ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Effect of Relevance Factor of Maximum *a posteriori* Adaptation for GMM-SVM in Speaker and Language 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

Gaussian mixture model - support vector machine (GMM-SVM) with nuisance attribute projection (NAP) has been found to be effective and reliable for speaker and language recognition. In maximum *a posteriori* (MAP) adaptation of GMM, the relevance factor is the parameter that regulates how much the adaptation data affect the base model, which impacts the final recognition performance. In our previous work, the data-dependent relevance factor and adaptive relevance factor have been introduced. In this paper, we provide insights into different types of relevance factor for MAP in the context of application as formulated under Speaker Recognition Evaluation (SRE) and Language Recognition Evaluation (LRE) by the National Institute of Standards and Technology (NIST).

Index Terms: maximum *a posteriori*, supervector, Gaussian mixture model, support vector machine

1. Introduction

Gaussian mixture model (GMM) that relies on acoustic spectral features has shown reliable performance for text-independent speaker and language recognition [1]. Especially GMM-supervector with the application of support vector machine (SVM) achieves very effective performance [2]. In GMM approach, a model is obtained by maximum *a posteriori* (MAP) adaptation from a universal background model (UBM) [3]. A UBM is usually trained through expectation-maximization (EM) algorithm from a background data to cover a wide range of languages, speakers, sessions and channels.

In [4], joint factor analysis (JFA) was introduced to compensate the channel variation through eigenchannel modeling in GMM-supervector and to emphasize the speaker-dependent component by using low dimension speaker factor through eigenvoice modeling. However, in GMM-SVM system, eigenchannel is not as effective for channel compensation as nuisance attribute projection (NAP) [5]. It is observed that GMM-supervector system can work effectively with NAP without eigenchannel and eigenvoice analysis. It has been noticed that the relevance factor in MAP determines how much the observed training data influence the model adaptation, thus the resulting GMM model. In our previous work, the data-dependent relevance factor and adaptive relevance factor have been reported to be effective in both speaker and language recognition [6] [7]. In this paper, we study the different types of relevance factor and systematically investigate their performances in the same frameworks in GMM-SVM speaker and language recognition systems where NAP is applied. In particular, we use Bhattacharyya-based GMM-SVM as a classification platform to compare the effectiveness of different relevance factor in MAP on NIST Speaker Recognition Evaluation (SRE) 2008 and Language Recognition Evaluation (LRE) 2009 and 2011 tasks.

In the remainder of the paper, we present three types of relevance factor in MAP in Section 2. We describe the Bhattacharyya-based kernel for GMM-SVM speaker and language recognition in Section 3. The performance evaluation is reported in section 4. We summarize the paper in Section 5.

2. Three Types of Relevance Factor in MAP for GMM

An UBM can be denoted by the following set of parameters,

$$\mathrm{u} = \{\bar{\omega}_i, \bar{\mathbf{m}}_i, \bar{\Sigma}_i; i = 1, 2, ..., C\} \tag{1}$$

where C is the number of Gaussian components. The adapted GMM, λ, takes a similar form

$$\lambda = \{\omega_i, \mathbf{m}_i, \Sigma_i; i = 1, 2, ..., C\} \tag{2}$$

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

Adapting u to λ, we assume that the GMM parameters are subject to some prior distribution. With the MAP criterion, λ is selected such that it maximizes the posterior probability.

2.1. Fixed Relevance Factor

In conventional MAP, λ is obtained by

$$\breve{\lambda} = \arg\max_{\lambda} P(\lambda|\mathbf{X}) = \arg\max_{\lambda} \left[f(\mathbf{X}|\lambda)g(\lambda)\right] \tag{3}$$

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; and κ is the number of feature vectors.

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 is then the multiplicative combination of g_1 and g_2. We have the mean and covariance parameters of the ith Gaussian adapted as [3] [1],

$$\mathbf{m}_i = \alpha_i \breve{\Xi}_i + (1 - \alpha_i)\bar{\mathbf{m}}_i \tag{4}$$

$$\Sigma_i = \alpha_i \breve{\mathbf{S}}_i + (1 - \alpha_i)\left[\bar{\Sigma}_i + \bar{\mathbf{m}}_i\bar{\mathbf{m}}_i^*\right] - \mathbf{m}_i\mathbf{m}_i^* \tag{5}$$

where $\breve{\Xi}_i$ and $\breve{\mathbf{S}}_i$ are respectively the first and second order sufficient statistics; α_i are the adaptation coefficients given by [1]

$$\alpha_i = \frac{N_i}{N_i + \gamma_i} \tag{6}$$

[1] In fact, only mean vector satisfies (6) [3].

