A NEW COMMON COMPONENT GMM-BASED SPEAKER RECOGNITION METHOD

Yih-Ru Wang and Chen-Yu Chiang

Department of Communications Engineering,
National Chiao Tung University, Taiwan, Republic of China
yrwang@cm.nctu.edu.tw

ABSTRACT

In this paper, a new common component GMM (CCGMM)-based speaker recognition approach is presented. It first defines a divergence measure to calculate the similarity of the speech signals of two speakers. Then, a CCGMM training algorithm which simultaneously maximizes the likelihood of CCGMM and the inter-speaker divergence is proposed. Performance of the proposed approach was examined using a telephone-speech database (MAT) containing 2962 speakers. A speaker recognition rate of 90.0% was achieved. The recognition rate raised to 96.1% when it was combined with the conventional GMM-based scheme.

1. INTRODUCTION

For text-independent speaker identification/verification problem, the most successful generative models are Gaussian Mixture Models (GMM) [1]. A GMM with several (say 32) mixture components is used to model the speech signal characteristics for each speaker. The maximum likelihood (ML) criterion is usually used in recognition decision. Besides, the maximum a posteriori (MAP) model adaptation technique is also popular in speaker verification [2]. In a MAP-adapted system, the GMM model of a speaker is adaptively generated from a universal background model (UBM) by using his speech signal. Generally, a UBM with large number (say 2048) of Gaussian components is used. For a speaker recognition system involving a large number of speakers, the computational complexity will be very high because of the use of a large number of mixture components in the likelihood computation for both the GMM and the MAP-adapted GMM approaches.

To solve the problem, a new speaker recognition system using a GMM with common mixture components, referred to as CCGMM [6], is proposed in this paper. It uses a large GMM for all speakers, but it uses a divergence measure to simplify the likelihood computation. The divergence (or called Jeffrey divergence) [3-5] was used to measure the similarity between two probability distributions, or more precisely, the relative entropy of two probability distributions. The divergence measure had been successfully used in simple speech signal classification such as speech segmentation or voice activity detector (VAD). In those applications, the distribution of speech signal was assumed to be Gaussian. The divergence measure was also integrated into a speaker recognition system [5]. In this paper, the divergence is used to measure the dissimilarity between speech signals of two speakers represented by CCGMM models. A training algorithm to simultaneously maximize the likelihood of CCGMM and the inter-speaker divergence is proposed.

The remainder of the paper is organized as follows. In Section 2, the CCGMM-based divergence measure is described in detail. In Section 3, the CCGMM training algorithm for building a speaker recognizer is discussed. Section 4 presents the experimental results of the proposed speaker recognition method by using a speech database containing 2962 speakers. Some conclusions are given in the last section.

2. CCGMM AND DIVERENCE MESUARE

In the past, the divergence measure [3-5] is used to measure the dissimilarity of two random variables based upon the information theory. It is derived from the average discriminating information between the two random variables. It can be expressed by

$$D(p_1,p_2)=\int\left[p_1(O)-p_2(O)\right]\ln\frac{p_1(O)}{p_2(O)}dO\,, \tag{1}$$

where $p_1(O)$ and $p_2(O)$ are the distributions of the two random variables which can be the speech signals of two speakers. In order to find the divergence measure in Eq. (1), the estimates of the distribution functions of the two random variables must be done first. In some previous studies [3, 4], the distributions of random variables were assumed to be Gaussian. In this paper, a mixture Gaussian distribution is used to model the distribution of the speech signal for each speaker, i.e.,

$$p_s(O\mid\lambda_s)=\sum_{i=0}^{M-1}c_{is}N(O|\mu_{is},\Sigma_{is})\quad\forall s\,, \tag{2}$$

where $\lambda_s=\{(c_{is},\mu_{is},\Sigma_{is});i=0,\cdots,M-1\}$ is the parameter set of speaker s. Then the divergence measure between two speakers, s and s', becomes

$$D(p_s,p_{s'})=\int\left[\sum_i c_{is}\cdot N(O|\mu_{is},\Sigma_{is})-\sum_i c_{is'}\cdot N(O|\mu_{is'},\Sigma_{is'})\right]\ln\frac{\sum_i c_{is}\cdot N(O|\mu_{is},\Sigma_{is})}{\sum_i c_{is'}\cdot N(O|\mu_{is'},\Sigma_{is'})}dO\,. \tag{3}$$

