IEICE TRANS. INF. & SYST., VOL.E92–D, NO.4 APRIL 2009

LETTER

Using a Kind of Novel Phonotactic Information for SVM Based Speaker Recognition*

Xiang ZHANG[†a)], *Nonmember*, **Hongbin SUO**[†], *Student Member*, **Qingwei ZHAO**[†], *and* **Yonghong YAN**[†], *Nonmembers*

SUMMARY In this letter, we propose a new approach to SVM based speaker recognition, which utilizes a kind of novel phonotactic information as the feature for SVM modeling. Gaussian mixture models (GMMs) have been proven extremely successful for text-independent speaker recognition. The GMM universal background model (UBM) is a speaker-independent model, each component of which can be considered as modeling some underlying phonetic sound classes. We assume that the utterances from different speakers should get different average posterior probabilities on the same Gaussian component of the UBM, and the supervector composed of the average posterior probabilities on all components of the UBM for each utterance should be discriminative. We use these supervectors as the features for SVM based speaker recognition. Experiment results on a NIST SRE 2006 task show that the proposed approach demonstrates comparable performance with the commonly used systems. Fusion results are also presented.

key words: *speaker recognition, Gaussian mixture model, universal background model, support vector machine*

1. Introduction

Currently, the dominant features used for speaker recognition are cepstral features extracted over short time spans. The modeling of these features is typically based on log-likelihood ratio of Gaussian mixture models (GMMs), or discriminatively using support vector machines (SVMs). Gaussian mixture modeling-universal background model (GMM-UBM) [1] and Gaussian supervectors-SVM (GSV-SVM) [2] are two basic and commonly used approaches for speaker recognition. The fusion of these two systems shows excellent performance in task of speaker recognition. However, both of the approaches use only short-term cepstral features. They ignore high-level information, such as the particular word usage or the acoustic variability of phonetic events when comparing different speakers.

Phone-conditioned [3] and word-specific [4] cepstral models are a direct attempt to make models invariant to the choice of words. The work in [3] utilized a speech recognition front-end to hypothesize the phonetic content of the utterance and then phone dependent models were refined in place of a global GMM. The research in [4] focused on using word units from a speech recognizer as acoustic unit, and aimed to compare the same acoustic unit as spoken by different speakers. However, these approaches have the drawback of fragmenting the speech data and require sufficiently accurate speech recognition.

In this letter, a new approach for speaker recognition which uses a kind of phonotactic information for modeling is presented. The Gaussian components of the UBM can be considered as modeling some underlying phonetic sounds [5]. We assume that the utterances from different speakers should get different average posterior probability on the same Gaussian component. This reflects how the pronunciation of the same sound unit may differ from one speaker to another. Thus, we concatenate these average posterior probabilities on all components of the UBM into a vector. Together with the SVM modeling technique, this vector can be used as the feature for discriminative classification. This system is much less computationally complex than the traditional phonotactic systems based on speech recognition.

This paper is organized as following: In Sect. 2, we give a simple review of two basic approaches for speaker recognition. In Sect. 3, our proposed system is presented in detail. Section 4 gives out the evaluation corpus and experiment results. Finally, we conclude in Sect. 5.

2. Baseline Systems

This section briefly reviews two basic speaker recognition systems, which serve as the baseline systems in this work. The first baseline system is a GMM-UBM [1], in which speaker models are trained by adapting from a UBM using maximum a posteriori (MAP) adaptation [6]. Only means of Gaussian components are adapted.

The second baseline system is GMM mean supervectors system followed by support vector machines (GSV-SVM). GSV-SVM is based on GMMs which differ only in means. The mean vectors of all mixture components are concatenated to form one supervector for each utterance. (Each mean is normalized by the corresponding standard deviation) [7]. We use these supervectors for SVM modeling.

Manuscript received September 16, 2008.
Manuscript revised November 17, 2008.

[†]The authors are with the ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, China.