where the relevance factor γ_i is a constant parameter in the normal-Wishart density with which the Gaussian parameters are modeled [3]; N_i is the occupation count given by

$$N_i = \sum_{t=1}^{\kappa} \rho_i(\mathbf{x}_t) \tag{7}$$

and $\rho_i(\mathbf{x}_t) = \omega_i \aleph(\mathbf{x}_t; \mathbf{m}_i, \Sigma_i) \big[\sum_{j=1}^{C} \omega_j \aleph(\mathbf{x}_t; \mathbf{m}_j, \Sigma_j)\big]^{-1}$ where $\aleph$ denotes Gaussian probability.

2.2. Data-dependent Relevance Factor

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 (or language-dependent) supervector $\Phi\mathbf{z}(\lambda)$:

$$\mathbf{m}(\lambda) = \bar{\mathbf{m}} + \Phi\mathbf{z}(\lambda) \tag{8}$$

where 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 the 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 $\beta = \zeta^{-1}(\lambda)\Phi^*\Sigma^{-1}N(\breve{\Xi} - \bar{\mathbf{m}})$. Substituting the result to (8), we arrive at

$$\hat{\mathbf{m}} = \bar{\mathbf{m}} + \Phi\breve{\mathbf{z}} = \breve{\alpha}\breve{\Xi} + \bar{\mathbf{m}}(1 - \breve{\alpha}) \tag{9}$$

where $\breve{\alpha} = (\Phi^{-2}\Sigma + N)^{-1}N$. As compared to the conventional MAP of Eq. (4), Eq. (9) shows that $\breve{\alpha}$ is the adaptation coefficient. Therefore, the relevance factor can be given by

$$\breve{\gamma} = \Phi^{-2}\Sigma \tag{10}$$

The relevance factor in (10) is data dependent since the parameter Φ is obtained with EM algorithm based on a training dataset as follows.

The M-step for Φ

$$\Phi = N(\breve{\Xi} - \bar{\mathbf{m}})\mathbf{E}\big[\mathbf{z}^*(\lambda)\big]\Big(\mathbf{E}\big[\mathbf{z}(\lambda)\mathbf{z}^*(\lambda)\big]N\Big)^{-1} \tag{11}$$

the E-step:

$$\mathbf{E}\{\mathbf{z}(\lambda)\} = [\mathbf{I} + \Phi^*\Sigma^{-1}N\Phi]^{-1}\Phi^*\Sigma^{-1}N(\breve{\Xi} - \bar{\mathbf{m}}) \tag{12}$$

$$\mathbf{E}\{\mathbf{z}^*(\lambda)\mathbf{z}(\lambda)\} = [\mathbf{I} + \Phi^*\Sigma^{-1}N\Phi]^{-1} + \mathbf{E}\{\mathbf{z}(\lambda)\}^2 \tag{13}$$

2.3. Adaptive Relevance Factor

Since SVM is a discriminative classifier, the supervector used to represent a certain speaker or language is required to be relatively stable without being affected by the duration variation of utterance.

In [7], we introduce the idea of adaptive relevance factor in which we feed in additional term to (10) so that it could adapt to the duration variation from one utterance to the other, as follows

$$\tilde{\gamma} = \theta_0\kappa\Phi^{-2}\Sigma \tag{14}$$

where κ denotes the duration of the utterance, and θ_0 is a constant which is determined empirically based on a given database. This adaptation of duration ensures that the point in supervector space is not seriously drifted by the duration variation of the corresponding utterance.

3. Speaker and Language Recognition

In this study, we use the following Bhattacharyya-based kernel referred to as the GMM-UBM Mean Interval (**GUMI**) in [8] is used to evaluate the performance of the various relevance factors.

