



A Hybrid Modeling Strategy for GMM-SVM Speaker Recognition with Adaptive Relevance Factor

Chang Huai You, Haizhou Li, Kong Aik Lee

Human Language Technology Department
 Institute for Infocomm Research, A*STAR, Singapore 138632

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

Abstract

In Gaussian mixture model (GMM) approach to speaker recognition, it has been found that the maximum *a posteriori* (MAP) estimation is greatly affected by undesired variability due to varying duration of utterance as well as other hidden factors related to recording devices, session environment, and phonetic contents. We propose an adaptive relevance factor (RF) to compensate for this variability. In the other side, in realistic application, it is likely that the different channel corresponds to its different training and test conditions in terms of quantity and quality of the speech signals. In this connection, we develop a hybrid model that combines multiple complementary systems, each of which focuses on specific condition(s). We show the effectiveness of the proposed method on the core task of the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) 2008.

Index Terms: speaker recognition, Gaussian mixture model, maximum *a posteriori*,

1. Introduction

GMM is one of the most popular acoustic modeling approaches in text-independent speaker recognition due to its reliable performance. Usually a GMM is trained using the maximum *a posteriori* (MAP) adaptation from a universal background model (UBM) [1] while the UBM is trained using the expectation-maximization (EM) algorithm from a background dataset covering a wide range of speakers, sessions and channels. GMM support vector machine (GMM-SVM) has been proven a state-of-the-art technique in the speaker recognition. It is usually realized by adapting only the mean vectors, while the weights and covariance matrices remains the same as the UBM. Concatenating the mean vectors leads to the so-called supervector representation, which can be used as input to SVM [2]. In this paper, we adopt the previously proposed GMM-UBM-Mean-Interval (GUMI) kernel [3] in which the supervector carries the information from both mean and covariance.

In MAP estimation, the degree of adaptation depends on the amount of available data. The training and test data may be recorded under different channels, recording devices, and with different phonetic contents. The MAP adaptation process is greatly affected by this unwanted session variability that causes mismatch between training and testing. It is believed that such mismatch can be compensated partly by empirically adjusting the relevance factor (RF), which controls the degree of adaptation. In MAP estimation, if a mixture component has a low occupancy count on new data, the corresponding mean statistics will be de-emphasized, and emphasized otherwise. As a result, the estimates of the mean and covariance starting from the UBM

varies undesirably depending on the specificity of the available data consisting of the duration and channel-related factor.

In the GMM-UBM speaker recognition system, the RF is not so sensitive due to the nature of generative modeling [1] and therefore can be fixed. However, SVM works in a discriminative manner. In speaker recognition, a GMM supervector is used to represent the speaker characteristics via utterance and serves as an input vector to the SVM. This requires the elimination of the negative effect of the duration variation in order to manifest the saliency of the speaker characteristics. Thus, we propose an adaptive scheme for the RF that changes according to the amount of the feature data.

In addition to the extrinsic channel effects, the most recent NIST evaluation has focused on intrinsic variability of speaking styles - conversational versus interview styles. To handle this form of variability, a hybrid model is proposed to combine multiple complementary systems, each focusing on a specific speaking style. In particular, the hybrid system is designed to generate gender-dependent and gender-independent models by considering the different training and test conditions corresponding to individual channels so that the speaker recognition system can have robustness to the unknown channel variation. We evaluate the proposed method on NIST SRE 2008 core task [4].

In the remainder of the paper, we brief on the MAP estimation in section 2. We then propose the adaptive RF and the hybrid scheme for speaker recognition in section 3. The performance evaluation is reported in section 4. We summarize the paper in section 5.

2. MAP Estimation

The UBM is trained using a large data set to form a speaker-independent model [1]. The expectation maximization (EM) algorithm is usually used for this purpose. The selection of dataset is done by considering different sessions, channels and speakers. The UBM can be denoted as

$$u = \{\omega_i^{(u)}, \mathbf{m}_i^{(u)}, \Sigma_i^{(u)}; i = 1, 2, \dots, M\} \quad (1)$$

while the speaker-dependent GMM, λ , has the same form

$$\lambda = \{\omega_i^{(\lambda)}, \mathbf{m}_i^{(\lambda)}, \Sigma_i^{(\lambda)}; i = 1, 2, \dots, M\} \quad (2)$$

where \mathbf{m}_i , Σ_i , ω_i , ($i = 1, \dots, M$) are respectively the mean vector, the covariance matrix, and the weight of the i th Gaussian component.