The above equation can be simplified by using common component GMM (CCGMM) models. In a CCGMM, a set of common Gaussian components, $\{N(O \mid \mu_i, \Sigma_i); i = 1, \cdots, M\}$, was used. Thus, the speakers' distributions become

$$p_s(O \mid \lambda_s) = \sum_{n=0}^{M-1} c_{is} N(O|\mu_i, \Sigma_i) \quad \forall s, \tag{4}$$

where $\lambda_s = \{(c_{is}, \mu_i, \Sigma_i); i = 0, \cdots, M-1\}$ is the parameter set for speaker *s*. The divergence measure in Eq. (3) can then be simplified as

$$D(p_s, p_{s'}) = \sum_i (c_{is} - c_{is'}) \int_{R_i} N(O|\mu_i, \Sigma_i) \ln \frac{\sum_i c_{is} \cdot N(O|\mu_i, \Sigma_i)}{\sum_i c_{is'} \cdot N(O|\mu_i, \Sigma_i)} dO, \tag{5}$$

where R_i is the region of the *i*th mixture component. We now make an assumption to approximate the speakers' probability distributions in the region of a single Gaussian by

$$\sum_i c_{is} N(O|\mu_i, \Sigma_i) \approx c_{is} N(O \mid \mu_i, \Sigma_i) \quad \forall O \in R_i \text{ and } \forall s. \tag{6}$$

Then, the divergence measure between two distributions can be approximated by

$$D(p_s, p_{s'}) \approx \sum_i (c_{is} - c_{is'}) \ln \frac{c_{is}}{c_{is'}}. \tag{7}$$

Comparing the above divergence measure with the original definition of divergence shown in Eq. (1), we find that they have the same form. Eq. (7) can therefore be treated as a divergence of two discrete random variables.

In order to get better approximation in Eq. (7), a global covariance matrix Σ is used for all mixture components. The CCGMM for speaker *s* becomes

$$p_s(O|\lambda_s) = \sum_i c_{is} N(O|\mu_i, \Sigma) \quad \forall s. \tag{8}$$

The above CCGMM with global covariance matrix can be treated as using a set of Parzen windows with Gaussian kernels to estimate the distributions of signal sources. In the speaker recognition case, it is used to model the speakers' speech characteristics. The objective is to efficiently encode the data samples using a set of Parzen weights. In other words, we wish to find a compact set of processing elements that can represent the source data in terms of its distribution.

The set of Gaussian kernels, $\{N(O \mid \mu_i, \Sigma); i = 0, \cdots, M-1\}$, can be found from the data of all speakers, i.e.,

$$\underset{\lambda}{MAX} \prod_{s,t} \sum_i c_{is} N(O_t^s|\mu_i, \Sigma), \tag{9}$$

where O_t^s is the feature vector of speaker *s* at time *t*. In fact, the Gaussian kernel set is the universal background model (UBM) with global variance. And the *i*th mixture weight, c_{is}, of speaker *s* can be found by the following re-estimation formula

$$\bar{c}_{is} = \frac{\sum_t c_{is} N(O_t^s; \mu_i, \Sigma)}{\sum_t \sum_j c_{js} N(O_t^s; \mu_j, \Sigma)}, \quad i = 0, \cdots, M-1. \tag{10}$$

After the CCGMM models of all speakers are found, the speaker recognition test can be performed by using either the maximum likelihood (ML) criterion or the minimum divergence measure (MDM) criterion.