$$\begin{aligned} &K_{\text{GUMI}}(\mathbf{X}_a, \mathbf{X}_b) \\ &= \sum_{i=1}^{C} \Big\{\Big[\Big(\frac{\Sigma_i^{(a)} + \bar{\Sigma}_i}{2}\Big)^{-\frac{1}{2}}(\mathbf{m}_i^{(a)} - \bar{\mathbf{m}}_i)\Big]^T \\ &\qquad \times \Big[\Big(\frac{\Sigma_i^{(b)} + \bar{\Sigma}_i}{2}\Big)^{-\frac{1}{2}}(\mathbf{m}_i^{(b)} - \bar{\mathbf{m}}_i)\Big]\Big\} \end{aligned} \tag{15}$$

where $\mathbf{X}_a$ and $\mathbf{X}_b$ denote the feature vector sequences a and b respectively. As an SVM kernel, the **GUMI** kernel carries the information from both mean and covariance. Besides the **GUMI** kernel of (15), we also use another GMM-SVM system with the Bhattacharyya-based kernel by considering only the mean statistical information. It is actually a simplified **GUMI** kernel where the covariance of the GMM is not updated. We name the kernel as **Bhatt**$_m$.

Let K be any of the above mentioned kernel, the discriminant score of the SVM is given by

$$f(\mathbf{X}) = \sum_{l=1}^{L} \alpha_l y_l K(\mathbf{X}_l, \mathbf{X}) + b \tag{16}$$

where L is the number of support vectors, and $\mathbf{X}_l$ is a feature vector sequence corresponding to 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.

4. Performance evaluation

The default value of θ_0 is set to 8.2×10^{-4}, the value is approximated by investigating the average length of the feature data and a reference value of fixed relevance factor. In this paper, the relevance factor with fixed value is denoted by **fix-RF**, the one with data dependence as given in Eq. (10) denoted by **dep-RF** and the adaptive relevance factor as given in Eq. (14) named as **adp-RF**. We carry out the study under the NIST SRE and LRE framework [10] [11] [12]. In all GMM-SVM systems, NAP is always applied on the supervector for channel compensation. The performance is measured in terms of equal error rate (EER) and minimum detection cost function (minDCF).

4.1. Relevance Factor in Speaker Recognition

We compared the different types of relevance factor under the speaker recognition framework using the **GUMI** kernel with 512 mixture components. First we study a series of fixed value within a reasonable range of relevance factor. For ease of comparison and intuitional convenience, we place the γ value of **fix-RF** and the constant θ_0 of **adp-RF** on the same axis. To do this, we introduce a parameter δ by taking $\gamma = 2^{(\delta-5)}$ for the fixed relevance factor and at the same time $\theta_0 = 2^{(\delta-5)} \times 8.2 \times 10^{-4}$, so that we can relate the two parameters γ of **fix-RF** and θ_0 of **adp-RF**, which are independent of each other on the same axis based on δ as listed in Table 1. We may ask whether there is an optimal value of the fixed relevance factor for MAP. If the optimal value exists, is the optimal value always the same regardless of change of training-test condition?

In NIST SRE 2008 short2-short3 task [10], we consider five training-test conditions: telephone-telephone (tel-tel), telephone-microphone (tel-mic), telephone-interview (tel-itv), interview-telephone (itv-tel) and interview-interview (itv-itv).

Table 1: The corresponding relationship of δ-γ and δ-θ_0 in our experiments

δ	1	2	3	4	5	6	7	8	9	10	11	12
$\log_2\gamma$	-4	-3	-2	-1	0	1	2	3	4	5	6	7
γ	0.0625	0.125	0.250	0.5	1.0	2.0	4.0	8.0	16.0	32.0	64.0	128.0
$\theta_0 \times 10^3$	0.0625	0.1025	0.205	0.41	0.82	1.64	3.28	6.56	13.12	26.24	52.48	104.96

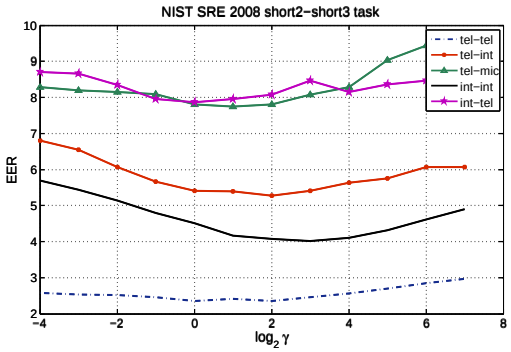

Figure 1: Investigation of the fixed relevance factor based on the performances of the GMM-NAP-SVM systems with different training-test conditions in terms of EER on NIST SRE 2008 short2-short3 task.

