The THU-EE Speaker Recognition Systems for NIST SRE 2010

Jia Liu, Liang He, Tao Hou, Yan Deng, Zhiyi Li
Yongzhe Shi, Meng Cai, Jiaming Xu, Xin Zhang, Wei-Qiang Zhang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liuj@tsinghua.edu.cn

Abstract

This paper describes the speaker recognition systems from the Department of Electronic Engineering, Tsinghua University (THU-EE) for NIST SRE 2010. Three systems are submitted for the evaluation, all of which are based on fusion of multiple subsystems. We describe each subsystem briefly and give their configurations. The processing speed of the primary system is also given in the paper.

Index Terms: THU-EE, NIST SRE 2010.

1. Introduction

This paper describes the speaker recognition systems from the Department of Electronic Engineering, Tsinghua University (THU-EE) for NIST 2010 Speaker Recognition Evaluation (SRE).

Our submissions are built on the following feature vectors:

- MFCC 13+Δ+ΔΔ: 13-dimensional MFCC concatenated with delta and double delta cepstrum, forming a 39-dimensional feature vector.
- PLP 13+Δ+ΔΔ+ΔΔΔ: 13-dimensional PLP concatenated with delta, double delta and triple delta cepstrum, forming a 52-dimensional feature vector.
- High level: 6-dimensional pitch contour, 5-dimensional energy and 1-dimensional duration polynomial coefficients are concatenated to form a 12-dimensional feature vector.
- MFCC HLDA 51: MFCC 13+Δ+ΔΔ+ΔΔΔ with HLDA dimension reduction, forming a 51-dimensional feature vector.
- PLP HLDA 51: PLP 13+Δ+ΔΔ+ΔΔΔ with HLDA dimension reduction, forming a 51-dimensional feature vector.
- TFC 15+13+11 fLFA: Time-frequency ceptral (TFC) [1] feature is extended for speaker recognition. 9 successive frames of basic feature vectors are first extracted to form a cepstrum matrix. A temporal (in horizontal direction) DCT is then performed on the cepstrum matrix and the first three columns with 15, 13, and 11 elements are concatenated to form a 39-dimensional feature vector. At last, feature domain latent factor analysis (fLFA) [2] is applied to reduce the effect of channel distortion.
- TFC 15+13+11 fSFA: TFC 15+13+11 with feature domain simplified joint factor analysis (fSFA) [3].

Our submissions are built on the following classifiers:

- GMM-UBM: Classical Gaussian mixture model - universal background model (GMM-UBM) [4].
- GSV-SVM: GMM super-vectors for support vector machines (GSV-SVM) [5].
- JFA: Joint factor analysis [6, 7].
- MLLR-SVM: Maximum likelihood linear regression transforms as features for SVM (MLLR-SVM) [8].
- PGMM: Phonetic GMM [9].

2. Detailed system descriptions

2.1. JFA-MFCC39

This subsystem uses factor analysis based 39-dimension MFCC. The UBM is gender-dependent with 1024 mixtures. The UBM training data come from SRE04, SRE05, SRE06 and Switch Board I, II and Cellular. The same data are used to train eigenvoice. SRE05, SRE06, Mixer 5 and SRE08 follow-up data are used to train eigenchannel. The eigenvoice factor is 300, eigenchanel factor is 100.

2.2. JFA-PLP52

This subsystem is similar to JFA-MFCC39 except for the feature and data used to train eigenvoice and eigenchannl. 51-dimensional PLP feature vector is used.

2.3. SVM-PLP51

This subsystem is built on GSV-SVM classifier using 51-dimensional PLP feature vector with HLDA. The UBM is gender-dependant and the number of mixture is 512 and the SVM used linear kernel function. The UBM training data come from SRE04 1-side training set. The HLDA training data come from SRE04 8-side training set. The SVM negative pool come from SRE04 1-side training set. The Nuisance attribute projection (NAP) [10] training data come from SRE04, SRE05 and SRE06. The number of NAP channel is 128.