3. MAXIMUM INTER-SPEAKER DIVERGENCE CCGMM RECOGNITION MODEL

In the above modeling of CCGMM, a set of Gaussian kernels are found from the data of all speakers. Now, we want to find a set of CCGMM recognition models, $\lambda_s = \{(c_{is}, \mu_i, \Sigma); i = 0, \cdots, M-1\}; \forall s$, which simultaneously maximizes the likelihood of each speaker's speech data and the inter-speaker divergence, i.e.,

$$\underset{\lambda_s}{MAX} \prod_t p_s(O_t^s|\lambda_s), \quad \forall s \tag{11}$$

and

$$\underset{\lambda_S}{MAX} \sum_s \sum_{s' \neq s} D(p_s, p_{s'}). \tag{12}$$

Since the simultaneous maximization of Eqs. (11) and (12) is complicated, we thus separate the problem into two steps:

(1) Find a new model $\lambda_s' = \{(c_{is}, \mu_{is}, \Sigma); i = 0, \cdots, M-1\}$ for each speaker *s* which

$$\underset{\lambda_s'}{MAX} \prod_t \sum_i c_{is} N(O_t^s|\mu_{is}, \Sigma), \quad \forall s. \tag{11'}$$

Here the initial model can be set to the CCGMM obtained in Section 2.

(2) Given with $\lambda_s' = \{(c_{is}, \mu_{is}, \Sigma); i = 0, \cdots, M-1\}$, find μ_i of the CCGMM model $\lambda_s = \{(c_{is}, \mu_i, \Sigma); i = 0, \cdots, M-1\}$ which maximizes the inter-speaker divergence, i.e.,

$$\underset{\lambda_s}{MAX} \sum_s \sum_{s' \neq s} (\bar{c}_{is} - \bar{c}_{is'}) \log \frac{\bar{c}_{is}}{\bar{c}_{is'}} \Bigg|_{\lambda_s'}. \tag{12'}$$

From Eq. (11'), we find the new model $\overline{\lambda'}_s = \{(\bar{c}_{is}, \bar{\mu}_{is}, \bar{\Sigma}); i = 0, \cdots, M-1\}$ of each speaker *s* by using the EM algorithm. The resulting re-estimation formula can be expressed by

$$\bar{c}_{is} = \frac{1}{T_s} \sum_t \left[p(i|O_t^s, \lambda_s) \right], \tag{13}$$

$$\bar{\mu}_{is} = \frac{\sum_t \left[p(i|O_t^s, \lambda_s^{'}) O_t^s \right]}{\sum_t \left[p(i|O_t^s, \lambda_s^{'}) \right]}, \quad (14)$$

$$\bar{\Sigma} = \frac{1}{S}\left(\sum_i \sum_s \bar{c}_{is} \left(\overline{O_{is}^2} - \bar{\mu}_i^2 \right) \right), \quad (15)$$

where T_s is the number of training samples for speaker s, S is the total number of speakers,

$$\overline{O_{is}^2} = \frac{\sum_t \left[p(i|O_t^s, \lambda_s^{'}) O_t^s (O_t^s)^T \right]}{\sum_t \left[p(i|O_t^s, \lambda_s^{'}) \right]}, \quad (16)$$

and

$$p(i \mid O_t^s, \lambda_s^{'}) = \frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})}. \quad (17)$$

After updating $\lambda_s^{'}$, we can find μ_i, for $i = 0, \cdots, M-1$, of the CCGMM models $\lambda_s = \{(c_{is}, \mu_i, \Sigma); i = 0, \cdots, M-1\}$ based on Eq. (12'), i.e.,

$$\underset{\bar{\mu}_i}{MAX} \sum_s \sum_{s' \neq s} (\bar{c}_{is} - \bar{c}_{is'}) \log \frac{\bar{c}_{is}}{\bar{c}_{is'}} \Big|_{\lambda_s^{'}}. \quad (18)$$

To find the solution, we first express the CCGMM weight $\bar{c}_{is}$ by

$$\bar{c}_{is} = \frac{1}{T} \sum_t \left[p(i|O_t^s, \lambda_s^{'}) \right] = \frac{1}{T} \sum_t \frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})}. \quad (19)$$

Then, find the derivation of CCGMM weight $\bar{c}_{is}$ by

$$\frac{\partial}{\partial \mu_i} \bar{c}_{is} = \frac{-1}{T} \sum_t \left[\left(\frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})} - \left(\frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})} \right)^2 \right) (O_t - \mu_i) \Sigma^{-1} \right]. \quad (20)$$

