A STUDY ON GMM-SVM WITH ADAPTIVE RELEVANCE FACTOR AND ITS COMPARISON WITH I-VECTOR AND JFA FOR SPEAKER RECOGNITION

Chang Huai You, Haizhou Li, Bin Ma, Kong Aik Lee

Institute for Infocomm Research (I^{2}R), A*STAR, Singapore 138632

{echyou,hli,mabin,kalee}@i2r.a-star.edu.sg

ABSTRACT

Recently, joint factor analysis (JFA) and identity-vector (i-vector) represent the dominant techniques used for speaker recognition due to their superior performance. Developed relatively earlier, the Gaussian mixture model - support vector machine (GMM-SVM) with nuisance attribute projection (NAP) has gradually become less popular. However, when developing the relevance factor in maximum *a posteriori* (MAP) estimation of GMM to be adapted by application data in place of the conventional fixed value, it is noted that GMM-SVM demonstrates some advantages. In this paper, we conduct a comparative study between GMM-SVM with adaptive relevance factor and JFA/i-vector under the framework of Speaker Recognition Evaluation (SRE) formulated by the National Institute of Standards and Technology (NIST).

Index Terms— maximum *a posteriori*, Gaussian mixture model, support vector machine, joint factor analysis, i-vector, PLDA

1. INTRODUCTION

The GMM-UBM technique has shown reliable performance for text-independent speaker recognition [1, 2, 3]. A GMM carries rich amount of information from its corresponding utterance. Besides the desired speaker information in the GMM, it also contains other information such as channel and duration of test utterance. Such distraction is considered as nuisance to the GMM and thus results in mismatch between the training and testing conditions.

Over the past few years, the use of SVM for speaker modeling in the GMM-supervector space [4] has shown significant performance improvement over the GMM-UBM baseline. The success of the method was mainly due to the proper combination of the generative Gaussian model and discriminative support vector in a constructive way. Furthermore, when the NAP [5] was introduced to deal with the channel effect, the channel mismatch problem was compensated to a large extent [6, 7, 8, 9]. Almost in the same period, the JFA approach [10] has shown the state-of-the-art performance in speaker recognition. It was reported to be effective due to its efficient analysis on speaker factors [11] and channel factors [12], where a GMM-supervector is viewed as a combination of different supervectors. JFA compensates the channel variation through eigenchannel modeling and emphasizes the speaker-dependent component by using low dimension speaker factor through eigenvoice modeling.

Presently, the i-vector technique that was originated from JFA brings a new height to speaker recognition and becomes the most popular [13, 14]. The i-vector extractor converts a sequence of features into a single low-dimensional vector in the total variability space, by which speech segment of variable length can be represented as fixed-length vector. In this regard, linear discriminant analysis (LDA) [15], probabilistic LDA (PLDA) [16, 17], and the heavy-tailed PLDA [18, 19] are useful for i-vector system.

With the JFA and i-vector become the de facto mainstream in speaker recognition, GMM-SVM has been becoming less popular. In our previous work [20, 21], we developed the adaptive relevance factor with respect to the application data instead of some fixed empirical value for GMM-SVM. As compared to conventional GMM-SVM, the GMM-SVM with the adaptive relevance factor demonstrates competitive properties. In this paper, we setup a database platform and provide an investigation to compare GMM-SVM with the adaptive relevance factor, JFA and i-vector under the same platform. The series of the NIST SRE [22] has provided a benchmarking platform for the research in text-independent speaker recognition for more than a decade. In these evaluations, the speaker recognition task has always prescribed trials where (i) genders are not mixed and (ii) the genders of the speakers involved are given. Thus, in this study, we use database from NIST SREs to evaluate the three popular techniques, i.e., GMM-SVM, JFA and i-vector; and adopt gender-dependent mode for all speaker recognition systems. We note that the three techniques are all GMM based, therefore, in our experiment, the three techniques share the same UBM for fair comparison.

In the remainder of the paper, the adaptive relevance factor of MAP for GMM-SVM is introduced in Section 2. The JFA and i-vector is briefly described in Section 3. The database assignment is described and the performance measure is reported in section 4. The conclusion is given in Section 5.