[*]This work is partially supported by MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030). The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195).

a) E-mail: xzhang@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E92.D.746

3. Outline of the Proposed Approach

3.1 Phonotactic Feature Extraction and SVM Modeling

In GMM-UBM speaker recognition system, the UBM is a weighted linear combination of M Gaussian component densities. For a D-dimensional feature vector, $\mathbf{x}$, the formula of the UBM is defined as

$$p(\mathbf{x}|\lambda_{UBM}) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}) \tag{1}$$

where λ_{UBM} represents the parameters of the UBM, $p_i(\mathbf{x})$, $i = 1, \ldots, M$, are the Gaussian component densities with mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$, M is the number of mixtures in the UBM, and w_i, $i = 1, \ldots, M$ are the mixture weights. Thus, the UBM model can be denoted as:

$$\lambda_{UBM} = \{w_i, \mu_i, \mathbf{\Sigma}_i\},\ i = 1, \ldots, M \tag{2}$$

The UBM is a speaker-independent background model, and each Gaussian component of the UBM can be considered as modeling a kind of underlying broad phonetic sounds. Thus, all the Gaussian components of the UBM can be considered to characterize a speaker-independent voice. Given some feature vectors from the hypothesized speaker, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, we first calculate the probabilistic alignment of the vectors into the UBM mixture components. That is, for mixture i in the UBM, the posterior probability is computed as following:

$$P(i|\mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{M} w_j p_j(\mathbf{x}_t)} \tag{3}$$

The posterior probability is the normalized likelihood ratio, so it can be seen as a kind of similarity. The larger the posterior probability is, the better the mixture can be used to represent that feature vector. Thus, the average posterior probability for each mixture can well represent the similarity between the hypothesized speaker's voice and the speaker-independent voice. The average posterior probability for mixture i is calculated as following:

$$b_i = \frac{1}{T} \sum_{t=1}^{T} P(i|\mathbf{x}_t) \tag{4}$$

Therefore, when given a hypothesized speaker's one utterance, all the posterior probabilities for the Gaussian components of the UBM can be considered as the voice-characteristic-discrepancy in the form of a probabilistic vector between the hypothesized speaker and the universal speaker (modeled by the speaker-independent model). These posterior probabilities are concatenated into a supervector:

$$\mathbf{b} = [b_1, b_2, \ldots, b_M]^t \tag{5}$$

The elements of this supervector can be seen as a histogram describing the characteristic of speech sound units for a hypothesized speaker's speech utterance. The supervectors from the same speaker are always similar, and different speakers would exhibit different patterns of histogram. This fact can be illustrated by Fig. 1. The utterances used are from NIST 2006 SRE corpus. The top two pictures are calculated using two different utterances from the same speaker, and the below ones are from another speaker. We can see that histograms from the same speaker are more similar than that from different speakers, such as in the labeled area, the pictures on the same row are quite similar, and the pictures on the same column which are from different speakers are less similar. SVMs are quite sensitive to this inconsistency.

Thus, this phonotactic information can be used as the feature for SVM based speaker recognition. The procedure for extracting the phonotactic features is shown in Fig. 2. $\mathbf{X}$ represents the cepstral feature vectors for a given utterance, and λ_i, $i = 1, \ldots, M$ are the Gaussian mixtures of the UBM. Using the cepstral features and the UBM, the average posterior probability for each mixture can be computed. We concatenated all these values into a supervector for SVM classification.

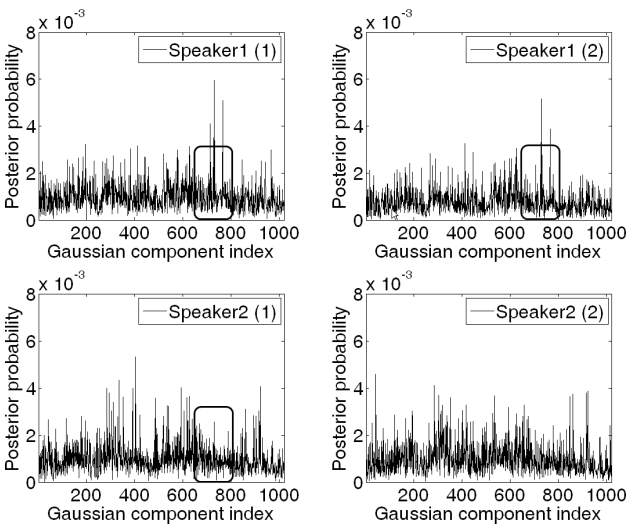

Fig. 1 Comparison of two speakers' phonotactic supervectors. (The x-axis is the Gaussian mixture index of the UBM, and the y-axis refers to the value of the posterior probability on each mixture.)

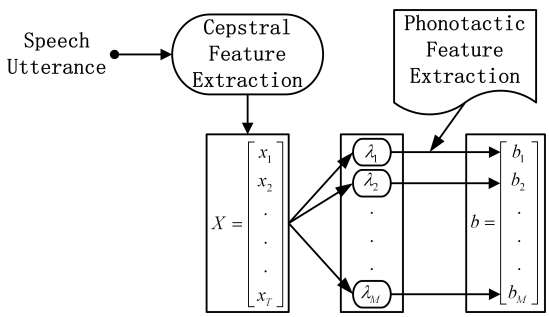

Fig. 2 The procedure of the phonotactic feature extraction.

