Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Text-Independent Speaker Verification via State Alignment

Zhi-Yi Li, Wei-Qiang Zhang, Wei-Wei Liu, Yao Tian, Jia Liu

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
lizhiyi06@mails.tsinghua.edu.cn, wqzhang@tsinghua.edu.cn, liu-ww10@mails.tsinghua.edu.cn

Abstract

To model the speech utterance at a finer granularity, this paper presents a novel state-alignment based supervector modeling method for text-independent speaker verification, which takes advantage of state-alignment method used in hidden Markov model (HMM) based acoustic modeling in speech recognition. By this way, the proposed modeling method can convert a text-independent speaker verification problem to a state-dependent one. Firstly, phoneme HMMs are trained. Then the clustered state Gaussian Mixture Models (GMM) is data-driven trained by the states of all phoneme HMMs. Next, the given speech utterance is modeled to sub-GMM supervectors in state level and be further aligned to be a final supervector. Besides, considering the duration differences between states, a weighting method is also proposed for kernel based support vector machine (SVM) classification. Experimental results in SRE 2008 core-core dataset show that the proposed methods outperform the traditional GMM supervector modeling followed by SVM (GSV-SVM), yielding relative 8.4% and 5.9% improvements of EER and minDCF, respectively.

1. Introduction

Text-independent speaker verification refers to determining whether the claim of identity is correct or incorrect to a text-unknown speech utterance. Nowadays, Gaussian mixture model (GMM) adapted from universal background model (UBM), as a classic method to cover the space of acoustic speech context, have been commonly used in speaker verification [1, 2]. This method can implicitly align the speech content to its corresponding mixture of UBM through maximum a posteriori probability (MAP) adaptation. However, in many practical applications, the acoustic components in training or testing speech data are limited even short, leading to the inadequate mixture cover. To make it up, some phonetic based methods [3, 4, 5] and some text-constraint based methods [6, 7] are proposed at a finer granularity. One typical work is phonetic GMM (PGMM), which models the sub-GMM-UBM systems for phonemes and does score fusion at final decision stage, performing slightly better than the GMM baseline [8, 9]. Another typical work estimates an MLLR transform per acoustic class to model speakers' characteristics [10, 11]. In fact, from the viewpoint of acoustic phoneme modeling by hidden Markov model (HMM) in speech recognition, different phoneme HMMs always share some common states and these states are considered to be the basic modeling units of speech context and can reflect more fundamental granularity.

On the other side, modeling the speech utterance to be a vector or supervector has been proved to be an efficient and popular way to present a varying number of feature vectors by a single vector, such as the input to support vector machine (SVM) [2, 12]. Among several proposed vector modeling methods , GMM supervector, which is derived by bounding the Kullback-Leibler (KL) divergence measure between GMMs, is still commonly used in practice so far, due to its well-done performance and simplicity, even though the i-vector based system can perform better in latest NIST speaker recognition evaluations (SRE) [13, 12].

Due to these considerations, this paper firstly present a state alignment based supervector modeling method for text-independent speaker verification. The proposed method try to convert a text-independent speaker verification problem to be a state-dependent one by taking advantage of state-alignment technologies commonly used in speech acoustic modeling. Firstly, phoneme HMMs are trained. Secondly, the clustered state Gaussian Mixture Models (GMM) is data-driven trained by the states of all phoneme HMMs. Next, the given speech utterance is modeled to sub-GMM supervectors in state level and be further aligned to be a final supervector. Besides, considering the duration differences between states, a weighting method is also proposed for kernel computation. In this paper, we use the SVM as the classifier for state-aligned supervectors because of its well robustness and simplicity without affecting its extensibility.

The paper is organized as follows. In Section 2, the proposed state aligned supervector modeling method is presented in details. Section 3 introduces the application as input to SVM classifier in text-independent speaker verification. In Section 4, experimental results are presented. Section 5 concludes the paper and outlines areas for future work.

2. State alignment based supervector modeling

2.1. State alignment based supervector modeling

At the beginning, we need to train the phoneme HMMs using Baum-Welch algorithm by some speech data with transcripts. Then, we train the state GMMs by data-driven clustering from phone HMMs as shown in Fig. 1. After that, all utterances of training and testing are decoded to state-labeled transcripts using the Viterbi HMM decoder.

