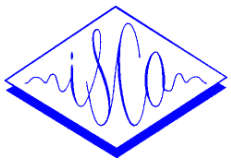

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEAKER RECOGNITION WITH DISCRIMINATIVE SPEAKER VQ MODELS

Kai Tat Ng† *Jian Su*†‡ *Bingzheng Xu*‡
email: EEKTNG@cityu.edu.hk
†Department of Electronic Engineering, City University of Hong Kong, Hong Kong
‡Institute of Radio and Automation, South China University of Technology, China

ABSTRACT

In this paper, we propose a discriminative VQ model for speaker recognition. The VQ training algorithm is developed within the framework of gradient descent learning. Unlike the conventional VQ scheme, which considers only the minimum distance to a speaker codebook, the new algorithm takes account of the distances to all competing classes and all codewords in a speaker codebook, and aims directly at minimizing the recognition error. Two sets of speaker recognition experiments based on the conventional learning scheme and the DVQ are conducted respectively on a database of 200 French speakers. We obtained 1.75% performance improvement in speaker recognition and 0.45% in verification over the conventional VQ algorithm.

1 INTRODUCTION

Speaker recognition with VQ has been well investigated [1] in some earlier studies. For a long time, speaker dependent VQ codebooks have been trained to represent well the distribution of speech features for each speaker. However, we know that the human brain discriminates speakers not only by their individual utterances but also by some critical features among speakers, especially when speakers are acoustically similar. Therefore, in this case, we should focus more on the discriminative features among speakers rather than the speech features themself[2].

In section 2, the conventional VQ-based speaker recognition scheme is briefed. In section 3, we propose a training algorithm of discriminative VQ (DVQ) for speaker recognition. The mathematical foundations are the misclassification measure formulation and the gradient descent algorithm, which stem from linear discriminant analysis [3] and have been introduced into speech recognition recently [4]. Speaker recognition experiments using the conventional learning scheme and the DVQ are conducted respectively for performance evaluations. Finally, the results along with discussions are reported.

2 VQ-BASED SPEAKER RECOGNITION

The VQ-based speaker recognition approach has been proposed with different representations of speech spectra. A number of experiments have demonstrated its successes and limitations[1].

As reported in some earlier studies, in a VQ-based speaker recognizer, each candidate speaker is represented by a codebook with $B(\geq 1)$ codewords.

During training, we conventionally use algorithms such as LBG[5] to generate a VQ codebook for each speaker. To be more specific, all training vectors of each speaker are assigned into B clusters, then the cluster centers are found by averaging all training vectors in each cluster or selecting the pseudo centroid for each cluster so that the overall distortion is reduced, the above two step repeat until the relative overall distortion is lower than a prescribed threshold.

During recognition test, an unknown input vector is compared to the reference codewords to produce B distance scores for each candidate speaker, then we come up with a minimum distance score for each speaker. After the overall distances between the test utterances and all speakers are evaluated, the speaker with the minimum average distance is identified as the recognition result. As for speaker verification, we accept the speaker if the normalized distance to his/her codebook is not greater than a threshold.

Now we go into two aspects of the above approach. First, the codebook of each speaker are produced by its own training data, so it may be designed to represent well the distribution of speach features for each speaker, but not to discriminate well the different characteristics among different speakers. A discriminative function which incorporates the parameters of competing speakers will be a solution to this problem and may result in better classification. Second, only the minimum distance is considered in the VQ decoding, which is discontinuous with respect to the overall reference parameter space and is not appropriate for parameter optimization. Moreover, this strategy is somewhat sensitive to the obtained VQ codebooks. It is known that the VQ procedure does not converge to a unique solution, in consequence, the system may

lack robustness. To this end, a new template distance function, which takes account of all codewords in a speaker codebook, would be more reasonable.

3 DISCRIMINATIVE TRAINING METHOD

Let S denote the dimension of a sample vector

$$\mathbf{x}_n = (x_1, x_2, \ldots, x_S)^t \tag{1}$$

which is known to come from one of K classes $\{C_j\}_{j=1}^K$. A set of training samples is given as $\psi = \{\mathbf{x}_n\}_{n=1}^N$. A classifier is defined by a set of parameters, denoted by Λ, and a decision rule. Given the training set ψ, the task of a minimum error classifier design is to find the parameter set and the accompanying decision rule, such that the probability of misclassifying any $\mathbf{x}_n$ is minimized. The subscript of $\mathbf{x}_n$ is suppressed later on for simplicity. Let $r_{j,b,s}$ be a cepstral coefficient in a codeword

$$\mathbf{r}_{j,b} = \{r_{j,b,s}\}_{s=1}^S. \tag{2}$$

The collection of VQ codebooks for K speakers is denoted by Λ,

$$\Lambda_j = \{\mathbf{r}_{j,b}\}_{b=1}^B \tag{3}$$

$$\Lambda = \{\Lambda_j\}_{j=1}^K \tag{4}$$

where B is the codebook size.

