IIR System Description for the 2010 NIST Speaker Recognition Evaluation Submission

Bin Ma, Hanwu Sun, Kong Aik Lee, Changhuai You, Donglai Zhu, Eryu Wang, Rong Tong, Chien-Lin Huang, Cheung Chi Leung, Ville Hautamaki, and Haizhou Li

Institute for Infocomm Research (IIR),
Agency for Science, Technology, and Research (A*STAR), Singapore
{mabin,hwsun,kalee,echyou,dzhu,tongrong,
clhuang,cclueng,vishv,hli}@i2r.a-star.edu.sg

1. INTRODUCTION

The Institute for Infocomm Research (IIR) team submitted one primary system for the 2010 NIST Speaker Recognition Evaluation (SRE) based on the fusion of multiple classifiers and acoustic features as described in this document (IIR_SystemDescription.pdf). This submission includes the results for seven train-test conditions as indicated in Table I. In particular, included in this submission (IIR.zip) are:

1. IIR_1_10sec_10sec_primary_llr.txt
2. IIR_1_core_10sec_primary_llr.txt
3. IIR_1_8conv_10sec_primary_llr.txt
4. IIR_1_core_ core_primary_llr.txt
5. IIR_1_8conv_core_primary_llr.txt
6. IIR_1_core_summed_primary_llr.txt
7. IIR_1_8conv_summed_primary_llr.txt
8. IIR_SystemDescription.pdf

The confidence scores of our submission can be interpreted as log-likelihood scores.

2. SYSTEM DESCRIPTION

The IIR system consists of four different classifiers using different acoustic features as listed in Table II. The first classifier is based on the generative GMM-UBM approach [1], while the remaining three classifiers are based on discriminative SVM techniques. In the following, the feature extraction process and classifiers designed are briefly described. The fusion technique is also reported.

2.1 Feature Extraction

2.1.1 PLP

The HTK toolkit is used to extract the PLP features. Speech samples were segmented into frames with a 20ms Hamming window progressing at a 10ms frame rate. Each speech frame was parameterized with 13 PLP coefficients and their first and second derivatives (i.e., a 39-dimensional feature vector). Further processing includes RASTA filtering, VAD detection [2], CMS and Gaussianization were applied.

Table I IIR submission includes seven training-test conditions.

		Test segment condition		
		10sec	core	summed
Training condition	10sec	✓		
	core	✓	✓	✓
	8conv	✓	✓	✓
	8summed			

Table II Classifiers (both generative⁺ and discriminative*) and acoustic features used for the *IIR* speaker recognition system.

Classifier	Features
GMM-UBM-JFA⁺ GMM-SVM-KL* GMM-SVM-BHATT* GMM-SVM-FT*	LPCC PLP MFCC

2.1.2 LPCC

The SPTK toolkit was used to extract the LPCC features. Speech samples were segmented into frames with a 30ms Hamming window progressing at a 10ms frame rate. Each speech frame was parameterized with 18 order LPCC coefficients. Using on the first derivative, a 36-dimensional feature vector was obtained. Further processing includes RASTA filtering, VAD detection [2], CMS and Gaussianization were applied.

2.1.3 MFCC

The Abacus toolkit was used to extract the MFCC features. 16 order MFCC was generated with a 30ms window at a 12.5ms frame rate. The 16-order MFCC, 16-order first and 14-order second derivatives were appended to form the final 46-dimensional feature vector. Spectral subtraction based noise reduction method [3] was used to assist the energy-based VAD to remove silence frames and to retain only the high quality speech frames for all the telephone and telephone-microphone data. For the interview-microphone style channel data, the frame selection is based on the logical AND between the energy based VAD and ASR transcripts provided by NIST. Finally, the selected feature vectors were processed by RASTA filtering and mean-variance-normalization (MVN).

2.2 Classifiers

2.2.1 GMM-UBM-JFA

Joint factor analysis (JFA) [4, 5] is a modeling technique used to treat the problem of speaker and channel variability, built on top of the classical GMM-UBM approach [1]. NIST SRE04 1side training data was used to generate a gender-dependent UBM model with 1024 Gaussian mixtures. Switchboard II data, SwitchBoard Cellular and SRE04 corpus were used to train a speaker space with 300 speaker factors. The diagonal matrix was trained using SRE04 1side train and test data. For channel space training, a telephone channel space with 100 channel factors was trained based on the telephone data from SRE04, SRE05, SRE06 and SRE08 data. Microphone channel space (50 channel factors) was trained based on the microphone data from SRE05, SRE06 and SRE08. Finally, interview data from the MIXER5, SRE08 and SRE08-followup were used to train an interview channel space with 100 channel factors. The full channel space (250 channel factors in total) was formed by appending the above three sub-spaces.

TZNORM was applied based on the train (TNORM) and test (ZNORM) conditions to different systems [6]. For ZNORM, we use SRE05 1side training utterances for telephone data and SRE05 microphone utterances for microphone and interview data. For TNORM, we use SRE06 1side training models for telephone data and SRE06 microphone models for interview data.