After the state GMMs are well trained, each state-dependent Universal Background Model (UBM) is obtained through Maximum A Posterior (MAP) adaptation from a common state-independent UBM. In this process, each physical state is treated as a cluster and the data with the same physical state labels are used to train a state-dependent UBM. Then, for every utterance with its state labels and the common state UBMs, their state GMMs can be obtained through MAP adaptation. At the end, sub GMM supervectors of all the states

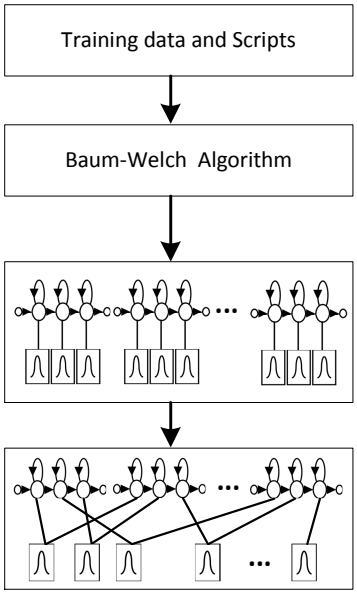

Figure 1: The process of training clustered state models.

are aligned and stacked to be a final state aligned supervector. The whole process is shown as Fig. 2.

Let's suppose that the number of state models is S and the feature dimension is D. the i-th state is modeled by a D-dimension GMM denoted as $\lambda_i = \{\omega_{i,j}, \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}; j = 1, \ldots, M_i\}$, as shown in (1):

$$p(\boldsymbol{x}|\lambda_i) = \sum_{j=1}^{M_i} \omega_{i,j} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}), \tag{1}$$

where M_i is the mixture number of Gaussian components. The mixture weights $\omega_{i,j}$ satisfies the constraint $\sum_{j=1}^{M_i} \omega_{i,j} = 1$. The D-dimension Gaussian density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be expressed as in (2):

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2} \right\}. \tag{2}$$

We also modeling the Gaussian supervector of every state by bounding the KL divergence measurement between two GMMs derived by Campbell [13]. And we can get the state aligned supervector $\boldsymbol{v}$ by stacking all S sub Gaussian supervectors state by state as shown in (3) and (4):

$$\boldsymbol{v} = [\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, \cdots, \boldsymbol{v}_S^T]^T, \tag{3}$$

$$\boldsymbol{v}_i = [\sqrt{\omega_{1,i}} \boldsymbol{\Sigma}_{i,1}^{(-1/2)} \boldsymbol{\mu}_{i,j}^T, \cdots, \sqrt{\omega_{i,M_i}} \boldsymbol{\Sigma}_{i,M_i}^{(-1/2)} \boldsymbol{\mu}_{i,M_i}^T]^T. \tag{4}$$

From the implementation point of view, this just means that all the Gaussian means need to be normalized before stacked into supervector.

2.2. Duration weight supervector modeling

Considering the duration difference between states, this paper also proposes a duration weight supervector modeling method for classification. The process of constructing weight supervector as follows. Given the i-th state specific UBM and

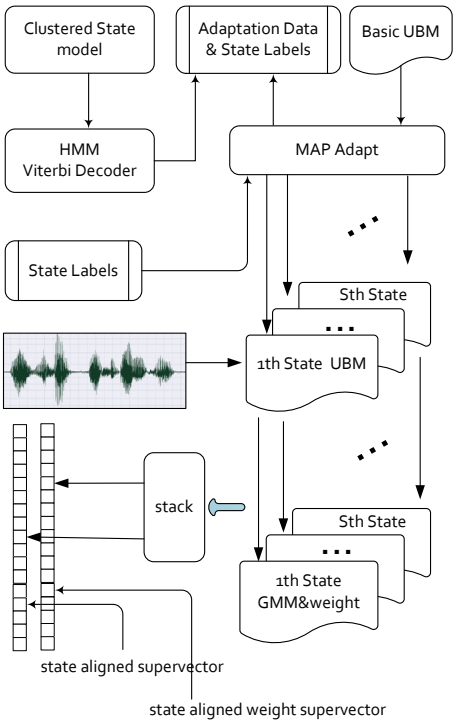

Figure 2: The process of state-aligned supervector modeling.