For the MAP adaptation of λ , prior knowledge is given by the prior distribution over λ , $P(\lambda)$. With the MAP criterion, λ is selected such that it maximizes the *a posteriori* probability,

$$\lambda = \arg \max_{\lambda} P(\lambda|\mathbf{X}) = \arg \max_{\lambda} [p(\mathbf{X}|\lambda)P(\lambda)] \quad (3)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\kappa]$ is the feature vectors used to train the GMM, λ ; \mathbf{x} is a J -dimensional perceptual linear predictive (PLP) feature vector; and κ is the number of feature vectors. The parameters of the i th Gaussian are adapted as follows [1],

$$\mathbf{m}_i^{(\lambda)}(j) = \alpha_i^{(m)}(j) \overline{\Xi}_i^{(\text{utt})}(j) + (1 - \alpha_i^{(m)}(j)) \mathbf{m}_i^{(u)}(j) \quad (4)$$

$$\begin{aligned} \Sigma_i^{(\lambda)} = & \alpha_i^{(\Sigma)} \mathbf{C}_i^{(\text{utt})} + (1 - \alpha_i^{(\Sigma)}) [\Sigma_i^{(u)} + \mathbf{m}_i^{(u)}(\mathbf{m}_i^{(u)})^T] \\ & - \mathbf{m}_i^{(\lambda)}(\mathbf{m}_i^{(\lambda)})^T \end{aligned} \quad (5)$$

where $\overline{\Xi}_i^{(\text{utt})}$ and $\mathbf{C}_i^{(\text{utt})}$ are respectively the first and second order sufficient statistics, i.e.,

$$\overline{\Xi}_i^{(\text{utt})} = \frac{1}{\eta_i} \sum_{t=1}^{\kappa} \frac{\omega_i f(\mathbf{x}_t | \mathbf{m}_i^{(u)}, \Sigma_i^{(u)}) \mathbf{x}_t}{\sum_{j=1}^M \omega_j f(\mathbf{x}_t | \mathbf{m}_j^{(u)}, \Sigma_j^{(u)})} \quad (6)$$

$$\mathbf{C}_i^{(\text{utt})} = \frac{1}{\eta_i} \sum_{t=1}^{\kappa} \frac{\omega_i f(\mathbf{x}_t | \mathbf{m}_i^{(u)}, \Sigma_i^{(u)}) \mathbf{x}_t \mathbf{x}_t^T}{\sum_{j=1}^M \omega_j f(\mathbf{x}_t | \mathbf{m}_j^{(u)}, \Sigma_j^{(u)})} \quad (7)$$

and $f(\cdot)$ denotes Gaussian density function; $\alpha_i^{(\rho)}(j)$ ($\rho \in \{m, \Sigma\}$, $j = 1, \dots, J$) are the data-dependent adaptation coefficients, which are given by

$$\alpha_i^{(\rho)}(j) = \frac{\eta_i}{\eta_i + r_i^{(\rho)}(j)} \quad (8)$$

The RF $r_i^{(\rho)}$ is a parameter in the normal-Wishart density as which the Gaussian parameters are modeled. However, in conventional MAP, the RF is given as a fixed value, and the probabilistic count η_i is given by

$$\eta_i = \sum_{t=1}^{\kappa} \frac{\omega_i f(\mathbf{x}_t | \mathbf{m}_i^{(u)}, \Sigma_i^{(u)})}{\sum_{j=1}^M \omega_j f(\mathbf{x}_t | \mathbf{m}_j^{(u)}, \Sigma_j^{(u)})} \quad (9)$$

3. Robust Speaker Recognition System

3.1. Adaptive Relevance Factor

Assuming that the distribution of the mean supervector is Gaussian, then it can be shown as follows

$$\mathbf{m}_i^{(\lambda)} = \mathbf{m}_i^{(u)} + D_i \mathbf{z}_i(\lambda) \quad (10)$$

where $\mathbf{z}_i(\lambda)$ is the i th hidden variable vector distributed with normal probabilistic distribution, which contains the speaker information; D_i is the i th diagonal matrix that can be trained and it is speaker-independent. From the above assumption, the RF for mean (where $\rho \in \{m\}$) can be derived to be $r_i^{(m)} = D_i^{-2} \Sigma_i^{(u)}$, $i = 1, \dots, M$ ¹.