Since $\frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})} = p(i \mid O_t, \lambda_s^{'}) < 1$, we therefore can get the following approximation

$$\frac{\partial}{\partial \mu_i} \bar{c}_{is} \approx \frac{-1}{T} \sum_t \left[\frac{c_{is} p(O_t|i, \lambda_s^{'})}{p(O_t|\lambda_s^{'})} (O_t - \mu_i) \Sigma^{-1} \right] = \frac{-\Sigma^{-1}}{T} \bar{c}_{is} \left(\bar{\mu}_{is} - \mu_i \right). \quad (21)$$

We then let the derivative of $\sum_s \sum_{s' \neq s} (\bar{c}_{is} - \bar{c}_{is'}) \log(\bar{c}_{is} / \bar{c}_{is'}) \Big|_{\lambda_s^{'}}$ equal to zero,

$$\sum_s \sum_{s' \neq s} \left[\begin{array}{l} \left((c_{is}\mu_{is} - c_{is}\mu_i) - (c_{is'}\mu_{is'} - c_{is'}\mu_i) \right) \log \frac{c_{is}}{c_{is'}} \\ + \left(c_{is} - c_{is'} \right) \left(\mu_{is} - \mu_{is'} \right) \end{array} \right] = 0, \quad (22)$$

and obtain the optimal means of CCGMM

$$\mu_i = \frac{\sum_s \sum_{s' \neq s} \left(\left(c_{is}\mu_{is} - c_{is'}\mu_{is'} \right) \log \frac{c_{is}}{c_{is'}} + \left(c_{is} - c_{is'} \right) \left(\mu_{is} - \mu_{is'} \right) \right)}{\sum_s \sum_{s' \neq s} \left(c_{is} - c_{is'} \right) \log \frac{c_{is}}{c_{is'}}}, \quad \text{for } i = 0, \cdots, M-1. \quad (23)$$

The above mixture mean re-estimation formula will move μ_i toward the mean, μ_{is}, of the speaker with larger weights and away from the mean, $\mu_{is'}$, of the speaker with smaller weights. Lastly, Eqs. (13) - (17) and (23) can be iteratively applied to re-estimate the maximum inter-speaker divergence CCGMM models (MD-CCGMM) of all speakers.

4. EXPERIMENTS ON SPEAKER RECOGNITION SYSTEMS

4.1. Speech Database

The MAT (***M***andarin speech data ***A***cross ***T***aiwan) telephone-speech database [7] was used in the following experiments to examine the performance of the proposed CCGMM-based speaker recognition approach. The database was collected by allowing speakers to input their voices by using any telephone handset around Taiwan through the public switching telephone network. The input speech signal was sampled in 8kHz with 16-bit linear PCM format.

To extract recognition features, a spectral analysis was applied to the speech waveform for every 30ms frame with 10ms frame shift. It extracted 38 recognition features including 12 MFCCs with their first and second derivatives, and delta and delta-delta log-energies. A window length of 7 frames was used in the calculations of derivatives. Then, a silence detector was used to remove the silence segments in the beginning and ending of each sentential utterance.

The number of speakers used in the following experiments was 2962 including 1377 males and 1585 females. The lengths of the training and test data were 24 and 6 seconds, respectively, for each speaker.

4.2. Speaker recognition systems

First, a conventional GMM speaker identification system was implemented. A GMM of 32 mixture components with diagonal covariance matrix was built for each speaker. Totally, 94,784 (32*2962) mixtures were used in the GMM speaker recognizer. The recognizer used the ML criterion. As shown in Table 1, a recognition rate of 96.0% was achieved. The recognition rate is high because all speech signals of a speaker were collected in

the same phone call so that there was no environmental mismatch in this task.

We then examined the CCGMM scheme discussed in Section 2. A universal background model (UBM) with global diagonal covariance matrix was firstly constructed. The CCGMM model $\lambda_s = \{(c_{is}, \mu_i, \Sigma); i = 0, \cdots, M-1\}$ of each speaker was then constructed by using the UBM. When 1024 mixtures were used, $\frac{1}{N}\sum_{i=1}^{N}\left(MAX_{j\neq i} N(\mu_j|\mu_i,\Sigma)/N(\mu_i|\mu_i,\Sigma)\right)$ $\approx 2.28\cdot 10^{-2}$, the approximation in Eq. (6) seems reasonable. The speaker recognizer using ML criterion was then tested. As shown in Table 1, the recognition rate was only 81.9% for the case of using 2048 mixtures. By switching to the minimum divergence measure (MDM) criterion, the recognition rate was improved to 84.9%. Although the recognition rate of the CCGMM scheme is low as compared with the GMM scheme, its computational complexity is much lower.