feature vector $\boldsymbol{x}_t$ from the utterance, we can first determine the probabilistic alignment of the feature vector $\boldsymbol{x}_t$ into the j-th UBM mixture component of i-th state as shown in (5):

$$\Pr(j\text{-th mix}|\boldsymbol{x}_t, i\text{-th state}) = \frac{\omega_{i,j}\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j})}{\sum_{k=1}^{M_i} \omega_{i,k}\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k})}. \tag{5}$$

Then the zeroth-order sufficient statistic by $n_{i,j}$ using $\Pr(\boldsymbol{x}_t)$ can be computed in (6):

$$n_{i,j} = \sum_{t=1}^{T} \Pr(j\text{-th mix}|\boldsymbol{x}_t, i\text{-th state}). \tag{6}$$

And the weight value $w_{i,j}$, which stands for the contribution to the j-th mixture components in i-th state of duration information, can yield by $n_{k,i}$ as in (7):

$$w_{i,j} = (\frac{n_{i,j}}{n_{i,j} + \gamma})^{1/2}, \tag{7}$$

where γ is a fixed factor for weight scaling. Then, we can get the weight supervector $\boldsymbol{w}$ by stacking all $w_{i,j}$ of all S states as shown in (8) and (9):

$$\boldsymbol{w} = [\boldsymbol{w}_1^T, \boldsymbol{w}_2^T, \cdots, \boldsymbol{w}_S^T]^T, \tag{8}$$

$$\boldsymbol{w}_i = [w_{i,1}\mathbf{1}, w_{i,2}\mathbf{1}, \cdots, w_{i,M_i}\mathbf{1}]^T, \tag{9}$$

which is the weight supervector of i-th state, $\mathbf{1}$ is a D-dimension vector, in which all elements are 1. This just means that the Gaussian supervector of every state need to be weighted by their duration information before putting them into a classifier.

3. Application as input to SVM classifier

The typical application of supervector is using them as input to support vector machine (SVM) [13] and it is still one of the good classifier for speaker verification and is used commonly in practice, even though some current technology like i-vector can perform better in latest NIST speaker recognition evaluation (SRE). In this paper, we still use the SVM as the classifier for proposed method because of its well robustness and simplicity without affecting its extensibility. As one of the most robust classifiers for speaker verification, SVM is a binary classifier which models the decision boundary between two classes as a separating hyperplane as shown in (10) and (11):

$$f : \mathbb{R}^N \mapsto \mathbb{R}$$
$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x} + b, \tag{10}$$

$$f(\boldsymbol{x}) = \sum_k \boldsymbol{a}_k^T K(\boldsymbol{x}, \boldsymbol{x}_k) + b. \tag{11}$$

One of the most important thing in SVM is the selection of the kernel function. The kernel function $k(\boldsymbol{x}, \boldsymbol{y})$ is designed so that it can be expressed to $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$, satisfied to the Mercer's theorem, where $\phi(\cdot)$ is a mapping function from the input space to kernel feature space of high dimensionality.

In our work, we select the linear kernel for fair comparison with the well-known baseline GSV-SVM, except that combining the weight supervectors in kernel construction. Furthermore, because of having weight supervectors for matching the state-aligned training supervector and testing utterance, we need to train the SVM model for every trial pair. In kernel function, $\phi(\boldsymbol{x})$ and $\phi(\boldsymbol{y})$ can be rewritten as in (12) and (13), respectively:

$$\phi(\boldsymbol{x}) = \boldsymbol{w}_x \circ \boldsymbol{w}_y \circ \boldsymbol{x}, \tag{12}$$

$$\phi(\boldsymbol{y}) = \boldsymbol{w}_x \circ \boldsymbol{w}_y \circ \boldsymbol{y}, \tag{13}$$

where the operator $\circ$ denotes as element-wise multiplication of vectors.

4. Experimental results

4.1. Experimental setup

In this paper, all experiments are carried out on NIST SRE 2008 telephone male dataset for both training and testing. The core condition is named short2-short3 [14].

