THE I4U SYSTEM IN NIST 2008 SPEAKER RECOGNITION EVALUATION

Haizhou Li[1,3,4], Bin Ma[1], Kong-Aik Lee[1], Hanwu Sun[1], Donglai Zhu[1], Khe Chai Sim[1], Changhuai You[1], Rong Tong[1,4], Ismo Kärkkäinen [1], Chien-Lin Huang[1], Vladimir Pervouchine[1], Wu Guo[2], Yijie Li[2,4], Lirong Dai[2], Mohaddeseh Nosratighods[3], Thiruvaran Tharmarajah[3], Julien Epps[3], Eliathamby Ambikairajah[3], Eng-Siong Chng[4], Tanja Schultz[5], Qin Jin[5]

[1]Institute for Infocomm Research (IIR), Singapore,
[2]University of Science and Technology of China (USTC), China,
[3]The University of New South Wales (UNSW), Australia,
[4]Nanyang Technological University (NTU), Singapore,
[5]Carnegie Mellon University (CMU), USA
{hli, mabin}@i2r.a-star.edu.sg

ABSTRACT

This paper describes the performance of the I4U speaker recognition system in the NIST 2008 Speaker Recognition Evaluation. The system consists of seven subsystems, each with different cepstral features and classifiers. We describe the I4U Primary system and report on its core test results as they were submitted, which were among the best-performing submissions. The I4U effort was led by the Institute for Infocomm Research, Singapore (**I**IR), with contributions from the University of Science and Technology of China (**U**STC), the University of New South Wales, Australia (**U**NSW), Nanyang Technological University, Singapore (NT**U**) and Carnegie Mellon University, USA (CM**U**).

Index Terms— Speaker recognition, classifier, channel variability, system fusion

1. INTRODUCTION

The NIST 2008 Speaker Recognition Evaluation (SRE, http://www.nist.gov/speech/tests/sre/2008) is distinguished from previous evaluations by including, in the training and test conditions of the core test, not only conversational telephone speech data (telephone data) but also (i) conversational telephone speech data recorded over a microphone channel (microphone data), and (ii) conversational speech recorded over a microphone channel involving an interview scenario (interview data). This prompted participants to apply effective channel compensation techniques and to adopt adequate system development strategies.

Recent advancements in speaker recognition are partly attributed to the effective use of acoustic features in speaker characterization. Some focus on the short-term spectral features, while others exploit temporal characteristics. Popular speaker modeling techniques include Gaussian mixture modeling with universal background model (GMM-UBM) [1], generalized linear discriminant sequence (GLDS) kernel by expanding acoustic features using a monomial basis [2], support vector machine modeling on GMM supervectors (GMM-SVM) [3], and MLLR transforms as features for support vector machine modeling (MLLR-SVM) [4].

The I4U system is solely based on acoustic features. In addition to the conventional MFCC, LPCC and PLP features, we have also included short-time frequency with long-time window (SFLW) [8], frequency modulation (FM) features [9], and channel-compensated MFCC in some of the classifiers [10]. Besides the GMM-UBM, GLDS-SVM and GMM-SVM classification techniques, new endeavors have been attempted by applying feature transformation on GMM-SVM (FT-SVM) [5], probabilistic sequence kernel based SVM (PSK-SVM) [6], and Bhattacharyya kernel based GMM-SVM [7].

In view of the fact that channel and session variability is one of the major challenges in the NIST 2008 SRE, we apply compensation techniques in all classifiers, including nuisance attribute projection (NAP) [11], joint factor analysis (JFA) or eigenchannel compensation (EIG) [12]. We paid extra attention to the compensation strategy, which is critical to overall system performance.

2. SYSTEM DESCRIPTION

The I4U system consists of three main modules, namely (i) feature extraction, (ii) a parallel bank of seven classifiers and (iii) system fusion. Two of the classifiers are based on the generative GMM-UBM approach, while the other five are based on discriminative SVM techniques. We implemented different acoustic features in combination with various classifiers, to achieve subsystem diversity. For brevity, we only discuss in this paper the seven main classifiers (subsystems) that are pivotal to the overall performance, as summarized in Table I.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

Table I. Acoustic features and classifier techniques (both generative$^+$ and discriminative*)

Classifier	Acoustic Features
GMM-UBM-EIG$^+$	*SFLW*
GMM-UBM-JFA$^+$	*PLP*
GLDS-SVM*	*Channel-compensated MFCC*
GMM-SVM*	*MFCC, FM*
FT-SVM*	*MFCC*
PSK-SVM*	*MFCC*
Bhattacharyya kernel*	*LPCC*

