SESSION VARIABILITY SUBSPACE PROJECTION BASED MODEL COMPENSATION FOR SPEAKER VERIFICATION

Jing Deng, Thomas Fang Zheng and Wenhu Wu

Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University, Beijing, 100084, P.R. China
dengj02@mails.tsinghua.edu.cn, {fzheng, wuwh}@tsinghua.edu.cn

ABSTRACT

In this paper, a session variability subspace projection （SVSP） based model compensation method for speaker verification is proposed. During the training phase the session variability is removed from speaker models by projection, while during the testing phase the session variability in a test utterance is used to compensate speaker models. Finally, the compensated speaker models and UBM are used to recognize the identity of the test utterance. Compared with the conventional GMM-UBM system, the relative equal error rate reduction of SVSP is 16.2% on the NIST 2006 single-side one conversation training, single-side one conversation test.

Index Terms— Speaker recognition

1. INTRODUCTION

Though research in speaker recognition has made a great progress, mismatch caused by session variability is still a big factor leading to recognition errors. Session variability includes phenomena such as transmission channel effects, transducer characteristics, background noise, and intra-speaker variability.

Lots of methods have been proposed to solve this issue which can be categorized into three domains: feature domain, model domain and score domain. In the feature domain, typical methods are cepstral mean subtraction [1], RASTA filter [2], feature warping [3] and feature mapping [4], *etc*.; In the model domain, typical methods are speaker model synthesis [5], factor analysis [6,7] and nuisance attribute projection (NAP) [8], *etc*.; In the score domain, typical methods are Hnorm [9], Tnorm [10] and Znorm [11], *etc*..

Factor analysis and NAP are two recently proposed and now very popular methods that have provided impressive reductions in verification error rates [6-8]. Though NAP greatly reduces the complex of session variability computation compared with factor analysis, it cannot be used for GMM-UBM system directly. In this paper, the idea of projection in NAP and the idea of model compensation in factor analysis will be combined together to form a new method called *session variability subspace projection* (SVSP) *based model compensation*. The main idea of SVSP is to use the session variability in a test utterance to compensate speaker models whose session variability has already been removed during the training phase. On the one hand, it simplifies the computation of session variability by projection; on the other hand, it can be easily used for GMM-UBM systems.

This paper is organized as follows. The SVSP based model compensation method will be presented in Section 2. In Section 3, experiments and results will be described. Finally, conclusions and perspectives will be given in Section 4.

2. SVSP BASED MODEL COMPENSATION

The basic idea of SVSP can be easily seen from Figure 1 which consists of four steps: estimation of session variability subspace, speaker model training, speaker model compensation and test utterance verification. We will detail them in the following sections.

2.1 Estimation of Session Variability Subspace

Given a speaker's Gaussian mixture model, a GMM supervector can be formed by concatenating the GMM component mean vectors [6-8]. The supervector is a sum of a session-independent supervector with an additional session-dependent supervector [8], which can be described as

$$M(s,i) = m(s) + Uz(s,i)\,. \tag{1}$$

In equation (1), the GMM supervector $M(s, i)$ is dependent of the speaker s and the session i. $z(s, i)$ is the latent factor which is assumed to belong to a standard normal distribution. U is a low-rank matrix from the constrained session variability subspace of dimension R_C. The computation method of U can be found in [8]. Note that the eigenvectors used to form the U matrix are orthogonal. So the derived projection matrix P can be written as