The idea of MAP estimation for GMM was presented in [6]. The primary purpose of the MAP is to estimate the probability distribution of the given data with known prior distribution. Insufficiency of data leads to low reliability, and vice versa. So small value of α in (8) is given for insufficient data set is reasonable, this causes the estimated GMM closer to the UBM. Otherwise, the value of α is high, so that the estimated GMM would be displaced further from the UBM. This is reflected in equations (4), (5) and (8) when we set the RF to a fix value.

In GMM-SVM system, this requires the distance from the universal speaker to the particular speaker does not vary with

¹In [5] for speaker recognition study, the similar derivation reaches the same result by assuming the supervector to be modeled by eigenchannel factor, eigenvoice factor and the residual speaker-dependent factor.

the sufficiency of an utterance, i.e. the duration of the utterance. In SVM system, the supervector is prior labeled to be target or imposter. A position in the supervector space is used to represent a specific speaker identity. This requires that the GMM supervector must stably represent the characteristics of the particular speaker regardless of sufficiency of the utterance available. In this connection, we propose an adaptive RF as follows

$$\begin{aligned} \check{r}_i^{(\rho)} &= r_i^{(\rho)} \varphi(\kappa) \\ &= r_i^{(\rho)} \left\{ \varphi(\kappa_0) + \frac{\varphi'(\kappa_0)}{1!} (\kappa - \kappa_0) + \frac{\varphi''(\kappa_0)}{2!} (\kappa - \kappa_0)^2 + \dots \right\} \end{aligned} \quad (11)$$

where φ is a hidden invariant function, κ_0 is any neighboring point which can be approximated with the average length of the utterances. According to (9), when κ increases, the probabilistic count η_i increases. Taking the expectation of the η_i , we have

$$E(\eta_i) = E\left(\sum_{t=1}^{\kappa} \frac{\omega_i f(\mathbf{x}_t | \mathbf{m}_i^{(u)}, \Sigma_i^{(u)})}{\sum_{j=1}^M \omega_j f(\mathbf{x}_t | \mathbf{m}_j^{(u)}, \Sigma_j^{(u)})}\right) \propto \kappa \quad (12)$$

where E is the expectation operator. If we chose $\varphi(\kappa) \approx \theta_0 \kappa$ by ignoring the high order polynomial terms we can arrive at

$$E(\alpha_i^{(m)}) \propto \frac{E(\eta_i)}{E(\eta_i) + \theta_0 \kappa D_i^{-2} \Sigma_i^{(u)}} \rightarrow \text{constant vector} \quad (13)$$

where θ_0 is a constant value which can be obtained from the known database. It means the expectation of the α is stable when we have the RF $\check{r}_i^{(m)}$ as follows

$$\check{r}_i^{(m)} \approx \theta_0 \kappa D_i^{-2} \Sigma_i^{(u)} \quad (14)$$

This ensures that the distance measure between the GMM supervector and UBM supervector is not seriously affected by the length of the adaptation utterance.

3.2. GUMI Kernel

In our previous work [3] [7], we derived an Bhattacharyya-based distance between two GMMs as follows

$$\begin{aligned} \check{\Psi}_{\text{Bhatt}}(p_a || p_b) &= \frac{1}{8} \sum_{i=1}^M \left\{ \left[\left(\frac{\Sigma_i^{(a)} + \Sigma_i^{(u)}}{2} \right)^{-\frac{1}{2}} (\mathbf{m}_i^{(a)} - \mathbf{m}_i^{(u)}) \right]^T \right. \\ &\quad \left. \left[\left(\frac{\Sigma_i^{(b)} + \Sigma_i^{(u)}}{2} \right)^{-\frac{1}{2}} (\mathbf{m}_i^{(b)} - \mathbf{m}_i^{(u)}) \right] \right\} \\ &+ \frac{1}{2} \sum_{i=1}^M \text{tr} \left[\left(\frac{\Sigma_i^{(a)} + \Sigma_i^{(u)}}{2} \right)^{\frac{1}{2}} (\Sigma_i^{(a)})^{-\frac{1}{2}} \left(\frac{\Sigma_i^{(b)} + \Sigma_i^{(u)}}{2} \right)^{\frac{1}{2}} (\Sigma_i^{(b)})^{-\frac{1}{2}} \right] \\ &+ \sum_{i=1}^M \ln \left\{ \frac{1}{\sqrt{\omega_i^{(a)} \omega_i^{(b)}}} \right\} - \frac{M}{2} \end{aligned} \quad (15)$$

