International Symposium on Chinese Spoken
Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

ON USE OF GMM FOR MULTILINGUAL SPEAKER VERIFICATION: AN EMPIRICAL STUDY

Xi-Ke QING and Ke CHEN
National Laboratory of Machine Perception and The Center for Information Science, Peking University, Beijing
E-mail: chen@cis.pku.edu.cn, URL: http://speech.cis.pku.edu.cn

ABSTRACT

This paper presents an empirical study on multilingual speaker verification based on a sophisticated statistical model – Gaussian Mixture Model (GMM). The languages used include Mandarin, Cantonese, and English. Comparative results of speaker verification are presented in terms of different databases associated with different languages. Our simulation results indicate that GMM can be used as a unified model in multilingual speaker verification, which provides an easy-to-use way for building a multilingual speaker verification system.

1. INTRODUCTION

The demand for intelligent information access is highly increased as Internet and telecommunication are rapidly growing. As an easy access way, speech becomes one of the most important media for communication between human and machine. For intelligent speech information access, a security issue is often along with personal or private information access, which results in the need of speaker recognition. [1]

Speaker recognition is a task of automatic recognition of a person by his/her voice. As one form of speaker recognition, speaker verification is a process that the identity claim of a speaker is accepted or rejected. Furthermore, a speaker verification system could be either text-dependent or text-independent. Text-dependent means that the same text is used in training and test. In contrast, any text is allowed in either training or test in a text-independent system. In practice, speaker verification covers most of tasks associated with personal or private information access by voice.

On the other hand, research on intelligent speech information system is being conducted towards multilingual information access to meet requirements of global business. Therefore, multilingual issues in speaker verification are worth studying. Theoretically, a speaker recognition process is irrelevant to spoken languages. For the multilingual case, however, a set of elaborately selected utterances carrying rich phonetic information of multiple languages are necessary for training a speaker verification system. This process of generating training samples is often hard for engineers without the help of linguists. Therefore, most of existing speaker verification systems are monolingual without careful sample selection. A usual way for multilingual systems is to process different languages separately. Different models are reported to be useful for different languages, which leads to another problem how to select an appropriate model for a specific language. In a multilingual speech information system, a unified model could be often demanded to deal with multilingual speaker verification for ease-of-development if it exists. In this paper, we explore the possibility of GMM being such a unified model in an empirical way. As a result, we present multilingual speaker verification results in terms of both text-dependent and text-independent cases, and our simulation results indicate that GMM can be a unified model for multilingual speaker verification.

The rest of the paper is organized as follows. The next section presents an overview of a GMM based speaker verification system. Section 3 describes databases used in our simulations. Section 4 reports simulation results on different databases. Conclusions are drawn in the last section.

2. GAUSSIAN MIXTURE MODEL

Gaussian mixture model (GMM) is a sophisticated statistical model, which can be viewed as a universal estimator. GMM has been applied to speaker recognition to model speaker's characteristics. [2][3].

GMM is specified as

$$\boldsymbol{p}(\vec{\boldsymbol{X}} \mid \lambda) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{M}} \boldsymbol{p_i} \boldsymbol{b_i}(\vec{\boldsymbol{X}}), \qquad (1)$$

where $\boldsymbol{p_i}$ are mixture coefficients subject to $\sum_{i=1}^{M} p_i = 1$.

$\boldsymbol{b_i}(\vec{\boldsymbol{X}})$ are component Gaussian distributions:

$$\boldsymbol{b_i}(\vec{\boldsymbol{X}}) = \frac{\exp\{-\frac{1}{2}(\vec{\boldsymbol{X}} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{\boldsymbol{X}} - \vec{\mu}_i)\}}{(2\pi)^{(\boldsymbol{D}/2)} \mid \Sigma_{\boldsymbol{i}} \mid^{1/2}}.$$

Here, $\lambda = (p_i, \vec{\mu}_i, \Sigma_i)$, D is the number of dimension of $\vec{X}$, M is the number of components, λ is the parameter set, $\vec{\mu}_i$ is the mean of the *i*-th components, and Σ_i is the covariance matrix of the *i*-th components. Consequently, its log-likelihood function is defined as

$$\boldsymbol{L}(\lambda) = \sum_{t=1}^{T} \log \boldsymbol{p}(\vec{\boldsymbol{X}}_t \mid \lambda). \tag{2}$$

Here, $\vec{X}_t$ is input vector *t* and *T* is the number of input vectors.

