A GMM-SUPERVECTOR APPROACH TO LANGUAGE RECOGNITION WITH ADAPTIVE RELEVANCE FACTOR

Chang Huai You, Haizhou Li, Kong Aik Lee

Human Language Technology Department, Institute for Infocomm Research (I^{2}R), A*STAR
1 Fusionopolis Way, #08 Connexis (South Tower), Singapore 138632
phone: +65 64082763, fax: +65 64677601, emails: {echyou, hli, kalee}@i2r.a-star.edu.sg
http://hlt.i2r.a-star.edu.sg/

ABSTRACT

Gaussian mixture model (GMM) supervector has been proven effective for language recognition. While a speech utterance can be represented with a GMM which can be obtained through maximum *a posteriori* (MAP) criterion, it is observed that the supervector formed from the GMM encounters shifting problem in the supervector space due to varying duration of the utterance. We propose an adaptive relevance factor for the MAP estimation to mitigate the negative effect of the variability of individual utterances. Moreover, we develop a language recognition system with a Bhattacharyya-based kernel where the information from the mean vectors and covariance matrices are separately assigned into corresponding dissimilarities. We show the effectiveness of the proposed adaptive relevance factor and the Bhattacharyya-based kernel on the National Institute of Standards and Technology (NIST) language recognition evaluation (LRE) 2009 task.

1. INTRODUCTION

Language recognition is the process of recognizing the language of a spoken utterance. Common techniques used in language recognition include the acoustic and phonotactic modeling approaches. In this paper, we are interested in studying Gaussian mixture model, the most popular acoustic modeling approach for its reliable performance.

In GMM approach, a language model is obtained by maximum *a posteriori* (MAP) estimation from a universal background model (UBM) [1]. The UBM is usually trained through expectation-maximization (EM) algorithm from a background dataset covering a wide range of languages, speakers and channels. In this paper, we develop a language recognition system based on the Bhattacharyya-based kernel, which has been shown effective for speaker recognition [2]. Typical speech analysis generates sequences of feature vectors with variable length, while support vector machine (SVM) requires fixed-dimension inputs. The key aspect of applying SVM to speech is to provide an SVM kernel, which compares a sequence of feature vectors with others efficiently. Empirically we observed that the utterance-based supervector input to SVM is better than the fixed-length-segment-based supervector.

With MAP estimation, we adapt the GMM parameters according to the data available for adaptation. Due to variability of individual utterances, this approach may lead to the inconsistency among the training GMM supervectors, and thus result in the serious mismatch between training and testing supervectors. It is believed that such inconsistency can be compensated partly via data-driven adjustment of the relevance factor. In MAP estimation, if a mixture component has a low probabilistic count on new data, the sufficient statistics will be de-emphasized, otherwise will be emphasized. As a result, the displacement of GMM supervector from the UBM supervector varies undesirably due to the variability of new data available for adaptation, for example, the duration change of individual utterance. This is because the probabilistic count in MAP depends on the number of the involved feature vectors. The relevance factor is a way of controlling how much new data should be observed in a mixture before the new parameters begin replacing the old parameters.

In the GMM-UBM system, the relevance factor is less sensitive and therefore can be fixed. This is possibly due to the nature of generative modeling [3]. However, SVM works in a discriminative manner. In language recognition, a GMM supervector is used to represent the language property of an 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 language characteristics. Thus, we propose an adaptive scheme for the relevance factor that changes according to the amount of the feature data. Presently, language recognition based on acoustic model reaches state of the art performance using discriminative training techniques. In this paper, we modify the existing Bhattacharyya kernel to clearly separate the information assignment of the mean vectors and covariance matrices in different terms of the kernel. We evaluate the proposed scheme on the language detection task of the NIST LRE 2009 [4].

In the remainder of the paper, we introduce the conventional MAP estimation for language recognition in section 2. Then, we describe the proposed adaptive relevance factor and the Bhattacharyya-based language recognition in section 3. The performance evaluation is reported in section 4. We summarize the paper in section 5.