According to the two motivations mentioned in last section, we give the following three-step definitions following the three-step procedure proposed by Juang and Katagiri [3] to derive the updating rules for optimal recognizer design. The three-step definition emulates the classification/recognition operation as well as the performance evaluation, particularly in terms of classification errors, in a smooth function form:

a) The discriminant function $g_j(\mathbf{x};\Lambda)$,

$$g_j(\mathbf{x};\Lambda) = \ln\{\sum_{b=1}^{B} e^{-d_{j,b}\xi}\}^{-1/\xi} \quad \xi > 0 \tag{5}$$

where ξ serves as a smoothing factor to incorporate all codewords, $d_{j,b}$ is the weighted cepstral distance between the utterance and the bth codeword of speaker C_j,

$$d_{j,b} = \sum_{s=1}^{S} \frac{1}{\sigma_s^2}(r_{j,b,s} - x_s)^2 \tag{6}$$

$g_j(\mathbf{x};\Lambda)$ evaluates the log-likelihood of class C_j upon observing the pattern $\mathbf{x}$. The exponential form in Eq.(5) reflects the fact that we are operating on the distances and a smaller distance means a better match between the two patterns. The implied classification rule is defined as

$$\mathbf{x} \in C_k \qquad \text{if } g_k(\mathbf{x};\Lambda) = \min_j g_j(\mathbf{x};\Lambda) \tag{7}$$

By varing the value of ξ, the composite effect of $d_{j,b}$'s upon g_j can be adjusted. In the extreme case where ξ approaches ∞, g_j equals the minimal $d_{j,b}$ among all the B codewords.

The above definitions lead to a continuous discriminant function $g_j(\mathbf{x};\Lambda)$ which is approproate for a gradient operation. The discriminant function may also be a more reasonable measure than the conventional distance since the incorporation of smoothing factors implies soft, and perhaps more robust decision.

b) The misclassification measure,

$$d_k(\mathbf{x};\Lambda) = g_k(\mathbf{x};\Lambda) - \ln\{\frac{1}{K-1}\sum_{j,j\neq k} e^{-g_j(\mathbf{x};\Lambda)\eta}\}^{-1/\eta} \tag{8}$$

The second term of Eq.(8) is considered as an anti-discriminant function of the input vector $\mathbf{x}$ in class k, which is a collective representation of all the other competing classes with respect to class k. The composite effect of competing distance may be adjusted by varying η. One extreme case is when η approaches ∞, it becomes

$$d_k(\mathbf{x};\Lambda) = g_k(\mathbf{x};\Lambda) - g_{j'}(\mathbf{x};\Lambda) \tag{9}$$

where $C_{j'}$ is the class with the largest discriminant value among those classes other than C_k, because $(M-1)^{1/\infty} \cong 1$. Obviously in this case, $d_k(\mathbf{x};\Lambda) > 0$ implies misclassification and $d_k(\mathbf{x};\Lambda) \leq 0$ means correct decision. In this way, the decision rule becomes a judgement on a scalar value. The misclassification measure is so defined that it is continuous with respect to the classifier parameters and offers a fair amount of flexibility.

c) The objective function or the sample loss function $l_k(d_k)$

$$l_k(d_k) = \frac{1}{1+e^{-d_k}} \tag{10}$$

The above functions is smoothed zero-one cost functions suitable for gradient algorithms. Clearly, when $d_k < 0$ which implies correct classification, virtually no cost is incurred. On the other hand, a positive d_k leads to a penalty which becomes essentially a count of classification error. Finally, for any unknown $\mathbf{x}$, the classifier performance is measured by

$$l(\mathbf{x};\Lambda) = \sum_{k=1}^{K} l_k(d_k)1(\mathbf{x} \in C_k) \tag{11}$$

where $1(\cdot)$ is an indicator function:

$$1(\mathcal{Q}) = \begin{cases} 1, & \text{if } \mathcal{Q} \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The solution to Λ is found by minimizing the overall loss function,

$$L(\Lambda) = \sum_{\forall \mathbf{x}} l(\mathbf{x};\Lambda) \tag{13}$$

Resorting to the gradient descent algorithm, at least a local minimum of $L(\Lambda)$ is reached if Λ is adjusted by

$$\Lambda^{n+1} = \Lambda^n + \varepsilon_n \delta\Lambda^n \tag{14}$$

where ε_n, the updating gain, is subject to the constraints[3] of

$$\sum_n \varepsilon_n \to \infty \tag{15}$$

and

$$\sum_n \varepsilon_n^2 < \infty \tag{16}$$

The correction term $\delta\Lambda^n$ or adjustment of $r_{j,b,s}$ may be derived by differentiating the sample loss function with respect to $r_{j,b,s}$.