$$P = UU^t \text{ and } PU = UU^tU = U\,. \tag{2}$$

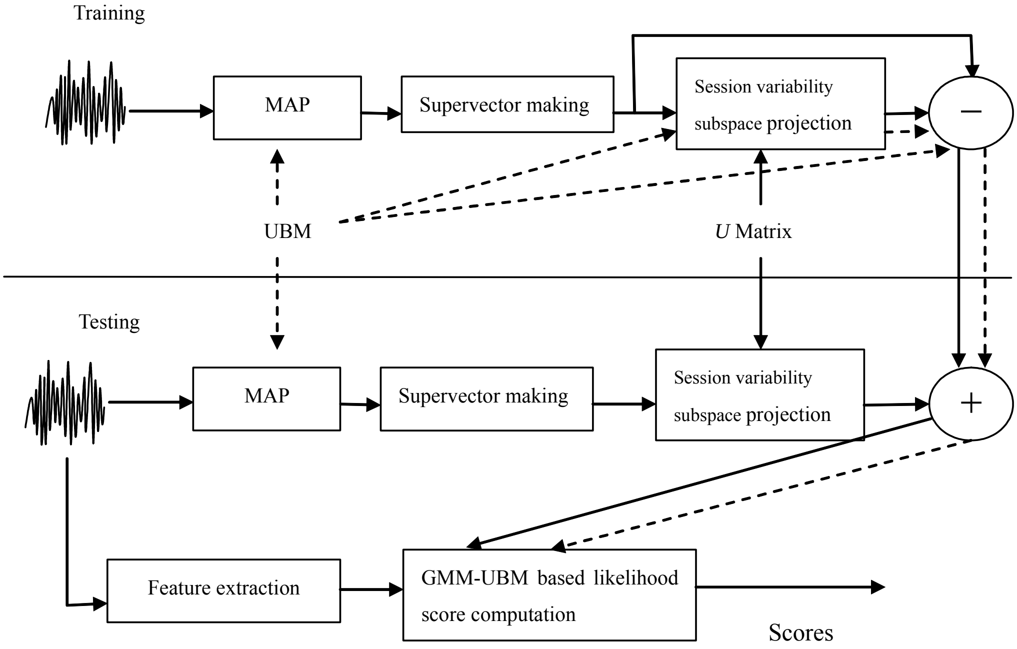
Training
MAP
Supervector making
Session variability subspace projection
−
UBM
U Matrix
Testing
MAP
Supervector making
Session variability subspace projection
+
Feature extraction
GMM-UBM based likelihood score computation
Scores

Fig.1. The schematic diagram of SVSP based model compensation for GMM-UBM systems.

2.2 Speaker Model Training

Given a speaker utterance, the speaker model is trained from UBM by the conventional MAP adaptation [12] only with mean vectors changed. Then a GMM supervector $M(s, i)$ is formed from this speaker model. After that, the session variability is removed from the GMM supervector by projection, which can be written as

$$\begin{aligned} M'(s) &= (I-P)M(s,i) \\ &= (I-P)m(s) \end{aligned} . \tag{3}$$

For a UBM, the derived supervector can be viewed as

$$M(ubm) = m \tag{4}$$

where m is a speaker-independent supervector. Similarly, the session variability is also removed from the UBM supervector, which can be written as

$$M'(ubm) = (I-P)m \tag{5}$$

2.3 Speaker Model Compensation

Given a test utterance j by speaker t, firstly a speaker model is adapted from UBM by the conventional MAP method, then a GMM supervector $M(t, j)$ is formed from it. After that, the session variability in the test utterance is calculated by

$$C(t,j) = PM(t,j) = Pm(t) + Uz(t,j) . \tag{6}$$

Finally, the speaker model $M'(s)$ can be compensated with $C(t, j)$ as

$$M'(s,j) = (I-P)m(s) + Pm(t) + Uz(t,j) . \tag{7}$$

Here, $M'(s,j)$ can be regarded as the model of speaker s in the test utterance's session condition.

Similarly, the UBM $M'(ubm)$ can be compensated with $C(t, j)$ as

$$M(ubm,j) = (I-P)m + Pm(t) + Uz(t,j) \tag{8}$$

where $M(ubm, j)$ can be regarded as the UBM in the test utterance's session condition.

2.4 Test Utterance Verification

As showed in Section 2.3, the compensated speaker model and UBM contain the same session variability as the test utterance. So the top-N log-likelihood ratio scoring [13], which is the basis of most current text-independent speaker verification systems, can be used to verify the identity of the test utterance. In the experiments of this paper, N is set to 4.

Given a sequence of features $\{f_l, l=1,2,\cdots, L\}$ and the derived GMM supervector $M(t, j)$, the l-th frame of feature f_l, which is closest to the k-th mixture component in the GMM supervector, can be expressed as

$$f_l = [m(t)]_k + [Uz(t,j)]_k + \Sigma_k d . \tag{9}$$

where $[.]_k$ means the k-th mixture component in a GMM supervector, Σ_k is the covariance matrix of k-th mixture component, and d is a variant belonging to a standard normal distribution. The score of f_l on the k-th mixture component of the GMM supervector $M(s, i)$ is

$$
\begin{aligned}
H\left(f_l \mid \left[M(s,i)\right]_k\right) &= \frac{1}{(2\pi)^{F/2}\left|\Sigma_k\right|^{1/2}} \\
&\cdot \exp\left\{-\frac{1}{2}\left(f_l-\left[M(s,i)\right]_k\right)^t \Sigma_k^{-1}\left(f_l-\left[M(s,i)\right]_k\right)\right\}
\end{aligned}, \qquad (10)
$$

which mainly depends on

$$
\left(f_l-\left[M(s,i)\right]_k\right)^t \Sigma_k^{-1}\left(f_l-\left[M(s,i)\right]_k\right). \qquad (11)
$$

