficients in the partition and normalized by the geometric mean of the variances of coefficients in this

partition. The weighting matrix depends on the statistical distributions of cepstral coefficients. The PNDM is a statistically optimized measure for the VQ codebook generation and feature vector quantization.

3. EXPERIMENTAL COMPARISON OF VARIOUS DISTANCE MEASURES

3.1 A PC-based Automatic Speaker Recognition System

The automatic speaker recognition system used in the experiments is designed as a single board based around a Texas Instruments TMS32010 digital signal processor. Besides the DSP chip, the hardware includes a preprocessor for the speech signal, a 12-bit A/D converter with 10kHz sampling frequency, a low-pass filter and an 8-bit D/A converter. An IBM personal computer serves as the host for the system and functions as a user interface. Figure 1 shows the configuration of the proposed system for automatic speaker recognition. This fast and cost effective system is capable of performing speaker recognition in real time.

3.2 Speech Database

To evaluate the performance of the proposed weighted cepstral distance measure, a 50 speaker database(30 males and 20 females) was used in the experiments. The database contains three types of vocabularies:

(1) Chinese names (usually 3 syllables): each speaker utters his/her name ten times and the names of others three times, respectively.

(2) Isolated digits: each speaker utters the isolated digits 0 to 9 in Chinese ten times.

(3) 4-digit sequences: each speaker utters twenty 4-digit sequences five times.

The database was prepared carefully and used in the experiments, especially for the purpose of comparison of various distance measures.

3.3 Statistical Characteristics of the Cepstral Coefficients

Figure 2 shows statistical distributions of the cepstral coefficients obtained with the database. The Fig.2(a) is the variance distributions of cepstral coefficients over all 50 speakers, and the next five are also the variance distributions of cepstral coefficients but corresponding to five individual speakers in the fifty speakers. Figure 3 shows the variance distributions of cepstral coefficients in six of the sixteen partitions of the feature space from one speaker. It can be seen from Fig.2 and Fig.3 that there is a large difference in the variance distributions among the partitions of

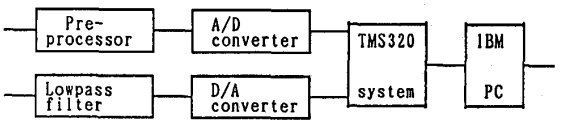

Fig.1. Block diagram of the proposed system for ASV/I.

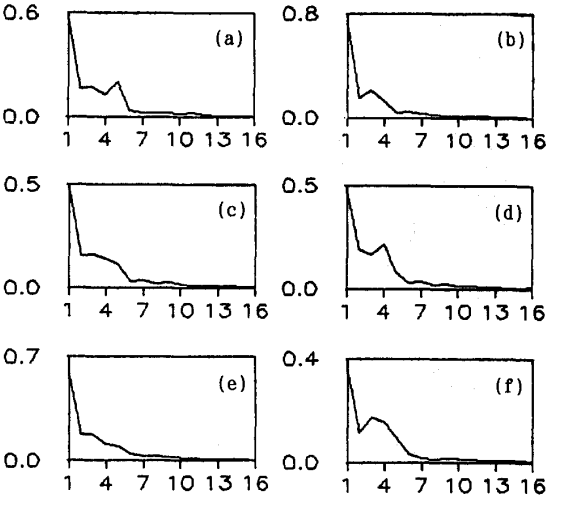

Fig.2. Variance distributions of cepstral coefficients from 50 speakers and five speakers in the 50 speakers.
X axis: Cepstral coefficients index
Y axis: Variance of cepstral coefficients

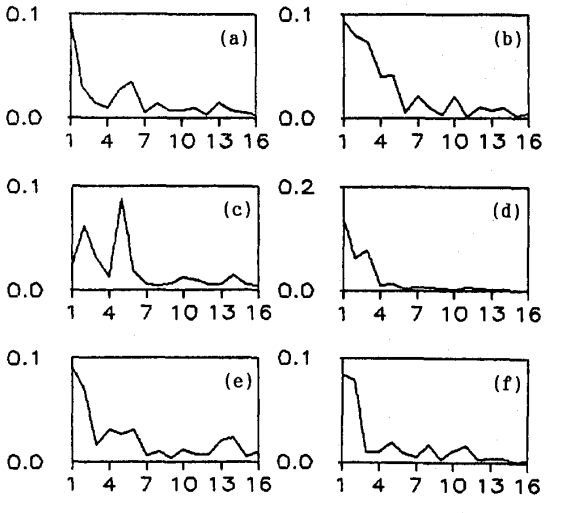

Fig.3. Variance distributions of cepstral coefficients in six of partitions of the feature space from one speaker.