Furthermore, the proposed algorithm is much less computationally complex than the traditional system based on phone-conditioned or word-specific models, and it is also faster than the GMM-UBM and GSV-SVM systems.

3.2 Further Improvement

When the large UBM is evaluated for a feature vector, only a few of the mixtures contribute significantly to the posterior value. This is because the UBM represents a distribution over a large space but a single cepstral feature vector will be near only a few components of the UBM. Thus, only the top C largest scoring mixture components are used for calculating the average posterior values, where C is an empirical value.

3.3 Selection of the SVM Kernel

After mapping the cepstral features to the phonotactic supervectors for SVM modeling and testing, the key issue is to find an appropriate SVM kernel that compares a supervector with others efficiently. Let N_{tgt} denote the number of frames of the target's data, and N_{non} denote the total number of frames of the background's data. According to [8], for $N_{non} \gg N_{tgt}$, after some feasible assumptions for simplification, the generalized linear discriminant sequence kernel (GLDS) is defined as following:

$$K_{GLDS}(\mathbf{X}, \mathbf{Y}) = \mathbf{b}(\mathbf{X})^t \mathbf{R}_{non}^{-1} \mathbf{b}(\mathbf{Y}) \tag{6}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are two sequences of speech cepstral feature vectors, and $\mathbf{b}(\mathbf{X})$ and $\mathbf{b}(\mathbf{Y})$ are the corresponding SVM supervectors. $\mathbf{R}_{non}$ is an $M * M$ matrix that is important to the performance of the SVM classification, where M is the number of the Gaussian mixtures of the UBM. $\mathbf{R}_{non}$ is calculated based on the background data set. In practice, it is useful to calculate only the diagonal of $\mathbf{R}_{non}$ [8]. According to [8], when using the proposed phonotactic supervectors as the SVM features, the i-th element on the diagonal is $1/N_{non} \sum_{j=1}^{N_{non}} P^2(i|\mathbf{z}_j)$, where $\mathbf{z}_j$ is the j-th frame of the background cepstral features. Thus, when adopting the GLDS kernel with this $\mathbf{R}_{non}$, the proposed phonotactic features should be discriminative.

For further simplification, we can rewrite Eq. (6) as following:

$$K_{GLDS}(\mathbf{X}, \mathbf{Y}) = (\mathbf{U}\mathbf{b}(\mathbf{X}))^t (\mathbf{U}\mathbf{b}(\mathbf{Y})) \tag{7}$$

where $\mathbf{U} = \mathbf{R}_{non}^{-1/2}$. In our work, the background data is a large subset of the data for training UBM. The data assigned to each mixture of the UBM is sufficiently adequate and well balanced for computing $\mathbf{U}$. We find that the values on the diagonal of $\mathbf{U}$ yield small dynamic range. Thus, in our work, we suppose $u_{ii} \approx u_{kk}$, for all $i \neq k$, where u_{ii} is the i-th diagonal element of matrix $\mathbf{U}$. Then, $\mathbf{U}$ can be viewed as an identity matrix weighted by a constant which can be ignored without influencing the total performance of the SVM classification. This simplification reduces the computational complexity dramatically and gives a very concise way of training and scoring. Therefore, in our actual work, a linear kernel ($K(\mathbf{X}, \mathbf{Y}) = \mathbf{b}(\mathbf{X})^t \mathbf{b}(\mathbf{Y})$) is finally used for SVM based speaker recognition, which shows satisfying performance in the experiments.

4. Experiments

4.1 Experimental Setup

We performed experiments on the 2006 NIST speaker recognition (SRE) corpus. We focused on the single-side 1 conversation train, single-side 1 conversation test, and the multi-language handheld telephone task (the core test condition) [9]. We used equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for evaluation [9].

For cepstral feature extraction, a 20 ms Hamming window with 10 ms shifts is used. Each utterance is converted into a sequence of 36-dimensional feature vectors, each consisting of 12 MFCC coefficients and their first and second derivatives. An energy-based speech detector is applied to discard vectors from low-energy frames. To mitigate channel effects, feature warping, cepstral mean subtraction and variance normalization are applied to the features.

The UBM consists of 2048 mixture components. For GMM MAP training, we adapt only the means with a relevance factor of 12. The UBM is trained using EM with the data from the corpora: NIST 01, NIST 02, NIST 04 and NIST 05. The background training set consisted of 1744 conversation sides from 340 speakers. These data are also chosen from the same corpora.