In equation (10), F is the dimension of the mixture component. In SVSP, the $M(s, i)$ is replaced with the compensated speaker model $M'(s, j)$, so equation (11) can be rewritten as

$$
\begin{aligned}
&\left(\left[(I-P)\left(m(t)-m(s)\right)\right]_k+\Sigma_k d\right)^t \\
&\cdot\Sigma_k^{-1}\left(\left[(I-P)\left(m(t)-m(s)\right)\right]_k+\Sigma_k d\right)
\end{aligned}. \qquad (12)
$$

As showed in equation (12), the session variability between a test utterance and a compensated speaker model is removed. Though there introduces a new item (*I-P*), it appears in the scoring procedure of every compensated speaker model and hence may not decrease the system's performance. Experimental results in Section 3 show that this item does not decrease the performance of a speaker verification system, while compensating the speaker models with session variability in a test utterance really improves the performance of a speaker verification system.

3. EXPERIMENTS AND RESULTS

The experiments were performed on the 2006 NIST speaker recognition (SRE) corpus [14] and focused on the single-side one conversation training, single-side one conversation test.

The features were extracted from speech signal at a frame size of 20 milliseconds every 10 milliseconds. The pre-emphasis factor was set to 0.97. Hamming windowing was applied to each pre-emphasized frame. After that, a 256-point FFT was calculated for each frame and a bank of 30 triangular Mel filters was used. Finally DCT was performed and 16-dimensional MFCC coefficients with the delta coefficients were obtained for each frame. After that, an energy based voice active detection was applied to discard low-energy frames. To mitigate channel effects, mean and variance normalization was applied to the extracted features.

Each gender-dependent UBM consisted of 1,024 mixture components and was trained from NIST SRE04. For the MAP training, only mean vectors were adapted with a relevance factor of 16. The data used for Tnorm were from NIST SRE05, which consisted of 368 females and 245 males. The data used for computing the U matrix were from the single-side 8 conversation training in NIST SRE05, which consisted of 295 females and 202 males.

The baseline system is a conventional GMM-UBM based speaker verification system.

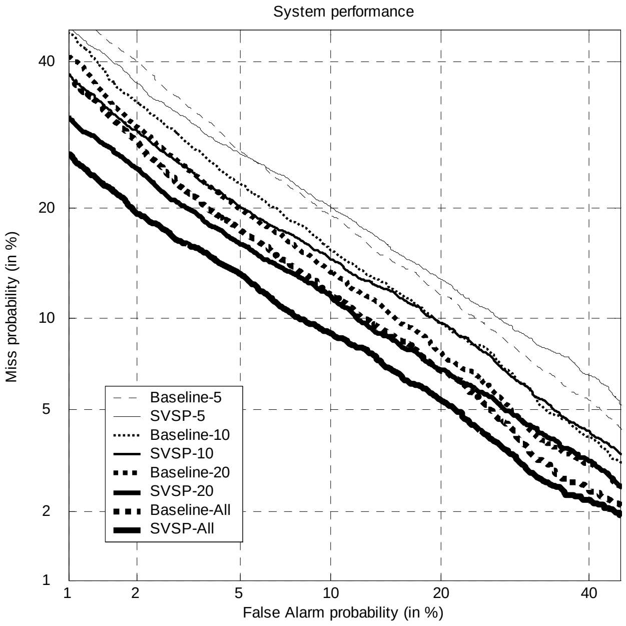

Fig.2. The DET plot for different test utterance lengths comparing baseline and SVSP.

3.1. Session Variability Subspace Size

An important part of SVSP is the size of session variability subspace which will affect the accuracy of the estimation of session variability. The results of different session variability subspace sizes are given in Table 1 where R_C is the size of the session variability subspace. The system used in this experiment was based on SVSP with Tnorm. Experimental results show that the system achieves the best DCF and EER when $R_C = 50$.

Table 1. Minimum DCF and EER results for different session variability subspace sizes.

