Eigenchannel Compensation and Symmetric Score for Robust Text-independent Speaker Verification

Yuan Dong, Jian Zhao, Liang Lu, Jiqing Lui
Beijing University of Posts and Telecommunications
Beijing, China

Xianyu Zhao, Haila Wang
France Telecom Research & Development Center
Beijing, China

Abstract—The negative effect of the session variability has become more and more severe for the performance of the speaker verification system. This paper discusses the eigenchannel compensation and investigates the symmetric scoring method to diminish the session variability and further enhance the performance. Experiments were conducted on the core tests of the 2006 and 2008 speaker recognition evaluation (SRE) corpora of the national institute of standards and technology (NIST) respectively. The experimental results demonstrate that the eigenchannel compensation can achieve excellent improvement and the symmetric scoring, as a measurement of cross similarity, can further improve the performance moderately. Overall, the system performance can be significantly improved, with equal error rate from 9.74% to 5.08% , 47.8% on SRE06 corpus and from 16.26% to 9.42% , 42.1% on SRE08 corpus while detection cost function from 0.0456 to 0.0263 , 42.3% on SRE06 corpus and from 0.0692 to 0.0449 , 35.1% on SRE08 corpus.

Keywords—eigenchannel compensation; symmetric score; session variability; speaker verification; NIST speaker recognition evaluation

I. INTRODUCTION

In the text-independent speaker verification field, the Gaussian mixture model - universal background model (GMM-UBM) [1, 2] strategy based on the statistical theory is the most prevalent and powerful method. The mismatch between mismatch between training and testing conditions is a severe obstacle against the recognition performance. Recently, session variability modeling has become more and more indispensable. The mismatch or the session variability can be caused by a number of phenomena including microphone/handset types, transmission channels, acoustic environments and variability introduced by speakers. Many methods have been proposed and used to diminish the mismatch, such as speaker model synthesis [3], feature mapping [4], score normalization [1] and so on. Speaker model synthesis is to synthesize a new speaker model that matches the channel condition of testing speech. The feature mapping approach focuses on mapping features from different channels into a common channel independent feature space. Score normalization works as a transformation of speaker verification output scores in order to boost the discrimination of the decision threshold by mapping the score distributions of individual speaker models into a common area, and can be

used to reduce many kinds of effects of speaker-dependent and speaker-independent variances.

This paper discusses the eigenchannel compensation [5] technology which is one of the prevalent and effective methods to overcome the negative effect of session variability and proposes a new method of symmetric score in order to further enhance the speaker verification system.

The rest of the paper is organized as follows. The next section describes the principle of the eigenchannel compensation. Then the implementation of symmetric score is followed. System description is presented in section 4. The experimental results on both of the national institute of standards and technology (NIST) speaker recognition evaluation (SRE) 06 and SRE08 data corpora are illustrated in section 5. Finally, the conclusion is given at the end.

II. EIGENCHANNEL COMPENSATION

In eigenchannel filtering for text independent speaker verification, the GMM mean vectors for the incoming utterance can be factored in two parts. One part characterizes the underlying speaker which is assumed to be independent of session variations; the other part is a constrained offset to represent the effect of session variations. This can be represented by,

$$M_r = M_s + Ux_r \quad (1)$$

where, M_r is a supervector which is a concatenation of mean vectors in the GMM characterizing the acoustic observations of a particular recording r , M_s is the supervector identifying the underlying speaker, x_r is a low-dimensional vector of latent session variability factors in the recording and U constrains the session variability whose columns span the session variability subspace in the model space. Session variability subspace U can be trained with any solution of eigen decomposition, e.g. principle component analysis, on a large development set whose utterances represent as many as the factors of session variability. Given a particular recording r , any one of parameter estimation methods applicable to determining eigen-decomposition such as maximum likelihood (ML) estimation, maximum a posteriori (MAP) estimation, and

expectation-maximization (EM) algorithm can be adopted to estimate the latent session variability coefficients x_r .

After obtaining these session variability factors, the constrained offset, Ux_r , can be generated which represents the session variability. The session variability compensation can be carried out either in the model space or in the feature space.

In the model space, the compensated speaker model and UBM are used to calculate the likelihood ratio score, where T_r is the number of frames in r , as

$$LLR_r = \frac{1}{T_r} \log \frac{p(O_r | M_s + Ux_r)}{p(O_r | M_{UBM} + Ux_r)} \quad (2)$$