2. GMM-SVM WITH ADAPTIVE RELEVENCE FACTOR

An UBM can be denoted by the set of parameters, $\mathrm{u} = \{\bar{\omega}_i, \bar{\mathbf{m}}_i, \bar{\Sigma}_i; i = 1, 2, ..., C\}$, where C is the number of Gaussian components. The adapted GMM, λ, takes a similar form $\lambda = \{\omega_i, \mathbf{m}_i, \Sigma_i; i = 1, 2, ..., C\}$ where $\mathbf{m}_i$, Σ_i, ω_i are respectively the mean vector, the covariance matrix, and the weight of the ith Gaussian component.

2.1. Adaptive Relevance Factor in MAP

In conventional MAP, λ is obtained by

$$\breve{\lambda} = \arg\max_{\lambda} \left[f(\mathbf{X}|\lambda) g(\lambda) \right] \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\kappa]$ is the sequence of feature vectors, which we call the adaptation data. $\mathbf{x}$ is a J-dimensional feature vector. $f(\mathbf{X}|\lambda)$ is the likelihood of $\mathbf{X}$ given a GMM λ. $g(\lambda)$ is prior density of the GMM λ.

Assuming that the weights that are required to be a conjugate distribution are modeled as a Dirichlet density $g_1(\omega_1, ..., \omega_C)$ while mean and covariance of GMM is a conjugate prior distribution with normal-Wishart density $g_2(\mathbf{m}_i, \Sigma_i)$. $g(\lambda)$ is the joint prior density

of g_1 and g_2. We have the mean and covariance parameters of the ith Gaussian adapted as follows [23],

$$\mathbf{m}_i = \alpha_i \breve{\Xi}_i + (1 - \alpha_i)\bar{\mathbf{m}}_i \tag{2}$$

$\breve{\Xi}_i$ is the first order sufficient statistics; α_i are the adaptation coefficients given by

$$\alpha_i = \frac{N_i}{N_i + \gamma_i} \tag{3}$$

The relevance factor γ_i is a constant parameter in the normal-Wishart density as which the Gaussian parameters are modeled [23]; N_i is the occupation count which is directly proportional to the duration of the feature sequence.

Let $\bar{\mathbf{m}}$ be the UBM-supervector. We assume that a GMM-supervector $\mathbf{m}(\lambda)$ is given by the sum of $\bar{\mathbf{m}}$ and a speaker-dependent supervector $\Phi\mathbf{z}(\lambda)$:

$$\mathbf{m}(\lambda) = \bar{\mathbf{m}} + \Phi\mathbf{z}(\lambda) \tag{4}$$

Φ denotes a diagonal transfer matrix and the vector $\mathbf{z}(\lambda)$ is speaker (or language) specific. To this end, we assume that Gaussian components in the GMM are functionally independent, and the vector $\mathbf{z}(\lambda)$ is of a standard normal distribution. Given the observed data $\mathbf{X}$, maximizing the posterior probability $P(\mathbf{z}(\lambda)|\mathbf{X})$ with respect to $\mathbf{z}$ gives $\breve{\mathbf{z}} = \arg\max_{\mathbf{z}} P(\mathbf{z}(\lambda)|\mathbf{X}) = \beta$, where $\beta = \zeta^{-1}(\lambda)\Phi^*\Sigma^{-1}N(\breve{\Xi} - \bar{\mathbf{m}})$, N denotes the occupation count, we arrive at

$$\hat{\mathbf{m}} = \bar{\mathbf{m}} + \Phi\breve{\mathbf{z}} = \breve{\alpha}\breve{\Xi} + \bar{\mathbf{m}}(1 - \breve{\alpha}) \tag{5}$$

where $\breve{\alpha} = (\Phi^{-2}\Sigma + N)^{-1}N$. As compared to the conventional MAP in (2), Eq. (5) shows that $\breve{\alpha}$ is the adaptation coefficient. Therefore, the relevance factor can be given by [10, 20, 21]

$$\breve{\gamma} = \Phi^{-2}\Sigma \tag{6}$$

The relevance factor in (6) is data dependent since the parameter Φ is estimated with expectation-maximization (EM) algorithm based on a training dataset [21].