R_C	DCF(×10^{-2})	EER (%)
10	3.9	10.8
30	3.7	10.1
50	3.6	9.3
100	3.8	11.2

3.2 Test Utterance Length

Another important part of SVSP is the length of test utterance, which will also affect the accuracy of the estimation of session variability. Figure 2 shows the impact of reducing the test utterance length for SVSP and the baseline system with a test utterance length of 5 seconds, 10 seconds, 20 seconds, and all of active speech. Experimental results indicate that at least 10 seconds of speech is required to estimate the session variability, while 20 seconds tests produce a better result with about 9.9% of relative EER reduction compared with the baseline system. With a longer test utterance, for example the full length, about 23.4% of relative reduction for DCF and 16.2% relative reduction for EER can be achieved. Similar results can be found in [7].

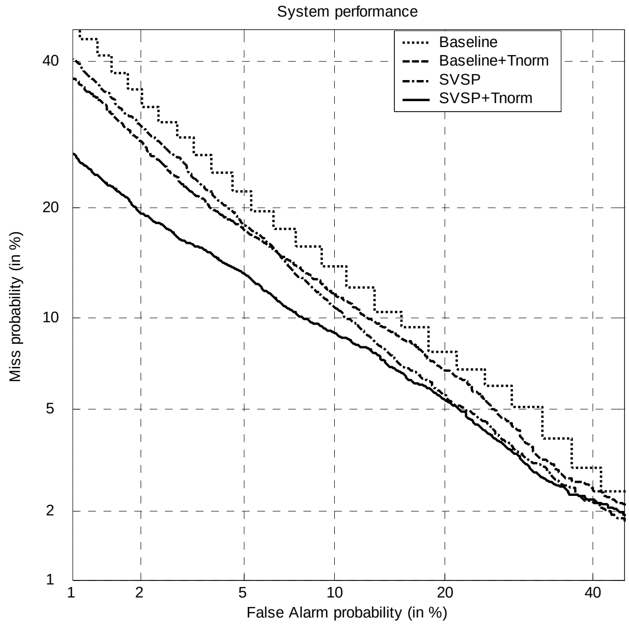

Fig.3. The DET plot for the NIST 2006 single-side one conversation training, single-side one conversation test comparing baseline and SVSP.

3.3 Comparison of different methods

Figure 3 shows the performance comparison of four systems: baseline, baseline with Tnorm, SVSP, and SVSP with Tnorm. Compared with the baseline, SVSP achieves a relative reduction of 15.4% for EER and 9.4% for DCF. With Tnorm both, the relative reduction is 16.2% for EER and 23.4% for DCF.

4. CONCLUSIONS

The results presented in this paper show the effectiveness of SVSP. This method simplifies the computation of session variability by projecting a GMM supervector onto the session variability subspace. During the training phase it removes the session variability from speaker models while during the testing phase it compensates speaker models with session variability estimated from a test utterance. After these processing, the speaker models, the UBM and the test utterance are in the same session condition.

Compared with the results of factor analysis and NAP methods in NIST 2006 SRE, SVSP seems not as effective as these two methods. This may be caused by the item (*I-P*) in equation (12) which may remove some useful speaker-dependent information from a speaker model. Further investigation on this issue will be carried out in the future.

5. REFERENCES

[1] S.Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Processing*, 1981. (29):254-272.

[2] Hermansky, H., Morgan, N, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing,* 1994. (2): 578-589.

[3] J.Pelecanos and S.Sridharan, "Feature warping for robust speaker verification," In *Proc. Speaker Odyssey 2001 conference*, June 2001, pp. 213-218.

[4] D. A. Reynolds, "Channel robust speaker verification via feature mapping," In *ICASSP*, 2003, (2): 53-56.

[5] R.Teunen, B.Shahshahani and L.P.Heck, "A modelbased transformational approach to robust speaker recognition," In *Proc. ICSLP*, 2000, pp. 213-218.

[6] P.Kenny, G. Boulianne, P.Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*,2005, 13(3):345-354.

[7] R.Vogt, S.Sridharan, "Experiments in session variability modeling for speaker verification," *ICASSP*, Toulouse, France, May 2006. (1):897-900.

[8] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A.Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *ICASSP* 2006, (1):97-100.

[9] D. A. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus.," In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, USA, May 1996. (1): 113-116.

[10] R. Auckenthaler,M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification system.," *Digital Signal Processing*, 2000. (10):1-16.

[11] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, New York, USA, April 1988. (1): 595-598.

[12] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10(1): 194-41, 2000.

[13] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, (2):963-966, 1997.

[14] National Institute of Standards and technology, "NIST speech group website," http://www.nist.gov/speech, 2006