Obviously, the distance is composed of two terms, i.e. the mean statistical dissimilarity and the covariance statistical dissimilarity. In order to avoid the unnecessary cross effect of the parameters, we consider that the mean statistical dissimilarity carries just the first-order adaptation data information with the mean vectors. According to [3], we have GUMI kernel as follows

$$\begin{aligned} K_{\text{GUMI}}(\mathbf{X}_a, \mathbf{X}_b) &= \sum_{i=1}^M \left\{ \left[\left(\frac{\Sigma_i^{(a)} + \Sigma_i^{(u)}}{2} \right)^{-\frac{1}{2}} (\mathbf{m}_i^{(a)} - \mathbf{m}_i^{(u)}) \right]^T \right. \\ &\quad \left. \times \left[\left(\frac{\Sigma_i^{(b)} + \Sigma_i^{(u)}}{2} \right)^{-\frac{1}{2}} (\mathbf{m}_i^{(b)} - \mathbf{m}_i^{(u)}) \right] \right\} \end{aligned} \quad (16)$$

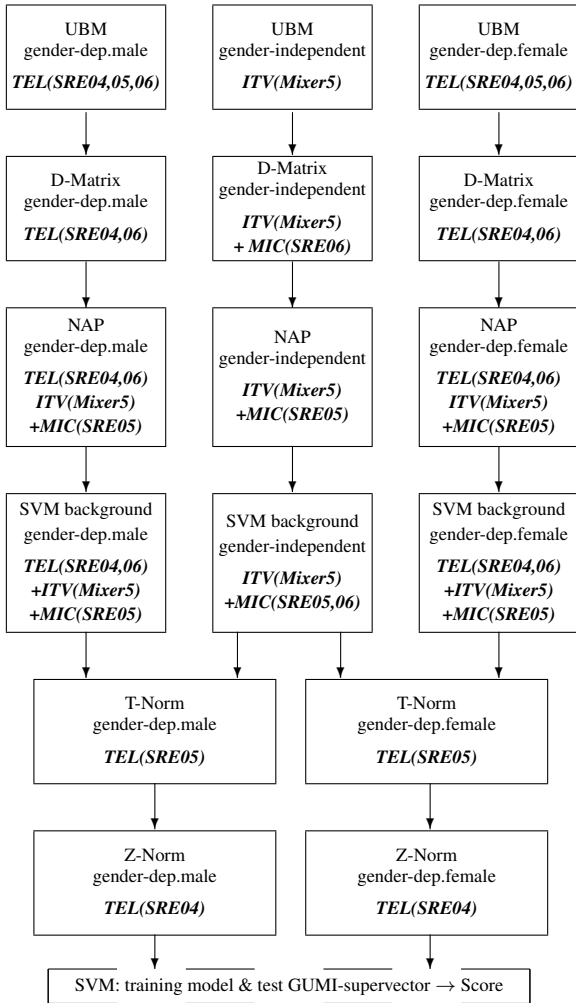


Figure 1: The proposed hybrid data assignment for GMM-SVM speaker recognition on SRE 2008 core task. In the lower part of each block, we indicate in italics the data set used at each stage.

3.3. Hybrid Strategy

In our study, we assume that different channel corresponds to different quality and quantity of the speech database available for model training and we also suppose that a test utterance for recognition is recorded through an unknown channel, which exists among the enrolled channels. Our task is to design a robust speaker recognition system against the unknown channels, e.g. telephony, interview and microphone channels.

It is observed that the gender-dependent model is of better performance than the gender-independent model when there are sufficient amount of data. On the other hand, for insufficient amount of training data, gender-independent model is found to be more suitable.