In discriminative classification using SVM, the supervector used to represent a certain speaker is required to be relatively stable without being affected by the duration variation.

In [20, 21], we introduce the adaptive relevance factor in which we feed in additional term to (6) so that it could adapt to the duration variation from one utterance to the other

$$\tilde{\gamma} = \theta_0\kappa\Phi^{-2}\Sigma \tag{7}$$

where κ denotes the duration of the utterance, and θ_0 is a constant which is determined empirically based on a given database.

Different from conventional GMM-SVM which uses a fixed relevance factor, the GMM-SVM with adaptive relevance factor applies (7) to each GMM-supervector. As we only obtain the parameter Φ once during training, the computation at run-time recognition process only involves an additional multiplication.

2.2. SVM

The discriminant score of the SVM is given by [7, 8, 24]

$$f(\mathbf{X}) = \sum_{l=1}^{L} \alpha_l y_l K(\mathbf{X}_l, \mathbf{X}) + b \tag{8}$$

where L is the number of support vectors, and $\mathbf{X}_l$ is the lth support vector, α_l is the weight assigned to the lth support vector with its label given by $y_l \in \{-1, +1\}$ and b is the bias parameter. K is the SVM kernel used to measure the similarity of the support vector and given vector. In this study, we use the Bhattacharyya-based kernel referred to as the GMM-UBM Mean Interval (**GUMI**) in [6, 7] as the SVM kernel.

3. JFA AND I-VECTOR

The JFA has been reported to have superior performance due to its robustness in channel compensation. Recently, Dehak et al. [13] proposed a feature extractor inspired by the JFA. Unlike JFA which models separately speaker and channel variability in a high dimension space of supervectors, Dehak's idea consists in finding a low dimensional subspace of the GMM-supervector space, named the total variability space that represents both speaker and channel variability. The vectors in the low-dimensional space are called i-vectors.

3.1. JFA

In JFA, the speaker variability is modeled by the eigenvoice, where several common factors are used to represent the spanned space of the speaker, while the channel variability is modeled using a set of latent variables channel factors. In particular, a speaker-dependent GMM-supervector s can be decomposed in joint factors as follows [10]

$$\mathbf{m} = \bar{\mathbf{m}} + V\mathbf{v} + U\mathbf{u} + D\mathbf{d} \tag{9}$$

where $\bar{\mathbf{m}}$ is a speaker-independent supervector from UBM, V is the eigenvoice matrix, $\mathbf{v}$ is the eigenvoice factors (or speaker factors) with normal prior distribution; U is the eigenchannel matrix, and $\mathbf{u}$ the channel factors with normal prior distribution; D is the residual diagonal matrix, and $\mathbf{d}$ denotes the speaker-specific residual factors with normal prior distribution.

As a result of the decomposition in (9), speaker adaptation can be performed by updating a set of speaker-dependent latent variables and minimizing the influence of channel effects in an utterance. In particular implementation, we train the eigenvoice matrix V by assuming U and D to be zeros; then train the eigenchannel matrix U given the estimate of V by assuming D to be zero; finally D matrix is trained given the estimates of V and U. In the training database design, for V matrix, we focused on obtaining the speaker-based principal dimensions; for the U matrix, the key is to obtain the channel (or nuisance) based principle dimensions. With the trained matrices V, U and D , the estimate of $\mathbf{v}$, $\mathbf{u}$ and $\mathbf{d}$ are obtained based on the posterior means given the particular utterance.

The score can be obtained by comparing the target speaker speech side and test segment statistics as follows

$$\begin{aligned} &Score = \\ &(V\mathbf{v}_{tar} + D\mathbf{d}_{tar})^T\Sigma^{-1}(\Xi_{test} - N_{test}\bar{\mathbf{m}} - N_{test}U\mathbf{u}_{test}) \end{aligned} \tag{10}$$

where $\mathbf{v}_{tar}$ and $\mathbf{d}_{tar}$ are the target speaker factors and residual factors; while Ξ_{test}, N_{test}, and $\mathbf{u}_{test}$ are the first order sufficient statistics, zero-order statistics (or occupation count), and the channel factors of the test speech utterance(s). We can see that the target speaker side is centered around speaker and residual factors, while the test speech has speaker-independent and channel factors removed. In score normalization, the z-norm and t-norm is used since they have been proven to effectively reduce the variability of the likelihood ratio scores that are used in the decision criterion.