2. MAP FOR LANGUAGE RECOGNITION

Usually, with EM algorithm, the UBM is trained using a large dataset to form a language-independent model [3]. The selection of dataset has to consider different languages, channels and speakers. The UBM can be denoted as

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

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

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

© EURASIP, 2010 ISSN 2076-1465

where $\mathbf{m}_i, \Sigma_i, \omega_i, (i=1,...,M)$ are respectively the mean vector, the covariance matrix, and the weight of the ith 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,

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

where $\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2, \cdots,\ \mathbf{x}_\kappa]$ is the feature vectors used to train the GMM, λ; $\mathbf{x}$ is a J-dimensional feature vector; and κ is the number of feature vectors. The parameters of the ith Gaussian are adapted as follows [3],

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

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

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

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

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

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

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

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

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

3. GMM SUPERVECTOR WITH ADAPTIVE RELEVANCE FACTOR

3.1 An Adaptive Relevance Factor for MAP

Assume each language can be modeled by a language-independent mean supervector, and the distribution of the mean supervector is Gaussian, then it can be expressed in the following form:

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

where $\mathbf{z}(\lambda)$ is a hidden variable vector distributed according to normal distribution and it contains the language-dependent information; D_i is a diagonal matrix that can be trained and it is language-independent. From the above assumption, the relevance factor for mean (where $\rho \in \{m\}$) can be derived to be $r_i^{(m)} = D_i^{-2}\Sigma_i^{(\text{u})}, i=1,...,M$ [1].

The idea of MAP estimation for GMM was presented in [1]. The primary purpose of the MAP is to estimate the probability density function of a certain group of the data given a prior distribution. It is reasonable that for insufficient data the reliability is low so the value of α in (8) is small and the estimated GMM should be close to the UBM. When the data becomes sufficient, the reliability of the sufficient statistics is high so the value of α in (8) is large, so that the estimated GMM should be displaced further from the UBM. This is reflected by equations (4), (5) and (8). Thus, when applying MAP to derive GMM supervector, to assure the reliability of the estimated model, the GMM supervector should be close to the UBM supervector when the feature data is insufficient, and vice versa.

However, in language recognition, usually GMM supervector is purposely used to represent the language of the utterance. It is generated from the universal languages which is represented by the UBM supervector. This requires the distance from the universal language to the particular language does not vary with the length of the utterance. In other words, the GMM supervector is required to stably represent the characteristics of the particular language regardless of length of the utterance spoken. In short, ideally, each utterance with the same language is expected to give the same GMM supervector regardless of the duration of the utterance. In this way, the supervectors can stably represent the language without being affected by the duration of an utterance. Therefore, we propose an adaptive relevance factor as follows

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

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

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

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

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

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

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

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

[1] In [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.

3.2 Bhattacharyya-Based GMM-SVM Kernel

In our previous work [6, 7], we derived an Bhattacharyya-based distance between two GMMs as follows

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

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 only carries the first-order of the adaptation data information with the mean vectors and the covariance statistical dissimilarity carries the second-order of new data information with the covariance matrices. Usually, the first term can be applied solely; we can assume that the covariance is not adapted and only exploit the mean information in the equation. By combining the two terms in (15), we arrive at the following kernel in practice

$$
\begin{aligned}
&K_{\text{Bhatt}}(\mathbf{X}_a,\mathbf{X}_b)\\
&=\sum_{i=1}^{M}\Big\{\Big[\frac{1}{2}\Big(\Sigma_i^{(\mathrm{u})}\Big)^{-\frac{1}{2}}(\mathbf{m}_i^{(\mathrm{a})}-\mathbf{m}_i^{(\mathrm{u})})\Big]^T\Big[\frac{1}{2}\Big(\Sigma_i^{(\mathrm{u})}\Big)^{-\frac{1}{2}}(\mathbf{m}_i^{(\mathrm{b})}-\mathbf{m}_i^{(\mathrm{u})})\Big]\Big\}\\
&\quad+\sum_{i=1}^{M} tr\Big[\Big(\frac{\Sigma_i^{(\mathrm{a})}+\Sigma_i^{(\mathrm{u})}}{2}\Big)^{\frac{1}{2}}(\Sigma_i^{(\mathrm{a})})^{-\frac{1}{2}}\Big(\frac{\Sigma_i^{(\mathrm{b})}+\Sigma_i^{(\mathrm{u})}}{2}\Big)^{\frac{1}{2}}(\Sigma_i^{(\mathrm{b})})^{-\frac{1}{2}}\Big]
\end{aligned}
\tag{16}
$$