We propose a hybrid strategy in consideration of the variation of the training and testing conditions corresponding to the channels. In particular, we use gender-dependent model for high signal-to-noise-ratio (SNR) and sufficient amount of training data set. This is to assure that the high performance does not drop down due to the great amount of mixture with the weak signal. However, we adopt gender-independent model for low SNR and insufficient amount of training data set, in a hope that the gender-independent training can overcome the problem

caused by the insufficient amount of data.

Fig. 1 illustrates the hybrid strategy developed for SRE 2008 core task, where telephone data set has sufficient amount of speech data with high SNR, while the interview data set has insufficient amount of speech data with low SNR, and both quantity and quality of the microphone data set are in between the two. The advantage of the hybrid strategy of the gender-dependent and gender-independent models is also contributed by the disparate data groups for training. It enhances the complementary of the models.

4. Performance Evaluation

4.1. Evaluation Condition on NIST SRE 2008

In the evaluation, 13-dimensional PLP coefficients, after voice activity detection (VAD), with their delta and double delta coefficients form the 39-dimension PLP feature. We investigate the performance of the proposed speaker recognition system on the NIST SRE 2008 core task. Actually, there are five categories of trial in the SRE 2008 short2-short3 core task, namely interview-to-interview (itv-itv), interview-to-telephone (itv-tel), telephone-to-telephone (tel-tel), telephone-to-microphone (tel-mic), and telephone-to-interview (tel-itv). In this paper, the purpose of our study is to design a robust system given an unknown-channel test utterance against the variation of the channels. There are three data groups corresponding to three different channels in the core-task. They are telephone group that comprises clean and strong speech signals, interview group that is of less data and weak noisy speech signal, and microphone data group that is in between the two.

4.2. Classifiers and their Performances

In our proposed Bhattacharyya-based system, we use 512 mixture components for either gender-dependent or gender-independent GMM. We trained the diagonal matrix D by using EM algorithm with the initial D being $D_i^{(0)} = (\Sigma_i^{(u)})^{-\frac{1}{2}}$. Although the adaptation of the RF in the (14) is only for the mean vectors i.e. $\tilde{r}_i^{(m)}$, we extend the same adaptive strategy to the adaptation of covariance matrix in (5), i.e. $\tilde{r}_i^{(\Sigma)} = \tilde{r}_i^{(m)}$. The value of θ_0 is set to 8.2×10^{-4} empirically².

Using the GUMI kernel, we evaluate the proposed hybrid GMM-SVM method named as **GUMI-Pro**. We compared the single GUMI system in either gender-dependent or gender-independent mode with fixed RF named as **GUMI-Dep** and **GUMI-Ind** respectively, and adaptive RF named as **GUMI-Dep-Ar** and **GUMI-Ind-Ar** respectively. We chose the Kullback-Leibler (KL) kernel [2] in the performance evaluation. For the fixed RF, we choose empirically the value of the RF to be 10 for the GMM-SVM system. Since the gender-dependent mode performs well, we only show the performance of the gender-dependent mode KL system named **KL-Dep**. All the above mentioned systems are implemented under the same platform, i.e. they share the same development databases of UBM, SVM background, nuisance attribute projection (NAP), T-Norm and Z-Norm. Besides the above-mentioned classifiers, we also involved the raw gender-dependent KL system which is without NAP or TZNorm.

Table 1 shows the equal error rate (EER) and minimum detection cost function (min DCF) values corresponding to the five categories respectively. Figs. 2 and 3 give the detection

²We obtained the θ_0 value by investigating the duration-histogram of the utterances in the training database and testing-development database and referring to the proper value of the fixed RF case.

Table 1: The comparison of the speaker recognition systems in terms of EER and minimum cost for SRE 2008 core task

EER (%)	itv-itv	itv-tel	tel-tel	tel-mic	tel-itv
KL-Dep-Raw	8.63	11.31	4.72	9.65	11.86
KL-Dep	4.56	8.09	3.36	8.08	7.66
GUMI-Dep	4.33	7.62	2.85	7.74	6.62
GUMI-Ind	3.38	7.85	4.32	7.88	7.65
GUMI-Pro	2.86	6.52	3.01	6.37	5.77
GUMI-Dep-Ar	3.27	5.90	2.52	6.36	4.42
GUMI-Ind-Ar	2.94	6.76	3.17	6.49	5.67
GUMI-Pro-Ar	2.52	5.56	2.68	5.61	4.30