2.1. Feature Extraction

In feature extraction, we first applied energy-based voice activity detection (VAD) to remove silence frames and to retain only the high quality speech frames. An input utterance was then converted to a sequence of feature vectors. Finally, the feature vectors were processed by mean-variance-normalization (MVN), RASTA, and feature warping. We used five different types of cepstral features and one FM feature, as shown in Table I. Both MFCC and LPCC features have 36 dimensions. The SFLW [8] is specially designed to account for both the short-time spectral characteristics and long-time resolution. The frequency modulation (FM) features [9] capture dominant frequency deviation in sub-bands. Feature-level channel compensation was performed on MFCCs as in [10], where channel adaptation factors were computed using probabilistic subspace adaptation [13].

2.2. Classifiers

Next we describe the classifier specifications and strategies. The development datasets are also summarized in Table II.

2.2.1. GMM-UBM-EIG

GMM-UBM-EIG is a GMM-UBM [1] with eigenchannel adaptation [12]. We used the NIST 2004 SRE (SRE04) 1conv4w (1 conversation) data to train the gender-dependent UBMs, each having 512 Gaussian mixture components. The telephone data from SRE04 and SRE06, together with the microphone data of SRE05, SRE06 and the interview data recorded in the LDC Mixer 5 project (distributed by NIST to participants), were used for eigenchannel adaptation. The number of eigenchannels was set to 30 empirically. The telephone data of SRE05 and SRE06 were used for score normalization of TNorm [14] and ZNorm [15] in both GMM-UBM-EIG and GMM-UBM-JFA classifiers.

2.2.2 GMM-UBM-JFA

Another GMM-UBM classifier is equipped with joint factor analysis (JFA) as the main channel compensation technique [12]. We used the SRE04 1conv4w data to train the gender-dependent UBMs, each having 1024 Gaussian mixture components. Switchboard II and SRE04 data were used to train the speaker space with 300 speaker factors. For channel space training, we used telephone data from SRE04, SRE05 and SRE06 to train the telephone channel space (100 channel factors), microphone data from SRE05 and SRE06 to train the microphone channel space (50 channel factors) and the Mixer 5 interview data to train the interview channel space (50 channel factors). The channel factors for different channel conditions were trained independently. We then combined these three subspaces to obtain a channel space of 200 channel factors. This subspace training scheme was motivated by considering that we have much fewer training data for the interview channel. Experimental results showed that the scheme was a wise and effective attempt to create a balanced representation of the three different channels.

2.2.3 GLDS-SVM

The GLDS-SVM classifier follows the architecture in [2]. The 36-dimension feature vectors extracted from an utterance were expanded to a higher dimensional space by employing all monomials up to order 3, thus resulting in a feature space of 9,139 dimensions. The expanded features were then averaged to form a single vector for each of the utterances. It is also assumed that the kernel inner product matrix is diagonal for computational simplicity. We used the same development dataset for all five SVM classifiers (see Table II). SRE04 data were used as the background data set; the telephone and microphone data from SRE04, SRE05 and SRE06, and the interview data of Mixer 5 were applied for NAP training; and the 1conv4w telephone data of SRE05 were used for TNorm. For the NAP training, the same scheme of individual channel space training and sub-space combination as that for JFA was adopted.

2.2.4 GMM-SVM

This classifier uses GMM supervectors to construct SVM kernels [3]. Given a speaker's utterance, a GMM is estimated by using MAP-adapted means of the UBM of 512 Gaussian mixture components. The mean vectors of mixture components in the GMM are then concatenated to form a supervector, which is used in the SVM kernels. NAP was used to compensate for the channel effects.

2.2.5 FT-SVM

This novel classifier uses parameters of a feature transformation (FT) function to form the supervectors [5]. The FT function is defined in such a way that the transformation matrices and bias vectors are controlled by different regression classes. The number of bias vectors can be set to be more than that of transformation matrices because the estimation of a bias vector is believed to be more robust. An iterative training procedure is carried out for the MAP estimation of the FT parameters from the UBM model. The UBM is a gender-dependent GMM with 512 Gaussian mixture components. The FT function has one transformation matrix and 512 bias vectors.