Parameter estimation in GMM is often performed by EM (Expectation Maximization) algorithm [2]. Here we summarize the EM algorithm as follows:

For the *i*-th element in the GMM, a posteriori probability is defined as follows:

$$\boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda) = \frac{\boldsymbol{p}_i \boldsymbol{b}_i(\vec{\boldsymbol{X}}_t)}{\sum_{k=1}^{M} \boldsymbol{p}_k \boldsymbol{b}_k(\vec{\boldsymbol{X}}_t)}. \tag{3}$$

Then

$$\boldsymbol{p}_{i,new} = \frac{1}{\boldsymbol{T}} \sum_{t=1}^{T} \boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda), \tag{4}$$

$$\vec{\mu}_{i,new} = \frac{\sum_{t=1}^{T} \boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda) \vec{\boldsymbol{X}}_t}{\sum_{t=1}^{T} \boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda)}, \tag{5}$$

$$\Sigma_{i,new} = \frac{\sum_{t=1}^{T} \boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda) \vec{\boldsymbol{X}}_t \vec{\boldsymbol{X}}_t^{'}}{\sum_{t=1}^{T} \boldsymbol{p}(\boldsymbol{i}_t = \boldsymbol{i} \mid \vec{\boldsymbol{X}}_t, \lambda)} - \vec{\mu}_{i,new} \vec{\mu}_{i,new}^{'}. \tag{6}$$

Here, $p_{i,new}$ is the new weight after iteration, $\vec{\mu}_{i,new}$ is the new mean value, and $\Sigma_{i,new}$ is the new covariance matrix.

Repeat (3)-(6) until the likelihood function (2) does not increase.

For use of GMM in speaker verification, we fit a GMM to speaker's utterances based on the EM algorithm. Once training is finished, a GMM will be viewed as a statistical model representing a specific speaker. Thus, each speaker owns a GMM trained by his/her utterances. Given a decision rule, the log-likelihood function in (2) is used for examining whether an unknown utterance would be produced by a specific GMM corresponding to a claimed identity or not.

3. DATABASE

For evaluate performance of GMM in a multilingual case, we adopt several databases associated with different dialects, where utterances in different languages were spoken by the same speakers in the same session. As a result, such a pair of bilingual databases will be always available for investigating performance of GMM. Doing so tends to investigate the sensitivity of GMM to utterances belonging to different languages on the same conditions. On the other hand, we need to investigate performance of GMM in different working styles in to avoid any incorrect outcome achieved due to different data-dependent cases. For this purpose, we adopt two set of pair databases in our simulations, respectively, corresponding to text-dependent and text-independent cases.

3.1 Text-dependent Databases

For all the databases used for text-dependent case, recording was performed in noisy rooms. 10 isolated digits were recorded in five recording session. In every session, digits were prompted in a random order such that each one occurred exactly ten times. These database are briefly described as follows:

a) TD_BJ_MAND and TD_BJ_ENG

The pair databases were recorded by speech processing laboratory at Peking University. The TD_BJ_MAND database consists of Mandarin utterances, while the TD_BJ_ENG database contains English utterances. There are 15 speakers, 13 male and 2 female, in the two databases.

b) TD_HK_CANT and TD_HK_ENG

The pair databases were recorded by the human-machine interactive laboratory at The Chinese University of Hong Kong. The TD_HK_CANT database includes Cantonese utterances, while the TD_HK_ENG database is composed of English utterances. There are 15 speaker as well, 8 male and 7 female, in the two databases.

