IOA 2010 Speaker Recognition Evaluation System Description

J. Zhang, L. Yang, X. Zhang, Y. Zhou, H. Suo and Y. Yan
Speaker and Language Recognition Group
ThinkIT Speech Lab, Institute of Acoustics, CAS, P. R. China
{jzhang, lyang, xzhang, yzhou, hsuo, yyan}@hccl.ioa.ac.cn

1. Introduction

IOA has submitted one system (i.e., primary system) to NIST SRE 2010 evaluation, only to the core-core condition. The system is a fusion of 8 sub-systems which are different in features, models and prosodic information usages.

2. Submitted System

The IOA primary system is a fusion of these gender dependent sub-systems:

- Factor analysis GMM UBM system with MFCC 12x3 features extracted from original waveforms. Z-norm followed by T-normalization was also performed on the scores.
- Factor analysis SVM GSV system with MFCC 12x3 features extracted from original waveforms. Z-norm followed by T-normalization was also performed on the scores.
- Factor analysis GMM UBM system with LPCC 18x2 features extracted from original waveforms. Z-norm followed by T-normalization was also performed on the scores.
- Factor analysis SVM GSV system with MFCC 18x2 features extracted from original waveforms. Z-norm followed by T-normalization was also performed on the scores.

Corresponding to the above 4 cases, sub-systems using features extracted from re-synthesized waveforms are also built. Each sub-system was calibrated first using the channel side information. Such calibrated sub-systems were fused by LLR.

3. Data Preparation

Speech and silence segmentation is performed by ASR transcriptions, which are provided by NIST. All words are linked to speech class. Segments labeled speech or silence are generated. Also our in-house phone decoder was used to remove additional silence on pre-processed waveforms. For each file which may contain no speech, a very low score is assigned to its corresponding trials and a decision FALSE is given

to these trials.

4. Acoustic Feature Extraction

For MFCC features, 12 mfcc coefficients (not including energy) are computed and cepstral mean subtraction was performed. First and second order derivatives over 5 frames are appended to each feature vector, which results in the dimensionality of 36.

For LPCC features, the speech is segmented into frames by a 30-ms Hamming window progressing at a 10-ms frame rate. Each speech frame is parameterized by the 18th order LPCCs and their first derivative (i.e., a 36-dimensional feature vector). Further processing including CMS and Gaussianization is applied to all the LPCCs.

Gender-dependent Gaussian mixture model (GMM) with 1024 Gaussian components is trained on these two sets of features.

5. Factor Analysis

In our system, the joint factor analysis model is used as the key channel compensation method. The model combines the priors underlying classical MAP, eigenvoice MAP and eigenchannel MAP. In our implementation, the channel factor loading matrix (U), the speaker factor loading matrix (V) and the residual factor loading matrix (d) are estimated in a cascading way. We use the speech data of Switchboard II and Switchboard Cellular to train the speaker factor loading matrix V with 300 speaker factors. For channel space UU*, telephone speech data of NIST SRE 2004-2006 was used to train the telephone channel space, microphone speech data of NIST SRE 2005-2006 was applied to train a microphone channel space and the MIXER5 interview speech data was used to train interview channel space. In the enrollment stage, the three factors, i.e., x, y and z are re-estimated.

6. GMM and SVM GSV

6.1 GMM modeling

We used the NIST SRE-2004 1-side data to train a gender-dependent UBM with 1024 mixtures. GMM models adapted from UBM by MAP-adaptation are used to model the target speakers (only means were adapted). Relevance factor of 12 is used for the MAP adaptation. 1024 UBM model is estimated by EM algorithm and MAP-adaptation is used to obtain the target model supervector. For GMM modeling, in general, the score before zt-norm for each trial is given by log likelihood ratio, log p(x | M) - log p(x | UBM), where x is the test segment and M is the target speaker model.

6.2 SVM GSV

The feature extraction and UBM training are done in the same way as above. Means of Gaussian components are adapted by MAP adaptation for each training, testing and background segment. Then the corresponding GMM supervectors are obtained. Note, the supervector is normalized by the corresponding standard deviation and weight. The linear kernel is used to classify GMM supervectors. The target segment is split to multi parts with 20 seconds interval. The SVM model of target is trained on these positive samples and the negative back ground samples. The background samples are chosen from NIST 04, 05 and 06 evaluation data. The SVM training and scoring are built with SVM-Light release tools.