2.2.6 PSK-SVM

The probabilistic sequence kernel (PSK) SVM system [6] consists of two major elements, namely, (i) a generative front-end that performs nonlinear mapping of cepstral features into a characteristic vector, and (ii) SVM models defining the hyperplanes that separate a target speaker from the background speakers. The front-end bases were obtained by aggregating the GMMs of 72 speakers selected from a background dataset which gave the largest scattering measure. Each GMM speaker model has 256 Gaussian components, resulting in $72 \times 256 = 18,432$ Gaussian bases. The background speakers' GMMs were pooled together with equal weights to form an ensemble of front-end bases. To form the characteristic vector, we only evaluated the top 10 Gaussians that had higher likelihood probabilities in each of the background GMMs, turning off the rest, for computational efficiency.

2.2.7 Bhattacharyya kernel

We developed a GMM-SVM system with the Bhattacharyya kernel. In the conventional Kullback-Leibler (KL) kernel, only mean vectors are adapted from the UBM. The Bhattacharyya-based SVM kernel accounts for both mean and covariance statistics based on the Bhattacharyya distance, which measures the discrepancy between two probability distributions. In this way, the Bhattacharyya kernel is believed to characterize the speakers in a better way. We trained a UBM of 512 GMM components, SVM background and NAP projection matrix on SRE04 data.

Table II. Classifier development datasets

Classifier	Development Dataset		
	UBM /Background	Channel Space	TZNorm
GMM-UBM-EIG	*SRE04*	*SRE04, SRE05, SRE06, Mixer 5*	*SRE05 (T)* *SRE06 (Z)*
GMM-UBM-JFA			
GLDS-SVM			*SRE05 (T)*
GMM-SVM			
FT-SVM			
PSK-SVM			
Bhattacharyya kernel			

2.3 Fusion

The I4U Primary submission adopted the linear fusion:

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

where s_i is the score from the *i*th classifiers and *N=7* is the total number of classifiers. We optimize the weights for minimum DCF on the development set.

$$(\widehat{w}_i, \widehat{w}_0) = \arg\min_{w_i, w_0} DCF(\sum_{i=1}^{N} w_i s_i + w_0) \tag{2}$$

w_i is adjusted iteratively using numerical optimization, subject to sum of 1.0 and w_0 is given by the threshold which yields minimum DCF.

The core test of SRE08 was designed to have a *short2* training condition, which involves either telephone or interview speech, and a *short3* test segment condition which involves telephone, microphone or interview speech. As interview-microphone is not included, there were 5 training-test conditions in combination. We created a development dataset for system fusion and threshold-setting using SRE06 1conv4w (telephone data), SRE06 1convmic (microphone data) and Mixer 5 (interview data) data sets. There are only 6 speakers (3 males and 3 females) in the Mixer 5 dataset, with each speaker having 6 sessions, and each session having 9 recordings of around 30 minutes. We split each Mixer 5 recording into 6 trials in the development dataset.

We created a development dataset to cover the imposter and genuine trials of telephone-telephone, telephone-microphone and interview-interview conditions, and imposter trials only for telephone-interview and interview-telephone conditions. The whole development dataset was divided into two disjoint halves, one as a *Tuning Set* for calibrating fusion weights and another as an *Evaluation Set* for evaluating the performance. We estimated the fusion weights and thresholds for each *short2-short3* pair separately. For those pairs for which we only had imposter trials, we adopted the fusion parameters estimated from the telephone-telephone development dataset.

3. EVALUATION RESULTS

Table III reports the performance comparison of the seven individual classifiers as well as fusion system on the telephone-telephone development dataset. We report the performance in terms of EER and minimum DCF score (minDCF). It is worth noting that GMM-UBM-JFA demonstrates an outstanding performance, which is attributed to the effective channel space training [12].

Table III. Performance of individual classifiers and the fused system on the telephone-telephone development dataset.

	Tuning Set		*Evaluation Set*	
	EER (%)	minDCF	EER (%)	minDCF
GMM-UBM-EIG	5.47	0.0270	5.22	0.0243
GMM-UBM-JFA	3.19	0.0168	3.11	0.0160
GLDS-SVM	4.30	0.0238	4.44	0.0208
GMM-SVM	4.47	0.0238	4.43	0.0205
FT-SVM	4.20	0.0222	3.66	0.0189
PSK-SVM	5.29	0.0266	4.77	0.0230
Bhattacharyya kernel	4.46	0.0246	5.16	0.0242
Best Individual	3.19	0.0168	3.11	0.0160
Fusion	2.49	0.0122	2.05	0.0122
Improvement	21.94%	27.38%	34.08%	23.75%