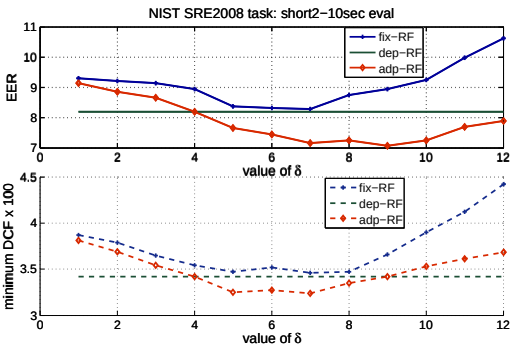

Figure 2: The performance of different types of relevance factor in terms of EER and minDCF on NIST SRE 2008 short2-10sec task.

The experimental results illustrated in Fig. 1 shows the investigation on the performances of the five different conditions with the fixed relevance factor. We can see that the best result is obtained on the point of $\log_2\gamma = 1$, *i.e.* $\gamma = 2$, while the itv-itv has its lowest EER at the point of $\log_2\gamma = 3$, *i.e.* $\gamma = 8$. The other situation also give the similar observation. It comes to a conclusion that for each particular speaker recognition with GMM-SVM in all of the five conditions, **adp-RF** is the best among three types, **dep-RF** are always better than **fix-RF**. However, the best value varies with different training-test condition. Fig. 2 shows the experimental result in terms of EER and minDCF for short2-10sec task. The best value of **fix-RF** is at the point of $\gamma = 4$, with EER=8.28% and minDCF = 0.0346. The EER and minDCF with **dep-RF** are respectively 8.18% and 0.0342. It also shows the best value of θ_0 for **adp-RF** is at $\theta_0 = 3.28 \times 10^{-3}$ where the EER and minDCF are 7.15% and 0.0324 respectively. The above results show that **dep-RF** is better than **fix-RF**, while **adp-RF** is the best among the three.

Table 2: The comparison of different types of relevance factor based on the **GUMI** kernel in terms of EER on SRE 2008 short2-short3 task, where $\delta = 7$.

SRE 2008: EER	*itv-itv*	*itv-tel*	*tel-tel*	*tel-mic*	*tel-itv*
GUMI:fix-RF	4.07%	8.08%	2.35%	7.81%	5.28%
GUMI:dep-RF	4.06%	7.85%	2.24%	7.34%	4.33%
GUMI:adp-RF	3.97%	6.59%	2.09%	6.72%	3.93%

Table 3: The comparison of the language recognition systems in terms of EER and minDCF for LRE 2009 30s closed-set task. (**GUMI:fix-RF** is with the relevance factor being set to 16.)

LRE 2009, 30s	EER	minDCF $\times$ 100
GUMI:fix-RF	5.41 %	5.21
GUMI:dep-RF	5.13 %	4.94
GUMI:adp-RF	4.47 %	4.43

4.2. Relevance Factor in Language Recognition

4.2.1. Language recognition

We choose LRE 2009 core task [10] as language recognition platform to investigate the effect of different types of relevance factor. We use 56-dimensional MFCC-SDC features with 7-1-3-7 delta-shift (corresponding to N-d-P-k parameters in [8]) plus 7 static cepstral computed after voice activity detection (VAD). In LRE 2009, there are 23 target languages. The **GUMI** kernel with 512 mixture components for GMM is adopted.

Tables 3 and 4 list the EER and minDCF for 30- and 10-second tasks, respectively. It can be seen that **GUMI:adp-RF** is apparently better than **GUMI:dep-RF**, which implies that the duration consideration is a right attempt. On the 30-second closed-set task of LRE 2009, with **GUMI:dep-RF**, the relative EER improvement of 5.17% with respect to **GUMI:fix-RF** is obtained. Similar improvement can also be seen on the 10-second closed-set task of LRE 2009 as well as those with minDCF measurement.

4.2.2. Language-pair recognition

In language pair recognition, we use LRE 2011 pair task [12] as an evaluation platform to study the effect of different types of relevance factor. We adopt NIST LRE 2011 30-second task to evaluate the performance of our language-pair recognition system. The number of target languages in LRE 2011 is $\eta = 24$, thus there are $\eta(\eta - 1)/2 = 276$ language pairs. As a result, each test segment is used for 276 trials. The system's overall performance measure is based on the top η target language pairs

Table 2: The comparison of different types of relevance factor based on the **GUMI** kernel in terms of EER on SRE 2008 short2-short3 task, where $\delta = 7$.