3.2 Text-independent Databases

The pair databases are the TI_BJ_MAND and TI_BJ_ENG databases recorded by speech processing laboratory at Peking University in a noisy room. Continuous utterances were recorded in four sessions and the interval between two consecutive sessions is about one week. The TI_BJ_MAND database consists of Mandarin utterances, while the TI_BJ_ENG database contains English utterances. There are 10 speakers, 8 male and 2 female, in. the two databases.

4. SIMULATIONS

It is well known that GMM is useful in text-independent speaker recognition [2][3]. However, it is rarely used in text-dependent speaker recognition. In order to possibility of GMM being a unified model, we first examine performance of GMM in text-dependent case in contrast to the performance of Hidden Markov Model (HMM). Then we investigate performance of GMM in terms of different databases described in Section 3.

In our simulations, preprocessing and feature extraction are briefly summarized as follows. In text-dependent case, speech is segmented into frames by 25 ms window processing at a 12.5 ms frame rate. In text-independent case, speech is passed through energy based detector so that only voiced parts remain and then are segmented into frames by 25 ms window processing without overlapping between frames. Each frame is normalized and a Mel-scaled cepstral feature vector is extracted accordingly. Consequently, 16-order Mel-scaled cepstrum is used in our simulations. Performance of GMM for speaker verification is evaluated by equal error rate (EER).

4.1 Results for Text-Dependent Case

First of all, we use a GMM of 16 components for text-dependent case in terms of the TD_BJ_MAND database. An HMM of two states and four Gaussian components per state is also used in the same circumstance. Results are shown in Fig. 1.

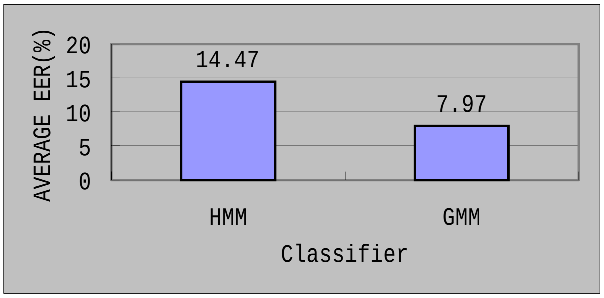

Fig. 1 Performance of HMM and GMM for text-dependent speaker verification on a Mandarin database.

From Fig. 1, it is evident that average EER of GMM is smaller than that of HMM. Note that the HMM produces the best performance among several HMMs of different structures in terms of the database. Our results here and more results [4] indicate that a proper GMM is applicable to text-dependent speaker verification as well and its performance is comparable with other sophisticated models commonly used in text-dependent case, e.g. HMM. Therefore, we can apply GMM to both text-dependent and text-independent cases. For text-dependent case, we fix the structure of GMM for multilingual databases, and the GMM of 16 Gaussian components are used in our simulations.

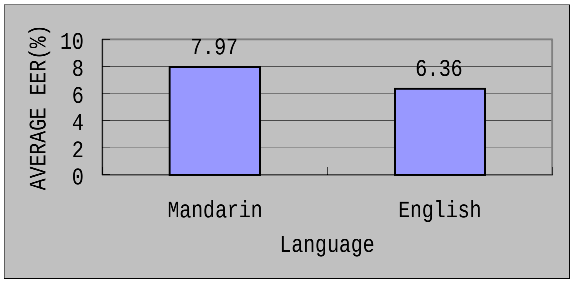

Fig.2 Results on the Mandarin and English databases.

Fig. 2 illustrates results on the TD_BJ_MAND and TD_BJ_ENG databases. Results in Fig. 2 show that the average EER for Mandarin is quite near that for English by the same GMM.

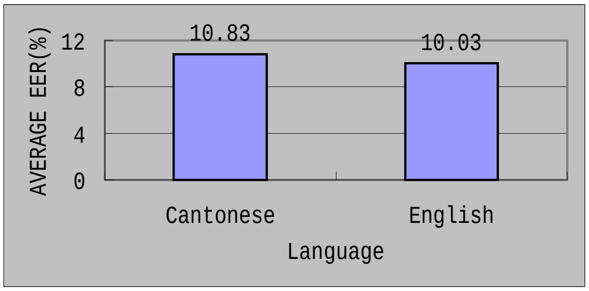

Fig.3 Results on the Cantonese and English databases.