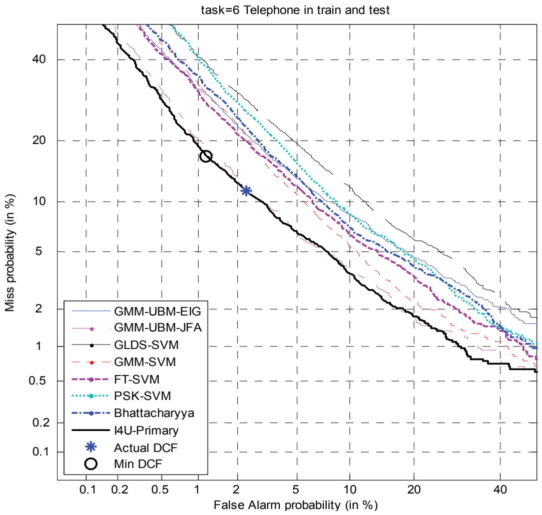

Figure 1. DET curves of individual classifiers and the fused system on the SRE08 core test (*short2-short3*).

Table IV. Performance of individual classifiers and the fused system on the *Evaluation Set* and SRE08 short2-short3 core test.

	Evaluation Set	SRE08 (*short2-short3*)	
	minDCF	minDCF	Actual DCF
GMM-UBM-EIG	0.0330	0.0347	-
GMM-UBM-JFA	0.0223	0.0260	-
GLDS-SVM	0.0475	0.0447	-
GMM-SVM	0.0290	0.0382	-
FT-SVM	0.0310	0.0416	-
PSK-SVM	0.0527	0.0475	-
Bhattacharyya kernel	0.0405	0.0442	-
Best Individual	0.0223	0.0260	-
Fusion	0.0073	0.0202	**0.0239**
Improvement	67.26%	22.30%	-

Figure 1 illustrates the detection error tradeoff (DET) curves of all seven classifiers and the fusion system (I4U Primary) in SRE08 with all short2-short3 trials being pooled together in the 5 training-test conditions. Table IV summarizes the performance in terms of minDCF in the SRE08 core test (short2-short3). For reference, we also report the performance on the Evaluation Set in the development dataset. It is observed that a 22.3% minDCF reduction over the best individual classifier was achieved by the fusion system.

4. CONCLUSION

The I4U Primary system effectively fuses multiple classifiers using acoustic features to achieve promising results. The superior performance is generally attributed to the JFA and NAP strategies for channel compensation with GMM-UBM-JFA being the best single classifier. We trained the channel factors in JFA and channel spaces in NAP independently for the telephone, microphone and interview channels, and combined them together. The I4U system has greatly benefited from this strategy, albeit with limited and unbalanced training samples across channels and speaker.

5. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Computer Speech and Language,* vol. 20, pp. 210-229, 2006.

[3] W. M. Campbell, D, Sturim, and D. A. Reynolds, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *Proc. ICASSP*, 2006.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg and A. Venkataraman, “MLLR transforms as features in speaker recognition,” in *Proc. ICASSP*, 2005.

[5] D. Zhu, B. Ma and H. Li, “Using MAP estimation of feature transformation for speaker recognition”, in *Proc. Interspeech, 2008*.

[6] K. A. Lee, C. You, H. Li, T. Kinnunen, and D. Zhu, “Characterizing speech utterances for speaker verification with sequence kernel SVM,” in *Proc. Interspeech*, 2008.

[7] C. H. You, K. A. Lee and H. Li, “An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition,” IEEE Signal Processing Letters, vol. 16, no. 1, pp. 49-52, Jan. 2009.

[8] C.-L. Huang, B. Ma, C.-H. Wu, B. Mak and H. Li “Robust speaker verification using short-time frequency with long-time window and fusion of multi-resolutions”, in *Proc. Interspeech*, 2008.

[9] T. Thiruvaran, E. Ambikairajah, and J. Epps, “Extraction of FM components from speech signals using an all-pole model”, *IET Electronics Letters*, vol. 44, no. 6, March 2008, pp. 449-450.

[10] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on ASLP*, vol. 15, no. 7, 2007.

[11] A. Solomonoff, W. M. Campbell and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proc. ICASSP*, 2005.

[12] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.

[13] S. Lucey and T. Chen, "Improved Speaker Verification through Probabilistic Subspace Adaptation," in *Proc. Eurospeech*, 2003, pp. 2021–2024.

[14] 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.

[15] K. P. Li and J. E. Porter, “Normalizations and selection of speech segments for speaker recognition scoring,” in *Proc. ICASSP*, vol. 1, pp. 595–598, 1988.