7. System based on re-synthesized waveforms

Our prosodic-related system is implemented with following considerations: 1) we suppose that noise, silence and other non-speech signals do not contain much speaker specific information. 2) prosodic information is the main source for discriminating one speaker from others. So in our implementation, original speech provided by NIST SRE was pre-processed by our in-house phone decoder to get contentful parts. Then in these contentful parts, we extracted their prosodic features, such as pitch, vibration amplitude and phase bias. Next, only the part having periodic components is kept and its corresponding prosodic features are used to re-synthesize the harmonic structure of the speech signals using sinusoidal modeling technique. So the re-synthesized waveform contains only the contentful and prosodic-related information for the specific speaker. In such a way, some garbage speech is discarded and the recognition system may work in a more robust way.

8. Calibration and Fusion

We have used calibration and fusion tool FOCAL based on logistic linear regression developed by Niko Brummer. We use the side-information conditional calibration and fusion. We calibrated the subsystems with side information about channel provided by NIST which categorized each trial into one of five classes: phonecall-mic to phonecall-mic, phonecall-tel to phonecall-tel, interview-mic to interview-mic, interview-mic to phonecall-tel and interview-mic to phonecall-mic.

The confidence scores in our submission may not be interpreted as log likelihood ratios.

9. Speed and Resources

All our experiments are carried out on several DELL PowerEdge R610 systems with Intel Xeon CPU E5520 @2.27GHz, 8GB memory. The verification process for the JFA based GMM system takes about 0.012xRT. Also, the verification process for JFA based GSV system takes about 0.018xRT. The overall fused primary system runs about 0.12 times real time.

Table 1: Per system requirements for core-core task
(Memory requirement in MB and Time requirement in Real time factor)

	Enrollment	Verification	Total
Time	0.25	0.015	0.265
Memory	200	300	500

10. Acknowledgments

Thanks to Yeming Xiao and Chunyan Liang for carrying out a lot experiments for us. Thanks to Xiang Xiao for designing various development data sets for testing purpose. Thanks also go to all members of Speaker and Language Recognition Group and ThinkIT speech Lab for supporting us during these periods.
This work is partially supported by The National Science and Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

References

[1] Kenny, P "Joint factor analysis of speaker and session variability: Theory and algorithms" - Technical report CRIM-06/08-13 Montreal, CRIM, 2005, http://www.crim.ca/perso/patrick.kenny/.

[2] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P.: “A Study of Inter-Speaker Variability in Speaker Verification”, IEEE Transactions on Audio, Speech and Language Processing, July 2008.

[3] Burget Lukáš, Fapšo Michal, Hubeika Valiantsina, Glembek Ondřej, Karafiát Martin, Kockmann Marcel, Matějka Pavel, Schwarz Petr, Černocký Jan: BUT system description: NIST SRE 2008, In: Proc. 2008 NIST Speaker Recognition Evaluation Workshop, Montreal, CA, NIST, 2008.

[4] Claudio Vair, Daniele Colibro, Fabio Castaldo, Emanuele Dalmasso, Pietro Laface: Channel Factors Compensation in Model and Feature Domain for Speaker Recognition , IEEE 2006.

[5] Matejka P., Burget L., Schwarz P. and Cernocky J., Brno University of Technology System for NIST 2005 Language Recognition Evaluation. Odyssey: The Speaker and Language Recognition Workshop, San Juan, Puerto Rico,Jun 2006

[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] Niko Brummer and Johan du Preez: Application-Independent Evaluation of Speaker Detection, Computer Speech and Language, 2005.

[8] WM Campbell, Generalized linear discriminant sequence kernels for speaker recognition. Acoustics, Speech, and Signal Processing, 2002. Proceedings.2002

[9] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech and Language Processing, 2007.

[10] Glembek, O., Burget, L., Dehak, N., Brummer, N., and Kenny, P., "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis", Proc ICASSP 2009, Taipei, Taiwan, April 2009.

[11] Serra, X., "Musical sound modeling with sinusoids plus noise", Musical signal processing, pp. 497–510, 1997.