Then, we further test the GMM on the TD_HK_CANT and TD_HK_ENG databases. The results are depicted in Fig. 3. Apparently, the similar performance is achieved by GMM for Cantonese and English in terms of text-dependent case.

Since those databases are designed for multilingual cases by speech uttered by the same speakers in different languages, our simulation results indicate that GMM is insensitive to languages for text-dependent case even though training samples are selected at random.

4.2 Results for Text-Independent Case

GMM has been successfully applied to text-independent speaker recognition [2][3]. For the same purpose, we have designed a pair of bilingual databases, TI_BJ_MAND and TI_BJ_ENG. In our simulations, we use a baseline GMM consisting of 32 components for all the speakers in different languages.

Fig. 4 illustrates the performance of such a GMM on the TI_BJ_MAND and TI_BJ_ENG databases. Obviously, averaging EER on two databases is very similar.

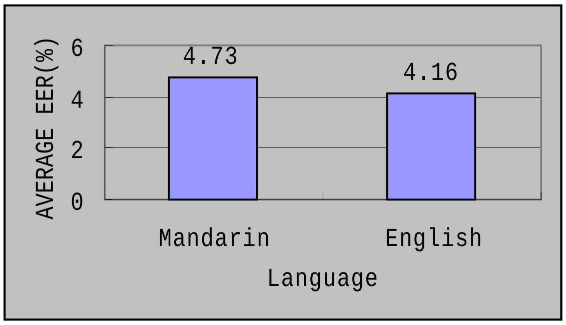

Fig.4 Results on the Mandarin and English databases.

Results here and others reported elsewhere [5] show that GMM produces good performance almost regardless of languages for text-independent case.

Our simulation results reported above show that GMM produces very similar performance for pair databases that consist of the same speakers in different languages and are recorded in the same sessions. With regard to performance, the slight difference between multilingual databases might be caused by many factors, e.g. uttering difference and psychological change when a speaker talks in native and foreign languages. For GMM itself, it captures only statistical characteristics of a speaker based on a spectral feature in our simulations. Therefore, the slight performance difference on different pair database does not prevent from our following argument: GMM can be used as a unified model for multilingual speaker verification.

5. CONCLUSION

We have presented our empirical study on use of GMM for multilingual speaker verification. Our simulation results shows that a fixed structural GMM model can perform very well for speaker verification regardless of different dialects and working styles (either text-independent or text-dependent), which suggests that GMM can be a unified model for multilingual speaker verification. In particular, GMM performs very well in the text-dependent case, to our best knowledge, which has not been reported in literature. On the basis of our empirical study, we suggest that GMM would be able to be a proper model for multilingual speaker verification, which provides an ease-to-use way for building a multilingual speaker verification system.

ACKNOWLEDGMENTS

We would like to thank L. Wang for providing her source code for the GMM implementation and the human-machine interactive laboratory at The Chinese University of Hong Kong for providing the TD_HK_CANT and TD_HK_ENG databases. This research was partially supported by a National Key Fundamental Research Project Grant (G1999032708) in China.

REFERENCES

[1] H. Meng et al, “ISIS: A multilingual spoken dialog system developed with CORBA and KQML agents,” Proceedings of International Conference on Spoken Language Processing, Beijing, 2000. (in press)

[2] D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph.D. Thesis, Georgia Institute of Technology, 1992.

[3] D. A. Reynolds, “Speaker identification and verification using Gaussian Mixture Speaker Models,” ESCA Workshop of Automatic Speaker Recognition, Identification and Verification, Martigny Switzerland, April 5-7, 1994, pp. 27-30.

[4] X. K. Qing, *On Use of GMM for Multilingual Speaker Verification: An Empirical Study*, M.E. Thesis, National Laboratory of Machine Perception and The Center for Information Science, Peking University, 2000. (in Chinese)

[5] K. Chen, L. Wang, and H.S. Chi, “Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification,” *International Journal of Pattern Recognition and Artificial Intelligence* 11(3): 417-445, 1997.