Then, the training algorithm proposed in Section 3 was applied to find a set of CCGMM models which maximized the inter-speaker divergence. It is referred to as the MD-CCGMM scheme. The MDM criterion was used. As shown in Table 1, a speaker recognition rate of 90% was achieved.

Table 1. Experimental results for GMM, CCGMM, and MD-CCGMM schemes.

Scheme	No. of mix.	Recog. Rate (%)
GMM	94786(32*2962)	96.0
CCGMM/ML	512	75.2
	1024	79.9
	2048	81.9
CCGMM/MDM	512	79.6
	1024	84.1
	2048	84.9
MD-CCGMM/ MDM	1024	88.0
	2048	90.0

Lastly, the MD-CCGMM scheme and the GMM scheme were combined as shown in Fig. 1. The recognition score was the weighted combination of the log-likelihood of the GMM recognizer and the divergence measure of the MD-CCGMM recognizer. The following recognition criterion was adopted

$$Arg\ \underset{s}{MAX}\left\{\frac{1}{T}\sum_t \log\left(p_s(O_t|\lambda_s)\right) - \alpha D(p(O), p_s)\right\}, \qquad (24)$$

where $p(O)$ was the distribution of the test data. To reduce the computational load, only the Top-N candidates obtained in the MD-CCGMM recognizer were sent to the GMM recognizer.

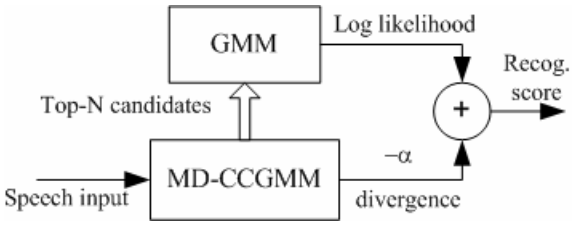

Figure 1. The speaker recognition system that integrated MD-CCGMM and GMM recognizers.

The recognition results were shown in Table 2. A recognition rate of 96.1% was achieved for the case of using Top-50 candidates. It can also be found from Table 2 that the number of mixtures needed to be calculated is significantly lower than the GMM scheme.

Table 2. Recognition results of integration MD-CCGMM and GMM recognizer.

Top N candidates	50	10	5
Recog. rate (%)	96.1	95.7	95.2
Total no. of mixtures need to be calculated	6248	2368	2208

5. CONCLUSIONS

In this paper, a new CCGMM-based speaker recognition approach was discussed. It defined a divergence measure to calculate the similarity of the speech signals of two speakers and proposed a sophisticated training algorithm to simultaneously maximize the likelihood of CCGMM and the inter-speaker divergence. Experimental results showed that it performed very well. An advantage of low computational complexity has been achieved. The proposed divergence measure can be used in speaker clustering and in speaker verification.

6. ACKNOWLEDGEMENTS

This work was supported by NSC of Taiwan, under the project with contract NSC 92-2213-E-009-046 and MOE under the project with contract A-93-E-FA06-4-4.

7. REFERENCES

[1] D. Reynolds and R. Rose, " Robust text-independent speaker identification using Gaussian mixture speaker models, ", IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[3] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems", *Proc. Roy. Soc. Lon., Ser. A*, no. 186, 453-461, 1946.

[4] J. T. Tou, R. C. Gonzales, " Pattern Recognition Principles ", Chap. 7-8, Addison-Welsley corp., 1974.

[5] J. P. Campbell, Jr., " Speaker Recognition : A Tutorial, ", Proc. Of IEEE, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[6] Yih-Ru Wang, Chi-Han Huang, " Speaker-and-environment change detection in broadcast news using common component GMM-based divergence measure", ICSLP-2004, Vol. II, pp. 1069-1072, Oct. 2004.

[7] Hsiao-Chuan Wang, Seide Frank, Chiu-Yu Tseng, Lin-Shan Lee, "MAT-2000 - design, collection, and validation of a Mandarin 2000-speaker telephone speech database", *ICSLP-2000*, vol.4, 460-463, 2000.