minDCF ($\times 100$)	itv-itv	itv-tel	tel-tel	tel-mic	tel-itv
KL-Dep-Raw	4.26	4.87	2.17	3.76	5.27
KL-Dep	2.08	3.09	1.35	2.76	2.87
GUMI-Dep	2.02	2.91	1.29	2.47	2.76
GUMI-Ind	1.72	3.38	1.95	3.17	3.44
GUMI-Pro	1.39	2.53	1.33	2.25	2.37
GUMI-Dep-Ar	1.57	2.33	1.15	2.13	1.84
GUMI-Ind-Ar	1.45	2.74	1.46	2.43	2.55
GUMI-Pro-Ar	1.25	2.16	1.18	2.00	1.79

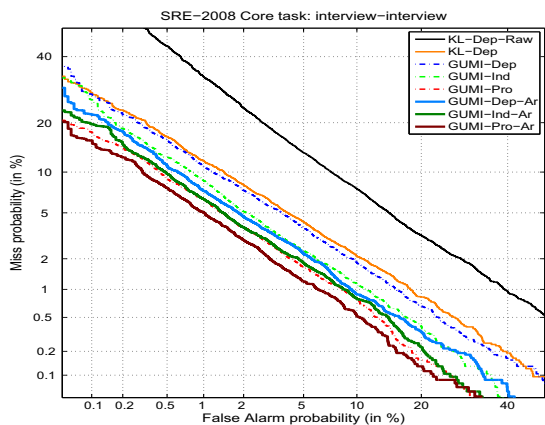


Figure 2: DET of the speaker recognition systems on NIST SRE 2008 interview-to-interview task

error trade-off (DET) curves for two of the five categories. It can be seen that the system **GUMI-Pro** is apparently better than **GUMI-Dep** and **GUMI-Ind** in most of the categories; it is observed that only telephone-to-telephone set is not improved by the proposed scheme, this is because the **GUMI-Dep** is specially designed for telephone-to-telephone subtask while **GUMI-Ind** is used for interview data. It is also noticed that the **GUMI-Pro-Ar** is obviously better than the **GUMI-Pro**, and the same findings are also noticed for gender-dependent and gender-independent set. It suggests that the adaptation is a right attempt. The hybrid strategy increases also the complementary factor from different modeling to raise the robustness of the entire system.

5. Summary

In this paper, we developed a robust GMM-SVM speaker recognition system to resolve the different problems in channel com-

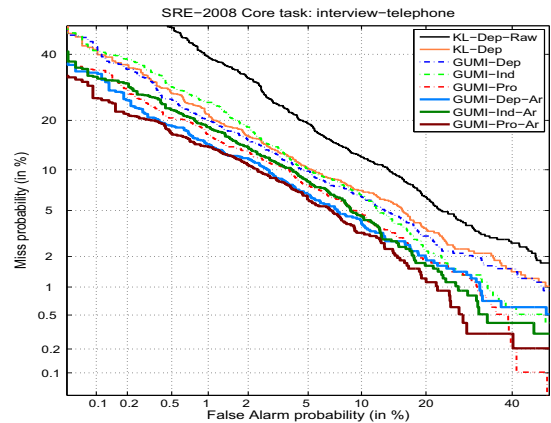


Figure 3: DET of the speaker recognition systems on NIST SRE 2008 interview-to-telephone task

penation. We introduced an adaptive RF approach to mitigate the negative effects to the speaker characteristics from the individual channel and utterance, especially the effect caused by the duration variability. In the proposed system, while NAP is used for inter-channel compensation, the adaptive RF can be viewed as the some channel-related and duration compensation. Moreover, we proposed the hybrid strategy that effectively fuses the different models that are complementary in improving the robustness of the entire system. Using GUMI kernel which carries the information from means and covariances of the GMM, we demonstrated the effectiveness of the proposed system on the NIST SRE 2008 core task.

6. References

- [1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19-41, 2000.
- [2] 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.
- [3] 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.
- [4] <http://www.itl.nist.gov/iad/mig/tests/sre/2008/>
- [5] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms," Montreal, CRIM, 2006.
- [6] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 291-298, 1994.
- [7] C. H. You, K. A. Lee and H. Li, "A GMM supervector kernel with the Bhattacharyya distance for SVM based speaker recognition", in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, 2009.