$$\frac{\partial l_k(d_k)}{\partial r_{j,b,s}} = \nu_k \pi_{j,k} \mu_j \phi_j \tag{17}$$

The updating rule is given as

$$r_{j,b,s}^{n+1} = r_{j,b,s}^n + \varepsilon_n \nu_k \pi_{j,k} \mu_j \phi_j \tag{18}$$

where

$$\nu_k = l_k(d_k)[1 - l_k(d_k)] \tag{19}$$

$$\pi_{j,k} = \begin{cases} \frac{e^{-g_j(x;\Lambda)\eta}}{\sum_{j',j'\neq k} e^{-g_{j'}(x;\Lambda)\eta}} & j \neq k \\ -1 & j = k \end{cases} \tag{20}$$

$$\mu_j = \frac{e^{-d_{j,b}\xi}}{\sum_{b'=1}^{B} e^{-d_{j,b'}\xi}} \tag{21}$$

$$\phi_j = \frac{2}{\sigma_s^2}(r_{j,b,s}^n - x_s) \tag{22}$$

The updating gain is approximated by

$$\varepsilon_n = \varepsilon_0(1 - \frac{n}{N}) \tag{23}$$

The ν_k factor represents the sensitivity to the approximate 0-1 error count function as measured in terms of the misclassification measure. It will lead to a substantial parameter adjustment if the absolute value of d_k is small which implies that the training token is close to the boundary of the competing word and is likely to be misrecognized. On the other hand, if the absolute value of d_k is large as in the case where the input token is either unlikely to cause confusion or obviously an extreme outlier, the amount of adjustment is, therefore, accordingly reduced.

Other factors, $\pi_{j,k}, \mu_k$, and ϕ_k, are related to discriminant function, cepstral distance respectively, and the basically work in a way similar to ν_k. When we consider the N-time presentations, i.e. iterations, of the data set ψ as an epoch, the initial gain of each epoch is ε_0 and the gain decreases by epoch.

In the extreme case when η and ξ both approach ∞, the parameter adjustment equations Eq.(18) can be simplified to

$$r_{k,b',s}^{n+1} = r_{k,b',s}^n - \varepsilon_n \nu_k \frac{2}{\sigma_s^2}(r_{k,b',s}^n - x_s) \tag{24}$$

$$r_{j',b',s}^{n+1} = r_{j',b',s}^n + \varepsilon_n \nu_k \frac{2}{\sigma_s^2}(r_{j',b',s}^n - x_s) \tag{25}$$

where the subscript b' is the index of the minimum distance codeword. Therefore, the adaptive adjustments in Eq.(24) is performed only on the parameters of b'th codeword of the speaker C_k. Similarly the subscript j' implies that only the parameters of speaker $C_{j'}$, which leads to the minimum discriminant score among all the speaker other than C_k, are adjusted.

Although the parameters are adaptively adjusted, an initial parameter set is required. In our simulations, the codebooks trained by class (speaker) with LBG procedure are used as the initial codebooks.

It is noted that each time a training sample $\mathbf{x}$ from class C_k is given, not only codebook Λ_k is adjusted, but also all other codebooks from the competing speakers. This training strategy pays more attention to the discrimination among speakers when compared to the conventional VQ scheme.

4 EXPERIMENTS

The database is collected from 200 French speakers, which consists of a five-word sentence naturally uttered five times by each speaker for about one second each time. A 12th mel-scale cepstral analysis is performed at intervals of 10ms with a 32ms Hamming window. A codebook size of $B = 16$ is selected empirically. Each speaker is represented by a VQ codebook, called a class. Two experiments were conducted. First, we used three of the five sessions as training data and others as test data, the results are listed in the Table 1. Second, we used four of the five sessions as the training data and the fifth one as the test data. Rotating the order of them, we came up with five assessment sets. The results are reported in Table 2. The averages of the five assessments are reported in Table 3.