SRE 2008: EER	*itv-itv*	*itv-tel*	*tel-tel*	*tel-mic*	*tel-itv*
GUMI:fix-RF	4.07%	8.08%	2.35%	7.81%	5.28%
GUMI:dep-RF	4.06%	7.85%	2.24%	7.34%	4.33%
GUMI:adp-RF	3.97%	6.59%	2.09%	6.72%	3.93%

Table 3: The comparison of the language recognition systems in terms of EER and minDCF for LRE 2009 30s closed-set task. (**GUMI:fix-RF** is with the relevance factor being set to 16.)

LRE 2009, 30s	EER	minDCF $\times$ 100
GUMI:fix-RF	5.41 %	5.21
GUMI:dep-RF	5.13 %	4.94
GUMI:adp-RF	4.47 %	4.43

The EERs listed in Table 2 are for NIST SRE 2008 short2-short3 task. In all of the five conditions, **dep-RF** is always better than **fix-RF**, and **adp-RF** is the best among three types.

4.2. Relevance Factor in Language Recognition

4.2.1. Language recognition

We choose LRE 2009 core task [10] as language recognition platform to investigate the effect of different types of relevance factor. We use 56-dimensional MFCC-SDC features with 7-1-3-7 delta-shift (corresponding to N-d-P-k parameters in [8]) plus 7 static cepstral computed after voice activity detection (VAD). In LRE 2009, there are 23 target languages. The **GUMI** kernel with 512 mixture components for GMM is adopted.

Tables 3 and 4 list the EER and minDCF for 30- and 10-second tasks, respectively. It can be seen that **GUMI:adp-RF** is apparently better than **GUMI:dep-RF**, which implies that the duration consideration is a right attempt. On the 30-second closed-set task of LRE 2009, with **GUMI:dep-RF**, the relative EER improvement of 5.17% with respect to **GUMI:fix-RF** is obtained. Similar improvement can also be seen on the 10-second closed-set task of LRE 2009 as well as those with minDCF measurement.

4.2.2. Language-pair recognition

In language pair recognition, we use LRE 2011 pair task [12] as an evaluation platform to study the effect of different types of relevance factor. We adopt NIST LRE 2011 30-second task to evaluate the performance of our language-pair recognition system. The number of target languages in LRE 2011 is $\eta = 24$, thus there are $\eta(\eta - 1)/2 = 276$ language pairs. As a result,

Table 4: The comparison of the language recognition systems in terms of EER and minDCF for LRE 2009 10s closed-set task. (**GUMI:fix-RF** is with the relevance factor being set to 16.)

LRE 2009, 10s	EER	minDCF × 100
GUMI:**fix-RF**	11.02 %	10.80
GUMI:**dep-RF**	10.65 %	10.37
GUMI:**adp-RF**	9.21 %	9.14

Table 4: The comparison of the language recognition systems in terms of EER and minDCF for LRE 2009 10s closed-set task. (**GUMI:fix-RF** is with the relevance factor being set to 16.)

Table 5: The comparison of the language pair recognition systems in terms of η-top average EER and minDCF for LRE 2011 30s task

LRE 2011, η-top average	EER	minDCF ×100
Bhatt$_m$:**fix-RF**=0.25	13.83 %	13.14
Bhatt$_m$:**fix-RF**=4	13.79 %	12.92
Bhatt$_m$:**fix-RF**=8	13.64 %	12.69
Bhatt$_m$:**fix-RF**=32	14.65 %	14.08
Bhatt$_m$:**dep-RF**	13.05 %	12.39
Bhatt$_m$:**adp-RF**	12.75 %	12.07

Table 5: The comparison of the language pair recognition systems in terms of η-top average EER and minDCF for LRE 2011 30s task

LRE 2011, η-top average	EER	minDCF ×100
Bhatt$_m$:**fix-RF**=0.25	13.83 %	13.14
Bhatt$_m$:**fix-RF**=4	13.79 %	12.92
Bhatt$_m$:**fix-RF**=8	13.64 %	12.69
Bhatt$_m$:**fix-RF**=32	14.65 %	14.08
Bhatt$_m$:**dep-RF**	13.05 %	12.39
Bhatt$_m$:**adp-RF**	12.75 %	12.07

each test segment is used for 270 trials. The system's overall performance measure is based on the top η target language pairs for which the minDCF operating points for 30-second segments are greatest. MFCC-SDC with 80 dimensionality obtained with configuration of 10-2-3-7 is used. The GMM consists of 1024 Gaussian mixtures.