In experimental setup, 39-dimension Mel Frequency Cepstral Coefficient (MFCC) feature vectors (13 static + Δ + $\Delta\Delta$) are extracted from the speech signal at frame shift 10 ms with 20 ms Hamming window and are subjected to feature warping.

The standard GSV-SVM system is built as the baseline. A 1024 mixture UBM is trained using SRE2004 1-side training dataset. Speaker models are obtained by maximum a posteriori (MAP) adaptation.

We use the Switch Board I data to train 47 phonemes HMMs with 3 valid states. Each state is modeled to 32-mixture GMM by HTK tools so that all the 3*47 logical states GMMs are clustered to 32 physical state GMMs.

We use LibSVM interface in Shogun toolkit [15] as our SVM classifier. HVite is used to decode all the speech utterances in SRE dataset [16]

The system performance are evaluated in terms of Equal Error Rate (EER) and Minimum Detection Cost Function (MinDCF) [14].

4.2. Results and discussions

We first give the duration distribution of states after decoding all utterances in the dataset as shown in Fig.3. It can be seen that different states actually have obviously different durations. So it is very necessary to weighting the supervector during training and testing.

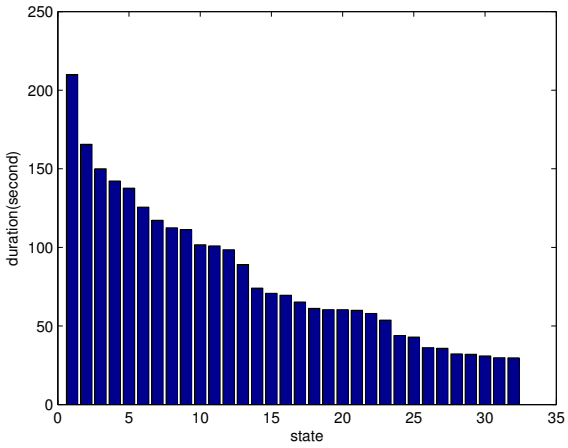

Figure 3: The duration of per state in data set.

Firstly, without considering the affection of language, the performance in SRE 2008 tel-tel English trial is shown in Table 1. The proposed method is not as good as the baseline without weighting. Through adjusting the weight factor to 80, experimental results show that the proposed supervector modeling method consistently outperforms the traditional method yielding relative 8.4% and 5.9% improvements of EER and minDCF, respectively.

Table 1: Performance comparison on tel-tel English dataset.

System	EER (%)	minDCF (%)	weight factor (γ)
Baseline	5.23	2.71	–
State aligned	6.43	3.17	–
State aligned	5.38	2.65	8
State aligned	4.98	2.53	64
State aligned	4.93	2.57	70
State aligned	**4.79**	**2.55**	**80**
State aligned	5.38	2.78	90

In addition, we also evaluate the performance in SRE 2008 tel-tel dataset. As shown in Table 2. It can also be presented that the proposed supervector method is comparable with the baseline. The reason of gain in Table 2 not as good as it in Table 1 may be that different language affects the accuracy of labels during the state decoder. Thus, we need to add some language compensation method to make it up.

As shown in the experimental result, one problem is that the weighting factors γ plays an important role in the approach such that the best gamma value need to be trained using the development dataset and then fixed and tested in the testing dataset in practical application.

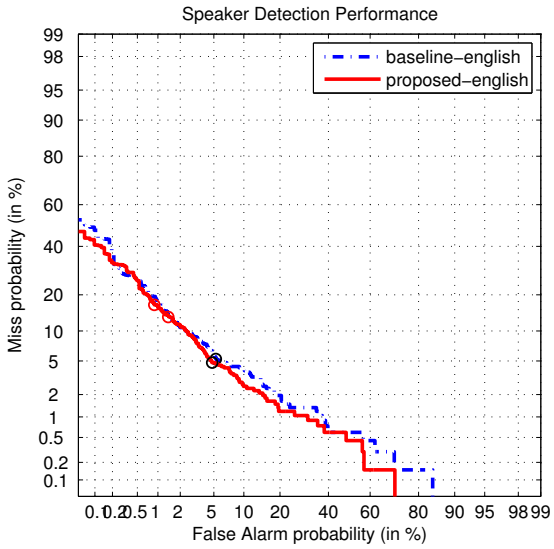

Figure 4: DET Performance comparison of baseline and proposed methods on tel-tel English dataset.