2.2.2 GMM-SVM-KL

The GMM-SVM system was designed based on the work reported in [7, 8]. Given an utterance for GMM adaptation, only mean vectors are adapted via MAP, while its weights and covariance matrices are kept unchanged. The mean vectors of mixture components in the GMM are then concatenated to form a supervector, which is used as input to SVM. The mean vectors are normalized by its standard deviation and weighted by the squared root of the weights of the Gaussian mixtures. This normalization step was motivated from the Kullback-Liebler (KL) divergence perspective, and the resulting kernel was referred to as the KL kernel. LibSVM [9] was used to train the SVM models and the NAP [7] with a corank of 60 was used for channel compensation in the supervector space.

The system was designed to be gender-dependent. The UBMs, with a model size of 1024, were trained using SRE04 data. The NAP loading matrix was trained using SRE04 telephone, SRE05 microphone, SRE06 microphone, MIXER5 and SRE08-followup data. SRE04 and MIXER 5 data were used to form the SVM background.

2.2.2 GMM-SVM-BHATT

This subsystem follows the conventional GMM-SVM architecture, as mentioned above, except that a different kernel metric was used, which we refer to as the Bhattacharyya kernel [10]. Different also from the KL kernel, the Bhattacharyya kernel allows the adaptation of both mean vectors and covariance matrices. Compared to the KL kernel, the Bhattacharyya kernel can better reflects the salient characteristics of the speaker GMM, where the supervector represents the relative distance to the UBM instead of an absolute point in the vector space. Similar training and development data as mentioned in Section 2.2.2 were used.

2.2.4 GMM-SVM-FT

The FT-SVM method characterizes a speaker by the difference between the speaker and a cohort of background speakers in the form of feature transform (FT) [11]. The FT is a linear regression function that projects speaker dependent features to speaker independent ones, also known as an affine transform. It consists of two sets of parameters, bias vectors and transform matrices. The former, representing the first order information, is more robust than the latter, the second order information. We propose a flexible tying scheme that allows the bias vectors and the matrices to be associated with different regression classes, such that both parameters are given sufficient statistics in a speaker verification task. We formulate a maximum a posteriori (MAP) algorithm for the estimation of feature transform parameters, that further alleviates the possible numerical problem. The FT parameters are then vectorized and compared via a support vector machine (SVM). Similar training and development data as mentioned in Section 2.2.2 were used.

2.3 Fusion

Our primary system adopted the following linear fusion model:

$$\hat{s} = w_0 + \sum_{i=1}^{N} w_i s_i \tag{1}$$

where s_i is the score from the ith subsystems and N is the total number of subsystems. We optimize the weights based on the minimum DCF criterion. The weights are tuned on development set designed based on the NIST SRE08 and SRE08 follow-up data. The optimum weights were found via brute-force search.

3. TRAINING DATA

The training data were drawn primarily from NIST SRE 2004, SRE 2005, and SRE 2006, Mixer 5 interview data, SRE 2008, Follow-up data for SRE2008, and Switchboard. Only English samples were used. We designed a development set to match the condition anticipated for SRE10. In particular, the probability of target was set to 0.01 for the development data drawn from the SRE08 and follow-up data. The other datasets were for UBM training, channel compensation, and score normalization.

4. PROCESSING SPEED

The processing speed of the system is measured based on the Xeon 2.13 GHz processor with 1 Gb RAM. The breakdown of the runtime factor is summarized in Table III for each of the classifiers.

Table III The runtime factors comparison of classifiers used in the system.

Classifier	Runtime Factor (×RT)
GMM-UBM-JFA	0.50
GMM-SVM-KL	0.08
GMM-SVM-BHATT	0.10
GMM-SVM-FT	0.28

5. REFERENCES

1] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, 10(1):19--41,2000.

[2] L. F. Lamel and L. R. Rabiner, “An improved endpoint detector for isolated word recognition,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 4, Aug. 1981.

[3] H. Sun, B. Ma and H. Li, “An efficient feature selection method for speaker recognition”, in *Proc. Chinese Spoken Language Processing (ISCSLP)*, pp. 181-184, 2008.

[4] 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, 2007.

[5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” Avail http://www.crim.ca/perso /patrick.kenny/.

[6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” Digital Signal Processing, vol. 10, no 1-3, pp. 42-54, Jan 2000.

[7] 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*, 2006, pp. 97-100.

[8] W. Campbell, D. Sturim, D.A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, 13(5): 308—311, May 2006.

[9] C. -C. Chang and C. -J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ ~cjlin/libsvm

[10] C. You, K. A. Lee, and H. Li, “GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition,” accepted for publication in *IEEE Trans. Audio, Speech, and Language Processing*.

[11] D. Zhu, B. Ma and H. Li, “Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition," in *Proc. ICASSP*, 2009.