In this investigation, we use the **Bhatt**$_m$ kernel to classify the language pair. The SVM models are trained by using one-to-one pair-modeling strategy. The EER and minDCF listed in Table 5 give a comparison among the three types of relevance factor. It can be seen that the adaptive relevance factor gives the best performance among the three. The above observation is also apparent from Fig. 3.

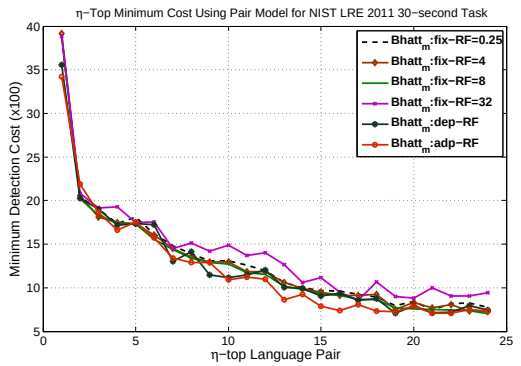

Figure 3: Comparison the different relevance factors on a top-η reference language pairs in terms of minDCF using the **Bhatt**$_m$ classifier.

5. Summary and Discussion

We used different Bhattacharyya-based kernels to build the GMM-supervector classification system to investigate the effect of different types of relevance factors in terms of the recognition accuracy. In GMM-SVM system with NAP for speaker and language recognition, we come to a conclusion that there is a best point for the fixed relevance factor; the best point of fixed relevance factor is changeable subject to the different training-test application; and the data-dependent relevance factor outperforms the fixed relevance factor; the adaptive relevance factor performs the best in the three types of relevance factor.

The conventional MAP algorithm is derived by maximizing the posterior probability of the GMM model given a speech sequence where the distribution of the Gaussian parameters are assumed to be Dirichlet and normal-Wishart distributed respectively; as a result the relevance factor is a fixed value. The supervector analysis assumes the prior probability of $\mathbf{z}(\lambda)$ be a normal distribution, it leads the relevance factor to be optimized by the training data. This is done by maximizing the posterior probability of the parameter $\mathbf{z}(\lambda)$ which characterizes the mean vectors of the GMM model.

In NIST SRE 2008 short2-short3 task, training and test utterance have the same nominal duration; while in short2-10sec task the training and test have different nominal durations. However, the utterances selected for SVM background supervectors have different duration. This is the reason why even for short2-short3 task **adp-RF** performs better than **dep-RF** does.

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] 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.

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

[5] D. Matrouf, N. Scheffer1, B. Fauve, and J.-F. Bonastre, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification," *INTERSPEECH 2007*, pp. 1242-1245, Antwerp, Belgium, Aug. 2007.

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

[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] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Int. Conf. on Spoken Lang. Process.*, pp. 89-92, 2002.

[9] http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf

[10] http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf

[11] http://nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf

5. Summary and Discussion

We used different Bhattacharyya-based kernels to build the GMM-supervector classification system to investigate the effect of different types of relevance factors in terms of the recognition accuracy. In GMM-SVM system with NAP for speaker and language recognition, we come to a conclusion that there is a best point for the fixed relevance factor; the best point of fixed relevance factor is changeable subject to the different training-test application; and the data-dependent relevance factor outperforms the fixed relevance factor; the adaptive relevance factor performs the best in the three types of relevance factor.

The conventional MAP algorithm is derived by maximizing the posterior probability of the GMM model given a speech sequence where the distribution of the Gaussian parameters are assumed to be Dirichlet and normal-Wishart distributed respectively; as a result the relevance factor is a fixed value. The supervector analysis assumes the prior probability of $\mathbf{z}(\lambda)$ be a normal distribution, it leads the relevance factor to be optimized by the training data. This is done by maximizing the posterior probability of the parameter $\mathbf{z}(\lambda)$ which characterizes the mean vectors of the GMM model.

In NIST SRE 2008 short2-short3 task, training and test utterance have the same nominal duration; while in short2-10sec task the training and test have different nominal durations. However, the utterances selected for SVM background supervectors have different duration. This is the reason why even for short2-short3 task **adp-RF** performs better than **dep-RF** does.

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] 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.

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

[5] D. Matrouf, N. Scheffer1, B. Fauve, and J.-F. Bonastre, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification," *INTERSPEECH 2007*, pp. 1242-1245, Antwerp, Belgium, Aug. 2007.

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

[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] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Int. Conf. on Spoken Lang. Process.*, pp. 89-92, 2002.

[9] http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf

[10] http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf

[11] http://nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf