The THUEE Speaker Identity Verification System for Evalita 2009 SIV-Application

Jia Liu, Liang He, Tao Hou, Zhiyi Li, Yongzhe Shi, and Wei-Qiang Zhang

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

Abstract. This paper focuss speaker identity verification system from the Department of Electronic Engineering, Tsinghua University (THUEE) for Evalita 2009 Speaker Identity Verification Application Track (SIV-A). Two systems are submitted for the evaluation. The primary one consists of GMM-UBM and GMM-SVM. The contrast one is based on high level feature. We describe each subsystem briefly and give their configurations. The processing speed of the primary system is also given in the paper.

Keywords: THUEE, EVALITA 2009, speaker verification, fusion, high level feature.

1 Introduction

This paper describes speaker identity verification system from the Department of Electronic Engineering, Tsinghua University (THUEE) for Evalita 2009 Speaker Identity Verification Application Track (SIV-A) [1]. We submit two systems for SIV-A. Our primary system is the fusion of three subsystems:

1. GMM-UBM subsystem [2]: simplified factor analysis (SFA) is deployed in feature domain [3].
2. GMM-SVM subsystem [4]: nuisance attribute projection (NAP) is used to degrade the impact of inter-session variability [5].
3. high level feature subsystem [6] : the modelling of this subsystem is the same to GMM-UBM, but we use high level feature instead of acoustic feature.

Our contrast system only consists of high level feature subsystem.

2 System Description

Both GMM-UBM subsystem and GMM-SVM subsystem are based on acoustic features. But these two subsystems use different modelling. The former takes GMM-UBM as baseline system and uses simplified factor analysis matrix to compensate acoustic feature and the latter takes GMM-SVM as baseline system and adopts nuisance attribute projection in model domain. High level feature subsystem uses the same modelling with GMM-UBM subsystem but is based on high level features.

2.1 GMM-UBM subsystem

Feature Extraction. G.723.1 VAD detector is applied to perform speech/silence segmentation [7]. We use 20ms window and 10ms shift to extract 12 MFCC coefficients plus C0. Cepstral mean subtraction and feature warping with a 3s windows are applied to improve system robustness [8]. Delta, acceleration are appended to each feature vector. 15% of low energy frames are discarded using a dynamic threshold.

Modelling. This subsystem is built gender-dependently. Feature domain simplified factor analysis (fSFA) is adopted as the main channel compensation technology.

We split Evalita 2009 SIV-A UBM Data into three sets: ubm set, lambda set and score normalization set. UBM with 256 Gaussian mixtures is trained on ubm set using 39-order MFCCs for factor analysis matrix and UBM with 2048 Gaussian mixtures is trained on the same set using compensated 39-order MFCCs for verification. Lambda set is used for factor analysis matrix and score normalization set is reserved for zt-norm. Since we use the same data to perform z-norm and t-norm, target trials exist, which is contradict to the requirement of zt-norm. So we remove target trial scores when calculating normalization parameters.

2.2 GMM-SVM subsystem

Feature Extraction. Feature extraction process is similar to GMM-UBM subsystem. The difference is that delta, acceleration and triple-delta coefficients are appended to each feature vector and Heteroscedastic linear discriminant analysis(HLDA) is employed to decorrelate features and reduce the acoustic feature dimensionality from 52 to 51.

Modelling. Gender-independent UBM with 1024 Gaussian mixtures is trained on ubm set. NAP matrix with 50 coranks is trained on lambda set.

Compared with classic GMM-SVM, our system makes a little modification. In classic GMM-SVM, adapted mean supervectors (and diagonal variance supervectors) are used as inputs of SVM classifier. The adapted supervector is combination of two parts: ubm mean (and variance) supervector and MAP adaptation supervector. In our system, we only use MAP adaptation supervector as inputs of SVM classifier. In both training and testing, we use the popular SVM Torch [9, 10].

2.3 High Lever Feature subsystem

Feature Extraction. Components of high level feature are diverse. Pitch, loudness and duration form 11-order high level feature.

2.4 Fusion

The scores of our three subsystems are scaled and shifted using likelihood ratio mapping. We use Niko Brummer's FoCal package to train linear parameters of affine transformation [11]. The Threshold for decision is trained on Evalita 2009 SIV-A Development Data.

3 Processing Time

Performance of all subsystems are measured separately on only one core of an Intel Core Quad CPU 2.4Ghz and 2GB memory. The real time (RT) factor of the primary fusion system is 1.22.

References

1. Aversano, G.: EVALITA 2009 - Speaker Identity Verification - Application Track Task Guidelines, http://evalita.fbk.eu/speaker.html
2. Reynolds, D., Quatieri, T., Dunn, R.: Speaker Verification using adapted Gaussian mixture models. Digital Signal Processing, vol. 10, issue 1, pp. 19–41 (2000)
3. Kenny, P.: Eigenvoice Modeling with sparse training data. IEEE Transactions on speech and audio processing, vol. 13, issue 3, pp. 345–354 (2005)
4. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support Vector Machines Using GMM Supervectors for Speaker Verification. IEEE Signal Processing Letters, vol. 13, pp. 308–311 (2006)
5. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in channel compensation for SVM speaker recognition. In: Proceedings of ICASSP (2005)
6. Campbell, J.P., Reynolds, D.A., Dunn, R.B.: Fusing high- and low-level features for speaker recognition. In: Proceedings of Eurospeech, pp. 2665–2668 (2003)
7. ITU. G.723.1 Annex A. Speech coders: Silence compression scheme. Geneva: ITU-T (1996)
8. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proceedings of A Speaker Odyssey, pp. 213–218. Crete, Grece (2001)
9. Vapnik, V.: The Nature of Statistical Learning Theory. SpringerVerlag, New York (1995)
10. Collobert, R., Bengio, S.: SVMTorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research, vol. 1, pp.143–160 (2001)
11. Brummer, N., de Preez, J.: Application-independent evaluation of speaker detection. Computer Speech and Language, vol. 20, pp. 230–275 (2006)