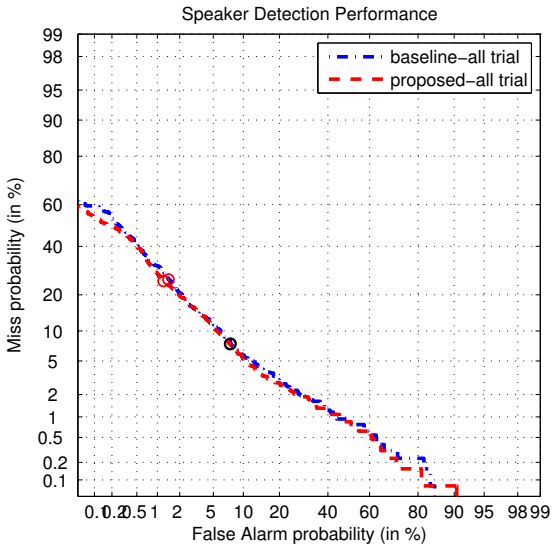

Figure 5: DET Performance comparison of baseline and proposed methods on tel-tel dataset .

Table 2: Performance comparation on tel-tel dataset .

System	EER (%)	minDCF (%)	weight factor (γ)
Baseline	**7.54**	3.97	–
State aligned	7.73	3.86	70
State aligned	7.62	**3.84**	80
State aligned	8.04	3.98	90

5. Conclusion and future work

In order to model the speech utterance at a finer granularity, this paper presents a novel state-alignment based supervector modeling method for text-independent speaker verification, which takes advantage of the state-alignment method in hidden Markov model (HMM) based acoustic modeling in speech recognition. The sub-supvectors obtained by data-driven clustered states are stacked to be a final state-alignment supervector. By this way, the proposed modeling method can convert a text-independent speaker verification problem to a state-dependent one. In addition, considering the duration differences between states, a weighting method is also proposed for kernel. In this paper, we still use the SVM as the classifier for proposed method because of its well robustness and simplicity without affecting its extensibility. Experimental results in SRE 2008 tel-tel English dataset show that the proposed methods outperform the traditional GMM supervector modeling followed by SVM (GSV-SVM), yielding relative 8.4% and 5.9% improvements of EER and minDCF, respectively. In the future, we intend to extend the proposed state-alignment idea to factor analysis such as i-vector based text-independent speaker verification.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 61370034, No. 61273268 and No. 61005019.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[3] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *Proc. NIPS*, Lake Tahoe, Dec. 2003, pp. 1377–1384.

[4] R. Faltlhauser and G. Ruske, "Improving speaker recognition performance using phonetically structured Gaussian mixture models," in *Proc. Eurospeech*, Scandinavia, Sept. 2001, pp. 751–754.

[5] R. Hansen, E. Slyh and T. Anderson, "Speaker recognition using phoneme-specific GMMs," in *Proc. Odyssey*, Toledo, May 2004, pp. 179–184.

[6] K. Boakye and B. Peskin, "Text-constrained speaker recognition on a text-independent task," in *Proc. Odyssey*, Toledo, May 2004, pp. 129–134.

[7] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Proc. IEEE Workshop on Signal Processing Applications for Public Security and Forensics (SAFE'07)*, Washington, April 2007, pp. 39–43.

[8] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1969–1978, Sept. 2007.

[9] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. Odyssey*, San Juan, June 2006.

[10] A. Stolcke, Sachin S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1987–1998, Sept. 2007.

[11] M. Ferras, C.-C. Leung, C. Barras, and J. Gauvain, "Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1366–1378, Aug. 2010.

[12] N. Dehak, P. Kenny, and P. Ouellet, "Front end factor analysis for speaker verification.," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[13] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, Toulouse, May 2006, vol. 1, pp. 97–100.

[14] National Institute of Standards and Technology, "The NIST Year 2008 Speaker Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/sre/2008/index.html.

[15] S. Sonnenburg, G. Raetsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. Bona, A. Binder, C. Gehl, and V. Franc., "The SHOGUN machine learning toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, June 2010.

[16] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (version 3.4)*, Cambridge University Engineering Department, Cambridge, UK, 2006.