In the feature space, eigenchannel filtering can be seen as a front end processing. Let o^{SD} be a feature vector affected by session variability, the filtering of o^{SD} to a session independent feature, o^{SI} , is then given by

$$o^{SI} = F^{SD \rightarrow SI}(o^{SD}) \quad (3)$$

$$= \sum_j p(v_j | o^{SD}) \left(\Sigma_j^{SI} (\Sigma_j^{SD})^{-1} (o^{SD} - \mu_j^{SD}) + \mu_j^{SI} \right)$$

where μ_j^{SI} and μ_j^{SD} are the mean vectors of the j -th component in the session independent and session dependent models, Σ_j^{SI} and Σ_j^{SD} are corresponding covariance matrices, $p(v_j | o^{SD})$ is the a posteriori probability of the j -th component given a feature vector. Assuming that the covariance matrices do not change under session variability, the above filtering procedure can be simplified to be,

$$o^{SI} = F^{SD \rightarrow SI}(o^{SD})$$

$$= \sum_j p(v_j | o^{SD}) (o^{SD} - \mu_j^{SD} + \mu_j^{SI}) \quad (4)$$

$$= \sum_j p(v_j | o^{SD}) (o^{SD} - (Ux)_j)$$

where $(Ux)_j$ stands for the j -th subvector in the supervector of Ux corresponding the j -th component.

The experiments in this paper employ the eigenchannel filtering in the model space and significant performance improvement can be achieved.

III. SYMMETRIC SCORE

In our previous published literature, an implementation of speaker adaptive test normalization based on a cross similarity measurement [6] has been proposed. In the recent exploration, it is found out that the symmetric score derived from the same

principle of the cross similarity measurement can make contribution to the recognition performance.

In the traditional scoring procedure of GMM-UBM system, given a test utterance X_i with N_i frames and a target model M_λ which is adapted from UBM with a corresponding enrollment utterance X_λ with N_λ frames, the likelihood score which can be seen as a kind of forward score can be calculated as

$$S_\lambda(X_i) = \frac{1}{N_i} \log \frac{P(X_i | M_\lambda)}{P(X_i | UBM)} \quad (5)$$

As for the symmetric score, besides the forward score above, a reverse score should also be calculated. First, a model M_i characterizing the speaker of the utterance X_i should be generated by model adaptation from UBM according to X_i . Then the reverse score is calculated as

$$S_i(X_\lambda) = \frac{1}{N_\lambda} \log \frac{P(X_\lambda | M_i)}{P(X_\lambda | UBM)} \quad (6)$$

Finally, the symmetric score $S_\lambda^{sym}(X_i)$ between M_λ and X_i can be calculated as follows:

$$S_\lambda^{sym}(X_i) = S_\lambda(X_i) + S_i(X_\lambda) \quad (7)$$

Then, the symmetric score can be used to determine the detection result of this trial by comparing with a predefined threshold.

IV. SYSTEM DESCRIPTION

Experiments were conducted with a GMM-UBM likelihood system on the condition of 1-conversation training, 1-conversation testing of the NIST SRE06 [7] and SRE08 [8] corpora respectively. The experiments focus on the performance comparison of eigenchannel and symmetric scoring. Four different training-testing conditions, including tel-tel, tel-mic, mic-tel, mic-mic conditions, are also analyzed for symmetric scoring. The label tel means the training or test utterances are recorded through a telephone channel while mic means through a microphone channel. Statistics on the number of models and trials on the SRE06 and SRE08 corpora are shown in Table 1 and Table 2.

TABLE I. TARGET MODEL STATISTICS OF SRE06 AND SRE08

	SRE06 models		SRE08 models	
	phn	mic	phn	mic
male	354	379	648	622
female	462	428	1140	538
total	816	807	1788	1160

TABLE II. TEST TRAIL STATISTICS OF SRE06 AND SRE08

condition	SRE06 trails	SRE08 trails
tel-tel	23292	37050
tel-mic	12033	15804
mic-tel	20755	11741
mic-mic	15918	34181
total	71998	98776

For cepstral feature extraction, 13-dimensional PLP vectors are extracted from the silence removed speech signal every 10 ms using 25ms window. Bandlimiting is adopted to retain the frequency range of filterbank outputs from 300Hz to 3400Hz only. The PLP features are processed with RASTA [9] filtering. Then delta, acceleration and triple-delta coefficients are appended to the static coefficients, forming a 52-dimensional feature vector. Finally, the 52-dimensional feature vector continues to be transferred by heteroscedastic linear discriminant analysis (HLDA) [10, 11] to 51 dimensions and histogram equalization [12, 13].

In the GMM-UBM system in this paper based on the likelihood ratio detection, a 2048 mixture gender-dependent UBM is trained from a large development set of telephone data selected from Switchboard I & II Cellular part and NIST SRE04 evaluation corpus, including 350 males and 468 females, approximately 44 hours in total. Speaker models of targets and test utterances for symmetric scoring are derived by Bayesian adaptation [1] from the UBM using the designated utterances, only the mean vectors adapted.

For the eigenchannel compensation, two different eigenbasis are used according to the type of test utterances. The eigenbasis for the telephone test condition are trained using 1790 telephone utterances selected from SRE04 corpus. 2502 more microphone utterances selected from SRE05 corpus are added, 4292 in total, to train the second eigenbasis for the microphone test condition in order to improve the performance of microphone test condition in the case that the available microphone data for system development are quite limited.