3.2. i-Vector

Comparing to the supervector used in GMM-SVM and JFA, the i-vectors are smaller in size to reduce the execution time of the recognition task while maintaining recognition performance similar to that obtained with JFA. A key ingredient to the success of this approach was the enormous quantity of data used to extract the i-vector feature set. In other words, the i-vector is a low-dimensional representation of an entire speech segment. It has been shown to respond well to generative modeling. Actually, the i-vector estimate is calculated by evaluating the posterior expectation of the hidden variables in the model conditioned on the Baum-Welch statistics extracted from the utterance. This posterior calculation provides a posterior covariance matrix as well as a posterior expectation. The posterior covariance matrix can be interpreted as quantifying the reliability of the point estimate. An i-vector system uses a set of low-dimensional total variability factors w to represent each utterance. Each factor controls an eigen-dimension of the total variability matrix T. The total variability factors w is the i-vector. In particular, the GMM-supervector m can be decomposed into speaker-independent supervector $\bar{m}$ and the speaker-dependent supervector $T\mathbf{w}$

$$\mathbf{m} = \bar{\mathbf{m}} + T\mathbf{w} \tag{11}$$

To train T, just using the same procedure used for training V in JFA but treat all utterance of all training speakers as belonging to different speakers. Thus T actually absorbs the information of V, U and D in JFA. For each utterance, the i-vector $\mathbf{w}$ can be obtained given T.

In fact, i-vector extractors are trained without speaker-level labeling. It indicates that further transformations should apply in order to increase their speaker discriminative capacity. In i-vector system, a score can be obtained by comparing the enrollment i-vector and the test i-vector. It was shown that by projecting i-vectors onto a Linear Discriminative Analysis (LDA) basis, trained using representative enrollment data and speaker-labels to defined classes, the performance can be improved significantly. More effective performance can be obtained by giving the score with PLDA where the i-vector is considered as the second layer input vector to PLDA system [16].

There are two versions of PLDA named Gaussian and heavy-tailed versions. Currently, Gaussian PLDA [16, 17] and heavy-tailed PLDA [18], performed either on i-vectors directly or on the LDA-projected length-normalized i-vectors, yield state-of-the-art speaker recognition results. i-vectors can be approximately Gaussianized by length normalization so that the performance of Gaussian PLDA with length normalization is similar to that of heavy-tailed PLDA without length normalization. The recent research results show that unity length normalization of the i-vector indicates that Gaussian PLDA is as effective as heavy-tailed PLDA. In this investigation, we chose Gaussian PLDA for the speaker recognition.

In particular, given a speaker and a collection of i-vectors $\mathbf{w}_1 j, ..., \mathbf{w}_R j$ (one for each recording of the speaker in jth style (or channel or session)), standard Gaussian PLDA assumes that the i-vectors are distributed according to

$$\mathbf{w}_{rj} = \varpi + \Omega \mathbf{h}_r + \Lambda \mathbf{q}_{rj} + \epsilon \tag{12}$$

incorporating speaker subspace Ω and channel subspace Λ. ϖ is the overall mean of the i-vectors. $\mathbf{h}$ and $\mathbf{q}$ are hidden variables representing the speaker factors and channel factors respectively; and they have standard normal priors. The residual ϵ_r is normally distributed with zero mean and diagonal covariance matrix. The PLDA is modeled by the parameters ϖ, Ω, Λ, and ϵ_r, which can be estimated through EM algorithm using the parameter training database.