4. PERFORMANCE EVALUATION

4.1 Evaluation on NIST LRE 2009

In the evaluation, 56-dimensional shifted delta cepstrum (SDC) features formed from Mel-frequency cepstrum coefficient (MFCC) with 7-1-3-7 delta-shift pattern (refer to *N-d-P-k* parameters in [10]) plus 7 static cepstral is computed after voice activity detection (VAD). We investigate the performance of the proposed language recognition system through the NIST LRE 2009 task. There are 23 target languages used for this evaluation, namely, Amharic, Bosnian, Cantonese, Creole (Haitian), Croatian, Dari, American English, Indian English, Farsi, French, Georgian, Hausa, Hindi, Korean, Mandarin, Pashto, Portuguese, Russian, Spanish, Turkish, Ukrainian, Urdu, and Vietnamese. Unlike LRE-2007 which has only telephony speech databases, LRE-2009 has two catagories of data sources named conversational telephone speech (CTS) and Voice of America (VOA) narrowband speech. In this experiment, the CTS training database are collected from CallFriend, OHSU, LRE07 Train, OGI22 and SRE06; and the VOA data are mainly from the VOA3 database provided by NIST and LDC, VOA7 provided by NIST and VOA8 downloaded from the web-site.

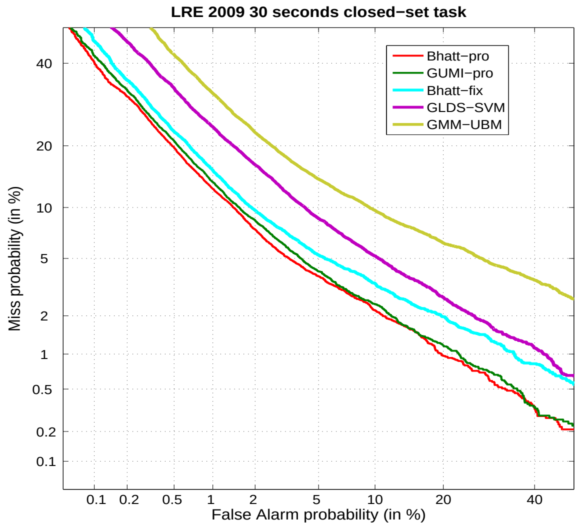

Figure 1: DET plot of the language recognition systems on NIST LRE 2009 30 seconds closed-set task

4.2 Core Classifiers

In our proposed Bhattacharyya-based system, we use 512 mixture components for GMM. We trained the diagonal matrix D by using EM algorithm with the initial D is $D_i^{(0)} = (\Sigma_i^{(\mathrm{u})})^{-\frac{1}{2}}$. Although the adaptation of the relevance factor in the (14) is only for mean vectors i.e. $\breve{r}_i^{(m)}$, we extend the same adaptation to variance matrix estimation in (5), i.e. $\breve{r}_i^{(\Sigma)}$=$\breve{r}_i^{(m)}$. The value of θ_0 is set to 8.2×10^{-4} via experiment [2]. In this evaluation, we developed the proposed Bhattacharyya-based system named as **Bhatt**. There are three sets of the implementation: **Bhatt-fix** is with the relevance factor being set to 16; **Bhatt-pro** is with the relevance factor being adapted by (14). In addition, for the performance comparison, we implement the following language recognition systems on the LRE 2009 task with the same training databases and feature processing. 1) **GLDS-SVM**: The GLDS-SVM system is based on the work reported in [8]. The feature vectors extracted from an utterance are expanded to a higher dimensional space by calculating all the monomials. All monomials up to order 3 are used, resulting in a feature space expansion from 56 to 32509 in dimension. 2) **GMM-UBM**: In the GMM-UBM system [3], firstly, 2048-mixture UBM was trained by using the universal background databases containing all of the target languages. The 2048-GMM of the target language is trained by using the training database. We used the Top-N (N is set to 20) method to reduce the computational complexity of the Gaussian probabilities. 3) **GUMI-pro**: A GMM-UBM mean interval (GUMI) GMM-SVM system [2] is developed with the adaptation of the relevance factor accroding to (11).

Outputs of individual classifiers are calibrated with separate linear backend followed by linear logistic regression (LLR). The calibrated scores are then combined via a final stage of LLR. Note that the development and evaluation data

[2] We obtained the θ_0 value by investigating the average length of the feature data and the best value of the fixed relevance factor.