V. EXPERIMENTAL RESULTS

In this section, the experimental results of the speaker verification systems are reported and analyzed. Results are mainly presented using Equal Error Rate (EER) and Detection Error Tradeoff (DET) curves. Along with DET curves, the minimum Decision Cost Function (DCF), defined by NIST as $DCF = 0.1 * Pr(miss) + 0.99 * Pr(false_alarm)$ [7], is also employed as an overall performance criterion.

Fig. 1, Fig. 2 and Table 3 show the performance criteria and the DET curves on all of the four conditions as a whole for the SRE06 and SRE08 corpuses. It is found that eigenchannel compensation can significantly improve the performance and that symmetric scoring method can further contribute to the system moderately, with EER from 9.74% to 5.08%, 47.8% improvement on SRE06 corpus and from 16.26% to 9.42%, 42.1% on SRE08 corpus while DCF from 0.045 to 0.0263, 42.3% on SRE06 corpus and from 0.0692 to 0.0449, 35.1% on SRE08 corpus.

Figure 1. Performance on the SRE06 corpus

Figure 2. Performance on the SRE08 corpus

As for the symmetric scoring method, when considering the four training and testing conditions separately, it can be found that symmetric scoring can moderately improve the performance on the channel match condition (tel-tel and mic-mic). Moreover, it helps most in the case of mic-tel. This may be due to that the “reverse” models trained on the testing telephone recordings provide the better discriminative capability than “forward” models trained on the microphone recordings. Similarly, it should be noted that for the case of tel-mic, the “reverse” models trained on the testing microphone recordings harm performance to some extent because of the lack of the representation ability of UBM to characterize the microphone data.

TABLE III. PERFORMANCE CRITERIA ON THE SRE06 AND SRE08 CORPUSES

system	SRE06 corpus		SRE08 corpus	
	EER(%)	DCF(x10)	EER(%)	DCF(x10)
baseline	9.74	0.456	16.26	0.692
baseline+	5.45	0.274	9.62	0.455
eigenchn				
baseline+	5.08	0.263	9.42	0.449
eigenchn+				
sym score				

TABLE IV. CONDITIONAL PERFORMANCE COMPARISON OF SYMMETRIC SCORING TECHNIQUE ON SRE06 CORPUS

condition	forward score		symmetric score	
	EER(%)	DCF(x10)	EER(%)	DCF(x10)
tel-tel	5.12	0.2493	4.85	0.2435
tel-mic	4.96	0.2383	4.81	0.2450
mic-tel	4.48	0.1769	3.20	0.1368
mic-mic	4.38	0.2987	4.13	0.2719

TABLE V. CONDITIONAL PERFORMANCE COMPARISON OF SYMMETRIC SCORING TECHNIQUE ON SRE08 CORPUS

condition	forward score		symmetric score	
	EER(%)	DCF(x10)	EER(%)	DCF(x10)
tel-tel	7.02	0.3775	6.71	0.3719
tel-mic	9.10	0.4621	9.37	0.5413
mic-tel	12.13	0.6058	8.78	0.4703
mic-mic	7.18	0.3929	6.93	0.3706

VI. CONCLUSION

This paper discusses the eigenchannel compensation method and investigates the symmetric scoring to cope with the session variability problem. Experiments were conducted on both the NIST SRE06 and SRE08 corpuses. The experimental results demonstrate that the eigenchannel compensation can bring the tremendous performance improvement and that

symmetric scoring can play an auxiliary role for the speaker verification system.

REFERENCES

- [1] F. Bimbot, J. Bonastre, and C. Fredouille, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, pp. 430–451, 2004.
- [2] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.
- [3] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformation approach to robust speaker recognition," *Proc. Int. Conf. Spoken Language Processing*, 2000.
- [4] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2003.
- [5] C. Vair, D. Colibro, F. Castaldo, "Channel factors compensation in model and feature domain for speaker recognition," *Proc. Speaker Odyssey*, 2006.
- [6] J. Zhao, Y. Dong, X.Y. Zhao, "Discussion on score normalization and language robustness in text-independent multi-Language speaker verification," *Lecture Notes in Computer Sciences 4681*, Springer, Germany, 2007, pp. 1121–1130.
- [7] The NIST year 2006 speaker recognition evaluation plan [2008-03-12]. <http://www.nist.gov/speech/tests/spk/-2006/sre-6evalplan-v9.pdf>.
- [8] The NIST year 2008 speaker recognition evaluation plan. [2008-05-23]. http://www.nist.gov/speech/tests/spk/-2008/sre08_evalplan_release4.pdf.
- [9] H. Hermansky, N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.
- [10] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," *Doctoral Dissertation*, John Hopkins University, Baltimore USA, 1997.
- [11] H. Yang, Y. Dong, X.Y. Zhao, "Discriminative transformation for sufficient adaptation in text-independent speaker verification," *Lecture Notes in Computer Science 4272*, Springer, Germany, 2006, pp. 558–565.
- [12] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, Prentice-Hall, New Jersey, 2002.
- [13] Á. Torre, A.M. Peinado, J.C. Segura, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 355–366, 2005.