Table 1. The training dataset list for GMM-SVM for SRE08 task

ITEM	Data Resource	#utts:f	#utts:m
UBM	$S04(t)_1$ + $S06(t)_1$	5651	4116
Φ-matrix	$S04(t)_2$+$S05(m)_2$+$S06(t)_2$ +$S06(m)_2$+$S08(i)_2$	7035	6786
NAP	$S04(t)_3$+$S05(t)_3$+$S05(m)_3$ + $S08(i)_2$	6801	6035
SVM-i	$S04(t)_4$	2532	2359
T-Norm	$S05(t)_5$+$S05(m)_5$	561	502
Z-Norm	$S06(t)_6$+$S06(m)_6$	361	245

To make inference of the identity of a given test segment, the posterior probability for both enrollment i-vector and test i-vector generated from the same speaker or from different speaker are computed based on PLDA model. So, the log-likelihood ratio for the same and different inference likelihood is obtained as the output of the PLDA system. It has been proven that ignoring the channel subspace Λ and using full covariance matrix of ϵ_r instead of the diagonal matrix can be effective for speaker recognition system. Therefore the PLDA system in the investigation adopts this way. Finally, the S-norm is applied for score normalization [18].

4. PERFORMANCE EVALUATION

In this investigation, the GMM-SVM system is implemented by using the GUMI kernel; and we use 512 mixture components with gender-dependent mode. For the channel compensation, the NAP is used and its rank is set to 60 for the GMM-supervector with 52 dimension of MFCC. We trained a diagonal matrix Φ by using EM algorithm with $\Phi_i^{(0)} = (\bar{\Sigma}_i)^{-\frac{1}{2}}$ as the initial matrix. The default value of θ_0 is empirically set to 8.2×10^{-4}. It is done by using a reference value that is obtained by setting the fixed relevance factor to 1. According to (7), by taking the average over all Gaussian components, we have a default value of θ_0 to be $\theta_0 \approx \kappa_0^{-1} \frac{1}{C} \sum_{i=1}^{C} (\Phi_i^2 \Sigma_i^{-1})$, with $\kappa_0 = \sum_{t=1}^{T_0} \kappa_t$ where κ_t is from a set of T_0 utterances representing the utterances in the application, therefore, κ_0 represents the average length of the feature data.

To build the experimental systems with a common platform, we designed the training database list for UBM, Φ-matrix, SVM background and T/Z-norm in Table 1. In the table, S0x(H)$_n$ indicates a group of the database, where x ∈ {4,6,8} denotes NIST year 2004, 2006, 2008, H∈{'t', 'm', 'i'} represents the channel[1], and n∈{1,2,...,6} indicates the speech data group index. With same x and H, the group with different n may have the database totally or partially disjointed. The GMM-SVM system with the fixed relevance factor is denoted as **conv-GS** (which represents the conventional GMM-SVM), the GMM-SVM with the adaptive relevance factor of (7) named as **adpt-GS**. The performance is measured in terms of equal error rate (EER) and minimum detection cost function (minDCF).

We setup JFA and i-vector systems with the same feature and database as GMM-SVM. For JFA system, the joint factors are composed by 300 speaker factors, 200 channel factors[2], and full rank diagonal matrix. The details of the training database list for JFA is

[1] 't' denotes the telephone channel; 'm' means the microphone channel; and 'i' denotes the interview channel.

[2] The 200 channel factors include 100 factors for telephone channel, 50 for microphone channel and the remainder 50 for interview channel.

Table 2. The training dataset list for JFA for SRE08 task

ITEM	Data Resource	#utts:f	#utts:m
UBM	$S04(t)_1$+$S06(t)_1$	5651	4116
EV	$S05(t)_3$+$S06(t)_2$	2575	2208
EC:tel	$S04(t)_2$+$S05(t)_3$+$S06(t)_2$	4142	3747
EC:mic	$S05(m)_3$+$S06(m)_2$	2355	2066
EC:itv	$S08(i)_2$	2095	2070
D	$S04(t)_4$	2532	2359
T-Norm	$S05(t)_5$+$S05(m)_5$	561	502
Z-Norm	$S06(t)_6$+$S06(m)_6$	361	245

Table 3. The training dataset list for the i-vector for SRE08 task. Note: 'TV' in the table means total variability.