2.4. SVM-MFCC51

This subsystem is similar to SVM-PLP51, except that 51-dimensional MFCC feature vector is used.

2.5. GMM-TFC39

This subsystem is built on GMM-UBM classifier using 39-dimensional TFC feature vector with fLFA. The UBM is gender-dependant and the number of mixture is 2048. The UBM training data come from SRE04 1-side training set. The LFA training data come from SRE04 8-side training set and SRE08 deferment set. The UBM for LFA is 512-mixture and the number of channel is 30.

2.6. MLLR-TFC39

This subsystem is built on MLLR-SVM classifier using 39-dimensional TFC feature vector with fSFA. The UBM training data come from SRE04 1-side training set. The SFA training data come from SRE04 8-side training set and SRE08 deferment set. The UBM for MLLR is 1024-mixture and the MLLR supervector dimension is $7 \times 39 \times 40$.

2.7. PGMM-MFCC39

This subsystem is built on PGMM classifier using 39-dimensional MFCC feature vector. During training for both UBM and target speaker, each utterance is divided into several segments according to the recognition result obtained by the same Hungarian phone recognizer [11]. Then these segments are clustered to form seven broad phone classes: vowels, diphthongs, plosives, affricates, fricatives, liquids and nasals. The maximum number of Gaussians per class is 1024 and the total mixture number is 3072.

2.8. GMM-HL12

This subsystem is built on GMM-UBM classifier using 12-dimensional high level feature vector. The UBM is gender-dependant and the number of mixture is 256.

3. Fusion and calibration

A set of 1000 Z-norm and 500 T-norm speakers from SRE05, SRE06 and Mixer5 is used to ZT-norm all the individual systems [12]. All the scores are then scaled and shifted using a linear score to likelihood ratio mapping. For this, we use Niko Brummer's FoCal package [13]. Training the linear parameters of this affine transformation is based on SRE08 data. The threshold for decision is trained on SRE08 test data.

4. Submission systems

The primary system consists of all the subsystems, but we alleviate the fusion weight of JFA-PLP52 subsystem, which may be overfitted on the development set. The contrast1 system consists of all the subsystems except for JFA-PLP52. The contrast2 system consists of all the subsystems optimized on the development set.

5. Processing time

Performances of all subsystems were measured separately on only one core of an Intel Core 2 Quad CPU 2.4GHz and 2 GB memory. Results are shown in Table 1. Note that the real time (RT) factor of the primary fusion system is less than the sum of all subsystems. It is because several subsystems share some common processing stages.

Table 1: *Per subsystem processing speed in real time factors.*

Subsystem	RT
JFA-MFCC39	0.005
JFA-PLP52	0.008
SVM-PLP51	0.004
SVM-MFCC51	0.004
GMM-TFC39	0.01
PGMM-MFCC39	0.08
MLLR-TFC39	0.05
GMM-HL12	0.001

6. References

[1] W.-Q. Zhang, L. He, Y. Deng, J. Liu, and M. T. Johnson, "Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, to be published.

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

[3] W. Guo, Y. Li, and L. Dai, "Simplified factor analysis in speaker verification," in *Proc. ICALIP*, Shanghai, Jul. 2008.

[4] 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.

[5] 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, France, May 2006, pp. I 97–I 100.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[8] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 2425–2428.

[9] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair, "Loquendo - Politecnico di Torino's 2008 NIST speaker recognition evaluation system," in *Proc. ICASSP*, Taipei, 2009, pp. 4213–4216.

[10] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *Proc. ICASSP*, 2005, pp. I 629–I 632.

[11] J. Gernocky, P. Matejka, P. Schwarz, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. Eurospeech*, 2005, pp. 2237–2240.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1-3, pp. 42–54, Jan. 2000.

[13] N. Brummer and J. de Preez, "Application-independent evaluation of speaker detection," *IEEE Transactions on Speech and Audio Processing*, vol. 20, pp. 230–275, 2006.