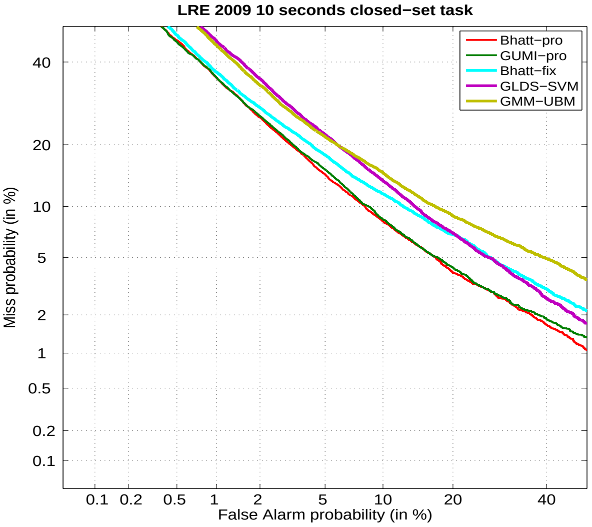

Figure 2: DET plot of the language recognition systems on NIST LRE 2009 10 seconds closed-set task

are grouped into 3-, 10- and 30-second utterances. We only conduct the experiments on 30- second utterances. Although the group of data are labelled under 30-second categories, the actual duration of utterances varies. The training is done on the 30-second development data using the FoCal Multiclass toolkit [11]. Log-likelihood ratio score from all classifiers are stacked together and a linear backend is trained. This is then followed by an LLR stage. The scores are converted into log-likelihood ratio for final decision with a threshold set at zero. A development set was designed for the training of backend, calibration. This development set consists of 8000 trials, and is built upon LRE07 augmented with additional trials taken from VOA3. The development set is split into two halves, one for training and the other one for cross-validation.

In this paper, we give the results of the closed-set task with nominal duration of 30 seconds and 10 seconds. Figs. 1 and 2 give the detection error trade-off (DET) plot of the LRE 2009 systems; while Tables 1 and 2 list the equal error rate (EER) and minimum detection cost function (min DCF) values corresponding to the Fig. 1 and Fig. 2 respectively. It can be seen that the system **Bhatt-pro** is apparently better than **Bhatt-fix**, it suggests that the adaptation is a right attempt. It is also noticed that the **Bhatt-pro** is better than the previous version of **GUMI-pro**, it means that the proposed **Bhatt** system is of something valuable over the conventional **GUMI** method.

5. SUMMARY

In this paper, we introduced an adaptive scheme for the relevance factor to mitigate the negative effects to the language characteristics from the individual utterance, especially the effect caused by the duration variability. We developed a Bhattacharyya-based GMM system for language recognition. In particular, there are two distance components named the mean and covariance statistical dissimilarities. We modified the mean statistical dissimilarity to carry the information from only the mean vectors so that each dissimilarity depends on a solely informative resource. This may bring more distinct informative sources for classification when we exploited both mean and covariance information. We demonstrated the effectiveness of the proposed system by using the LRE 2009 task.

Table 1: The comparison of the language recognition systems in terms of EER and minimum cost for LRE 2009 30s closed-set task

LRE 2009, 30s, closed-set	*EER*	*min. Cost × 100*
GMM-UBM	9.76 %	9.32
GLDS-SVM	6.88 %	6.81
Bhatt-fix	5.16 %	5.05
GUMI-pro	4.47 %	4.43
Bhatt-pro	4.21 %	4.14

Table 2: The comparison of the language recognition systems in terms of EER and minimum cost for LRE 2009 10s closed-set task

LRE 2009, 10s, closed-set	*EER*	*min. Cost × 100*
GMM-UBM	12.55 %	12.38
GLDS-SVM	11.79 %	11.77
Bhatt-fix	11.02 %	10.65
GUMI-pro	9.21 %	9.14
Bhatt-pro	9.04 %	8.98

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] http://www.itl.nist.gov/iad/mig/tests/lre/2009/

[5] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability : Theory and Algorithms," Montreal, CRIM, 2006.

[6] C. H. You, K. A. Lee and H. Li, "GMM-SVM Kernel with a Bhattacharyya-Based Distance for Speaker Recognition," accepted in IEEE Transactions on Audio, Speech and Language Processing.

[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.

[8] 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.

[9] 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.

[10] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds Language Recognition with Support Vector Machines. *Proc. Odyssey: The Speaker and Language Recognition Workshop* Toledo, Spain, ISCA, pp. 41C44, 31 May-3 June 2004.

[11] N. Brmmer,"FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores", available: http://niko.brummer.googlepages.com/focalmulticlass.