ITEM	Data Resource	#utts:f	#utts:m
iVP-I:UBM	$S04(t)_1$+$S06(t)_1$	5651	4116
iVP-I:TV	$S04(t)_4$	2532	2359
iVP-I:PLDA	$S04(t)_2$+$S05(m)_2$+$S06(t)_2$ +$S06(m)_2$+$S08(i)_2$	7035	6786
iVP-I:S-Norm	$S06(t)_6$; $S06(m)_6$	361	245
iVP-II:UBM	$S04(t)_1$+$S06(t)_1$	5651	4116
iVP-II:TV	$S04(t)_4$+$S05(m)_3$+$S06(m)_2$	4887	4425
iVP-II:PLDA	$S04(t)_2$+$S05(m)_2$+$S06(t)_2$ +$S06(m)_2$+$S08(i)_2$	7035	6786
iVP-II:S-Norm	$S05(t)_5$+$S06(t)_6$; $S05(m)_5$+$S06(m)_6$	922	747

shown in Table 2. For i-vector system, the total variability is trained with 10 iterations. For the i-vector extractor matrix, 400 total variability factors are used; for PLDA training, 200 speaker factors are used for each gender. In the experiment, two sets of data assignments in Table 3 were investigated. We denote the i-vector-PLDA for set-I as iVP-I and set-II as iVP-II. Both JFA and iVP systems are gender-dependent and share the same UBM as the GMM-SVM.

Table 4 shows the EER and minDCF of the two GMM-SVM systems as compared to the JFA, iVP-I and iVP-II in SRE 2008 short2-10sec evaluation [3] [25]. It can be seen that the iVP-II gives the best EER performance, while it has very close minDCF value with **adpt-GS** which is little bit lower than iVP-II.

[3]The short2-10sec evaluation means that the channel style of the training data is telephone with total duration of five minutes or interview with total duration of three minutes while the test data is captured through telephone channel with approximate 10 seconds of speech signal.

Table 4. The comparison of GMM-SVM with various relevance factors, JFA and i-Vector on SRE 2008 short2-10sec evaluation

close-set	EER	minDCF×100
conv-GS	8.28 %	3.46
adpt-GS	7.15 %	**3.24**
JFA	7.62 %	3.87
iVP-I	7.54 %	3.67
iVP-II	**5.74** %	3.26

Table 5. The comparison of GMM-SVM with various relevance factors, JFA and i-Vector on SRE 2008 short2-short3 evaluation.

EER (%)	*itv-itv*	*itv-tel*	*tel-tel*	*tel-mic*	*tel-itv*
conv-GS	4.07	8.08	2.35	7.81	5.28
adpt-GS	3.97	6.59	**2.09**	**6.72**	3.93
JFA	3.15	**5.71**	2.30	8.19	**3.87**
iVP-I	2.73	6.32	2.45	7.38	4.89
iVP-II	**2.54**	7.96	2.27	8.29	6.19
minDCF (×100)	*itv-itv*	*itv-tel*	*tel-tel*	*tel-mic*	*tel-itv*
conv-GS	1.96	2.73	1.18	2.42	2.17
adpt-GS	1.61	2.72	1.12	**2.15**	**1.70**
JFA	1.92	**2.52**	**0.90**	2.90	1.89
iVP-I	**1.41**	3.13	1.24	2.94	2.64
iVP-II	1.46	3.70	1.22	3.27	2.88

The equal error rate (EER) and minDCF listed in Table 5 are for NIST SRE 2008 short2-short3 task. There are five conditions in this task, i.e., '*itv-itv*', '*itv-tel*', '*tel-tel*', '*tel-mic*' and '*tel-itv*'. The first channel of each condition denotes the type of the channel from which the enrollment speech data were generated while the second channel means the type of channel from which the test segment was provided. **adpt-GS** is always better than **conv-GS** in terms of EER and minDCF. It also can be seen that the iVP has quite good performance in '*itv-itv*' channel pair, while JFA performs the best for '*itv-tel*' condition in terms of EER and minDCF. **adpt-GS** gives the best performance for both '*tel-tel*' and '*tel-mic*' conditions. For '*tel-itv*' condition, JFA shows the best in terms of EER while **adpt-GS** gives the best in terms of minDCF. Generally, the conventional GMM-SVM is not competitive with JFA and i-vector, while **adpt-GS** shows its strong competitiveness.