In the two experiments, the DVQ training algorithm was compared to the conventional VQ, referred to as CVQ, on the same data set. The performances of DVQ with respect to two settings of the controlling parameters, ξ and η, are given. The first setting is the extreme case of ξ and η approaching ∞ where the algorithm is reduced to a corrective training[1]. In the second one, we have $\xi = \eta = 2$ where the distances to all competing classes and all codewords are taken into account so that the discriminative function is continuous with respect to the parameter set Λ. In practice, we only used the four-smallest distances in Eq.(5) and the three-smallest distances in Eq.(8) for computational complexity considerations. It was found that the effect of discontinuity caused by the modification is insignificant. In the verification, the misclassification measure Eq.(8) was used as the normalized scoring for making decision. A Bayesian decision rule was then applied, which had been established during the training phase.

In tables 1, 2 and 3, *sr* is referred to as the speaker recognition error rate, *sv* as the verification error rate, *(c)* as the close test, test on the training data, and *(o)* as the open test, test on the test data. While *ts* is referred to as the test session.

	sr(c)	sr(o)	sv(c)	sv(o)
CVQ	0.00	2.25	0.69	1.34
DVQ($\eta, \xi = \infty$)	0.00	0.50	0.23	0.93
DVQ($\eta, \xi = 2$)	0.00	0.50	0.18	0.89

Table 1: *Error rate (%) in experiment 1*

ts	algorithm	sr(c)	sr(o)	sv(c)	sv(o)
1	CVQ	0.00	5.50	0.67	1.37
	DVQ($\eta, \xi = \infty$)	0.00	0.50	0.61	1.35
	DVQ($\eta, \xi = 2$)	0.00	0.50	0.45	1.20
2	CVQ	0.00	1.50	0.70	0.91
	DVQ($\eta, \xi = \infty$)	0.00	1.00	0.57	0.80
	DVQ($\eta, \xi = 2$)	0.00	1.00	0.24	0.47
3	CVQ	0.00	1.00	0.55	0.52
	DVQ($\eta, \xi = \infty$)	0.00	0.50	0.23	0.47
	DVQ($\eta, \xi = 2$)	0.00	0.50	0.21	0.45
4	CVQ	0.00	2.00	0.69	0.66
	DVQ($\eta, \xi = \infty$)	0.00	0.00	0.47	0.41
	DVQ($\eta, \xi = 2$)	0.00	0.00	0.43	0.37
5	CVQ	0.00	1.00	0.70	1.12
	DVQ($\eta, \xi = \infty$)	0.00	0.00	0.40	0.93
	DVQ($\eta, \xi = 2$)	0.00	0.00	0.44	0.99

Table 2: *Error rates (%) in experiment 2*

	sr(c)	sr(o)	sv(c)	sv(o)
CVQ	0.00	2.20	0.66	0.92
DVQ($\eta, \xi = \infty$)	0.00	0.40	0.44	0.79
DVQ($\eta, \xi = 2$)	0.00	0.40	0.35	0.70

Table 3: *Average error rate (%) in experiment 2*

5 CONCLUSIONS

In this paper, we have developed a discriminative VQ training algorithm and applied it to the VQ-based speaker recognition. By examining the evaluation reports, we note that the new algorithm outperforms the conventional one with 1.75% improvement in speaker recognition and 0.45% improvement in verification for the open test of the first experiment, with 1.8% and 0.22% improvement for the rotating test. The DVQ training algorithm is shown to be suitable for limited training data and large population speaker recognition purpose. The discriminative VQ outperforms the speaker dependent VQ model in verification experiments because it uses a normalized score of Eq.(8), which consider the scores from the competing speakers as the normalization background and keeps the decision strategy consistent with the training procedure of minimizing the classification error.

REFERENCES

[1] B.H. Juang L. Rabiner. *Fundementals of Speech Recognition.* Prentice Hall, New Jersey, 1993.

[2] H. Li, J.-P. Haton, and Y. Gong. On the MMI learning of Gaussian mixture speaker model. In *Proceedings of European Conference on Speech Technology*, Madrid, Spain, 1995.

[3] S.Katagiti B.H. Juang. Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing*, 40(12):3043–3054, December 1992.

[4] B.H. Juang P.C. Chang. Discriminative training of dynamic programming based speech recognizers. *IEEE Trans. on Speech and Audio Processing*, 1(2):135–143, April 1993.

[5] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for the vector quantizer design. *IEEE Trans. on Communication*, 28(1):84–95, Jan. 1980.