the feature vector space as well as among the cepstral coefficients. The great variety in the distributions shown in these figures testifies the necessity to define a statistically optimized distance measure for speaker-based VQ codebook generation and the minimum distance classification, so that the VQ codebooks would contain the information on the overall statistical characteristics of the feature space.

3.4 Results of Speaker Verification Experiment

The experiments on ASV were performed using the speech database. The test utterance for each trial consists of a Chinese name and a 4-digit sequence. The cepstral coefficients are first obtained from A standard linear prediction analysis every 15 msecs. Thus the test utterance is represented by a set of feature vectors, each is composed of 10 cepstral coefficients. For each speaker a VQ codebook is created by training on a set of corresponding feature vectors. The results of this experiment are shown in Table 1 in the form of verification rate for different distance measures and different VQ levels. Using the proposed weighted cepstral distance d_{pcep} the verification accuracies of 97.4% and 99.1% were achieved for 8-level codebook and 16-level codebook, respectively. For the sake of comparison, the experimental results using the Euclidean cepstral distance d_{cep} and the inverse variance weighted cepstral distance d_{wcep} are also shown in Table 1. The table indicates that the partition-normalized cepstral distance measure provides substantially better results than the Euclidean cepstral distance measure and the inverse variance weighted cepstral distance measure.

Level of VQ	4	8	16
d_{cep}	89.2%	93.6%	96.4%
d_{wcep}	91.3%	95.7%	98.2%
d_{pcep}	94.2%	97.4%	99.1%

Tabel 1. Verification rate for different distance measures with different levels of VQ.

4. DISCUSSION AND CONCLUSION

In this paper a new weighting function was proposed for the weighted cepstral distance measure in a speaker recognition system using speaker-based VQ approach. The most significant characteristics of the suggested PNDM are summarized as follows:

(1) Since the weighting function reflects the fine structure of the feature space, the corresponding partition-normalized distance measure is statistically optimized.

(2) By the weighting, each individual cepstral coefficient is variance-equalized.

(3) The weights in the PNDM can be adjusted individually for each partition and for each feature component across all codebooks.

(4) Based on the PNDM the features of different kinds can be combined directly to measure distortion without normalization preprocessing. Suppose that the feature space S is linearly scaled to S' by

$$x'_k = t_k x_k , \quad k=1,2,\ldots K$$

where $X' \in S'$, $X \in S$ and t_k is a constant. From Equations (4) and (7), we have

$$c'_{ijk} = t_k c_{ijk}, \qquad v'_{ijk} = t_k v_{ijk},$$

and

$$\lambda'_{ij} = \left(\sum_{k=1}^{K} t_k^2\right)\lambda_{ij} = T\lambda_{ij}, \qquad (11)$$

where $T=\left(\sum_{k=1}^{K} t_k^2\right)$ is still a constant. The relation between the distance d_P for the feature space S' and the distance d_P for S can be described by the following equation

$$d_P(X', C'_{ij}) = T d_P(X, C_{ij}). \qquad (12)$$

Thus it makes no difference to clustering vectors or to recognition that the distance measure is scaled up or down by a constant.

Preliminary experiments indicate that the PNDM is much effective for clustering feature vectors and for improving the performance in VQ-based speaker recognition. The suggested PNDM is applicable to both text-dependent or text-independent speaker recognition. Finally, we expect that the PNDM is equally effective for VQ-based speech recognition.

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, NO.6, pp.1304-1312: June 1974.

[2] S.Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. ASSP, ASSP-29, pp.254-272: 1981.

[3] G.Velius, "Variants of Cepstrum Based Speaker Identify Verification," IEEE ICASSP88, pp.583-586: 1988.

[4] Y.Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," IEEE ICASSP86, pp.761-764: 1986.

[5] D.K.Burton, "Text-dependent Speaker Verification Using Vector Quantization Source Coding," IEEE Trans. ASSP, ASSP-35, pp.133-143: Feb. 1987.