5. SUMMARY

In this paper, we develop the JFA, i-vector and GMM-SVM using the NIST-SRE series database. The database used for UBM, normalization and most of similar function are kept as close as possible. Therefore, we can compare fairly the three main techniques here.

We investigate their performances in speaker recognition task in terms of EER and minDCF. The result shows that the conventional GMM-SVM is less effective compared to JFA and i-vector. However, the GMM-SVM with adaptive relevance factor shows to be competitive. In the NIST SRE platform, it has been observed that the state-of-the-art techniques: i-vector, JFA and GMM-SVM with adaptive relevance factor shows their competitive advantages in different training-test condition. Relatively, GMM-SVM has low computational complexity and memory requirement.

Recently, we used the above mentioned **adpt-GS** and the JFA algorithms to contribute two sub-systems for the NIST SRE 2012 evaluation, and also the two sub-systems shared the same UBM and feature databases. Their performances in SRE 2012 evaluation are effective in terms of minimum DCF and actual DCF, and the **adpt-GS** sub-system performs better than the JFA sub-system in most of the SRE 2012 conditions. Generally, the actual cost of the **adpt-GS** gives very competitive performance among all our five sub-systems.

6. REFERENCES

[1] D.A. Reynold and R.C. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, pp. 72-83, January 1995.

[2] B. L. Pellom and J. H. L. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Process. Lett.*, vol. 5, no. 11, pp. 281-284, Nov. 1998.

[3] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, pp. 2072-2084, Sep. 2007.

[4] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308-311, 2006.

[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," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, vol. 1, pp. 97-100, Toulouse, 2006.

[6] 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.

[7] C.H. You, K.A. Lee and H. Li, "GMM-SVM Kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, no. 6, pp. 1300-1312, Aug. 2010.

[8] K.A. Lee, C.H. You, H. Li, T. Kinnunen, and K.C. Sim, "Using discrete probabilities with Bhattacharyya measure for SVM-based speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 861-870, May 2011.

[9] H. Li, B. Ma, K.A. Lee, H. Sun, D. Zhu, K.C. Sim, C.H. You, R. Tong, I. Karkainen, C-L Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, E-S Chng, T. Schultz and Q. Jin, "The I4U system in NIST 2008 speaker recognition evaluation," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 4201-4204, Taipei, Apr. 2009.

[10] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," CRIM, Montreal, *Technical Report, CRIM-06/08-13*, 2005.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, " A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 5, pp. 980-988, July 2008.

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

[13] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, Support vector machines versus fast scoring in the low dimensional total variability space for speaker verification, in Proceedings of Interspeech, Brighton, UK, Sep. 2009, pp. 1559-1562.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, " Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 788-798, May 2011.

[15] M. McLaren, and D. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 5456-5459, Prague, May. 2011.

[16] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conf. on Computer Vision*, pp. 1-8, Rio de Janeiro, Brazil, Oct. 2007.

[17] Y. Jiang, K.A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification," in *Proc. Interspeech*, 2012.

[18] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[19] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 48284831, 2011.

[20] C.H. You, H. Li, and K.A. Lee, "A GMM-supervector approach to language recognition with adaptive relevance factor," *18th Europ. Signal Process. Conf.*, EUSIPCO, pp. 1993-1997, Aalborg, Denmark, Aug. 2010.

[21] C.H. You, H. Li, B. Ma, and K.A. Lee, "Effect of relevance factor of maximum a posteriori adaptation for GMM-SVM in speaker and language recognition," *Interspeech 2012*, Portland, Sep. 2012.

[22] National Institute of Standards and Technology, *NIST Speaker Recognition*, site available: http://www.itl.nist.gov/iad/mig/tests/spk.

[23] J.L. Gauvain and C-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291-298, 1994.

[24] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech and Lang.*, vol. 20, pp. 210-229, 2006.

[25] National Institute of Standards and Technology, "The NIST year 2008 speaker recognition evaluation plan," available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/.