In both GSV-SVM and the proposed system, SVMTorch [10] with a linear inner-product kernel function is used for SVM training. In our experiments, the size of SVM features are 36*2048 for GSV features and 2048 for the proposed phonotactic features.

4.2 Experimental Results

When calculating the phonotactic features, only top C largest scoring mixture components for each cepstral feature frame are used. We compare the results of the system with different value of C in Table 1.

From Table 1, we can see that the proposed system achieves the best performance when C is tuned to be 250. This improvement can bring gains of 5.6% EER and 1.2%

Table 1 Performance for proposed system with different values of C.

top C	EER (%)	minDCF
total	**9.58**	**0.0408**
top 50	9.32	0.0431
top 100	9.22	0.0411
top 150	9.17	0.0409
top 200	9.12	0.0405
top 250	**9.04**	**0.0403**
top 300	9.07	0.0405

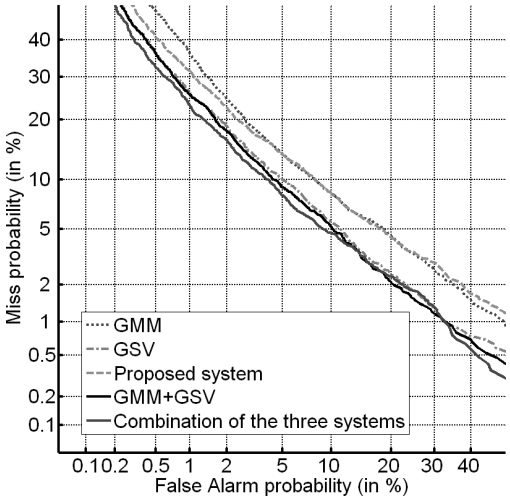

Fig. 3 DET curves for the proposed approach, GMM-UBM, GSV-SVM and various combinations.

Table 2 EER and minDCF for different systems.

system	EER (%)	minDCF
GMM (a)	9.01	0.0436
GSV (b)	7.37	0.0354
Proposed System (c)	9.04	0.0403
Baseline (a+b)	**7.06**	**0.0353**
All-combination (a+b+c)	**6.47**	**0.0323**

minDCF compared to the system using all the mixture components for calculating average posterior probabilities (total).

Figure 3 shows the detection error tradeoff (DET) curves for the various systems. We compare the proposed system with standard GMM-UBM and GSV-SVM systems. The exact values of EER and minDCF for different systems are shown in Table 2. It can be seen that the proposed system produces comparable performance to the two systems. Better performance can be obtained when combining the proposed system with the two standard cepstral systems. The system combining all the three approaches can achieve 8.4% relative improvement in EER and 9.3% relative improvement in minDCF, respectively, compared to the system only combining two baselines (a+b).

5. Conclusions

We have proposed a novel approach to speaker recognition. It uses a kind of phonotactic information as the feature for SVM modeling. This approach has been shown to have good performance on a NIST SRE 2006 task. Performance is found to be competitive with the commonly used GMM-UBM and GSV-SVM systems. Relative gains of 8.4% EER and 9.3% minDCF are obtained when combining the proposed approach with a system that integrates the two baselines.

References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol.10, no.1-3, pp.19–41, 2000.
[2] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol.13, no.5, pp.308–311, 2006.
[3] A. Park and T. Hazen, "ASR dependent techniques for speaker identification," Seventh International Conference on Spoken Language Processing, ISCA, 2002.
[4] D. Sturim, D. Reynolds, R. Dunn, and T. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," Acoustics, Speech, and Signal Processing, 2002. Proceedings (ICASSP'02), IEEE International Conference on, vol.1, 2002.
[5] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol.3, no.1, pp.72–83, 1995.
[6] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[7] P. Matejka, L. Burget, P. Schwarz, O. Glembek, M. Karafiat, F. Grezl, J. Cernocky, D. van Leeuwen, N. Brummer, and A. Strasheim, "STBU system for the NIST 2006 speaker recognition evaluation," Proc. ICASSP, 2007.
[8] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Comput. Speech Lang., vol.20, no.2-3, pp.210–229, 2006.
[9] "The NIST year 2006 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2006/index.html, 2006.
[10] R. Collobert, S. Bengio, and R. Williamson, "SVMTorch: Support vector machines for large-scale regression problems," J. Machine Learning Research, vol.1, no.2, pp.143–160, 2001.