Bayesian Speaker Recognition Using Gaussian Mixture Model and Laplace Approximation

Shih-Sian Cheng[1], *I-Fan Chen*[2], *Hsin-Min Wang*[1 2]

[1]Institute of Information Science, Academica Sinica, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academica Sinica, Taipei, Taiwan
sscheng@iis.sinica.edu.tw, ifanchen@gmail.com, whm@iis.sinica.edu.tw

Abstract

This paper presents a Bayesian approach for Gaussian mixture model (GMM)-based speaker identification. Some approaches evaluate the speaker score of a test speech utterance using a single data likelihood over the GMM learned by point estimation methods according to the maximum likelihood or maximum a posteriori criteria. In contrast, the Bayesian approach evaluates the score by using the expectation of the data likelihood over the posterior distribution of the model parameters, which is depicted by Bayesian integration. However, as the integration can not be derived analytically, we apply Laplace approximation to the derivations. Theoretically, we show that the proposed Bayesian approach is equivalent to the GMM-UBM approach when infinite training data is available for each speaker. The results of speaker identification experiments on the TIMIT corpus show that the proposed Bayesian approach consistently outperforms the GMM-UBM approach under very limited training data conditions, although the improvement is not very significant.

Index Terms: speaker identification, speaker recognition, Bayesian inference, GMM-UBM

1. Introduction

Speaker recognition, a natural and convenient way to authenticate a person's identity, involves two major tasks: identification and verification [1]. In automatic speaker identification (ASI), which is the focus of this paper, the recognition system outputs the speaker's identity for a given test speech utterance; whereas in automatic speaker verification (ASV), the claimed speaker identity of a test speech utterance is verified by the system. There is an increasing need to develop another technique of speaker characterization called speaker diarization [2], which attempts to group together speech segments produced by the same speaker within an audio stream. Different from the supervised nature of ASI and ASV, speaker diarization can be viewed as an unsupervised speaker recognition task.

The Gaussian mixture model (GMM) has been widely used in statistical speaker recognition [3, 4, 5, 6, 7, 8, 9]. In [3] and [4], the authors first successfully applied GMM to the speaker recognition task, where each speaker GMM was trained with the maximum likelihood (ML) criterion based on the associated training data. Subsequently, they applied GMM to ASV, where the speaker GMM was adapted from a universal background model (UBM) with the maximum a posteriori (MAP) criterion to alleviate the overfitting issue inherent in the ML estimation method [5]. This approach, which is known as the "GMM-UBM" method, has been one of the leading approaches for both ASI and ASV. In addition, GMM-based speaker identification has also been applied to speaker clustering in some speaker diarization systems [10, 11].

Although GMM-based ASI using the ML or MAP training criterion performs reasonably well, the generalization ability of these two criteria may still be limited due to the intrinsic nature of point estimation. Instead of evaluating the speaker score of the test speech utterance over an ML- or an MAP-derived model, we can evaluate it in a fully Bayesian manner. In other words, we can marginalize the score of the test speech utterance over the posterior distribution of the model parameters with the given training data [12]. However, the marginalized score can not be evaluated analytically when GMM is applied as the speaker model. Therefore, approximation methods, such as the variational inference [12, 13], are desired.

In this paper, we use *Laplace approximation* to evaluate the marginalized score. We show that, theoretically, the proposed Bayesian ASI approach is equivalent to the GMM-UBM approach when infinite training data is available for each speaker. Moreover, with the Bayesian framework, we can consider the GMM-UBM approach from a fully Bayesian perspective. The results of ASI experiments on the TIMIT corpus [14] show that the proposed Bayesian approach consistently outperforms the GMM-UBM approach under very limited training data conditions.

The remainder of this paper is organized as follows. In Section 2, we describe the Bayesian framework for ASI. Then, we present our implementations for Bayesian ASI in Section 3. The experiment results are detailed in Section 4. We then summarize our conclusions in Section 5.

2. Bayesian speaker identification framework

Suppose we have K target speakers $S_1, \cdots, S_K$. The enrollment training data for speaker S_i is denoted as $\hat{\mathcal{S}}_i = \{\mathbf{s}_i^1, \cdots, \mathbf{s}_i^{n_i}\}$, where n_i is the number of training samples. Given the test speech utterance $\mathcal{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_N\}$ with N samples, ASI can be performed by evaluating the log speaker posteriors over the test data and choosing the speaker with the largest posterior score as follows:

$$\begin{aligned} S_{output} &= \arg\max_{\hat{\mathcal{S}}_i} \ln p(\hat{\mathcal{S}}_i|\mathcal{O}) \\ &= \arg\max_{\hat{\mathcal{S}}_i} \ln \frac{p(\hat{\mathcal{S}}_i)p(\mathcal{O}|\hat{\mathcal{S}}_i)}{p(\mathcal{O})} \\ &= \arg\max_{\hat{\mathcal{S}}_i} [\ln p(\hat{\mathcal{S}}_i) + \ln p(\mathcal{O}|\hat{\mathcal{S}}_i)]. \end{aligned} \quad (1)$$

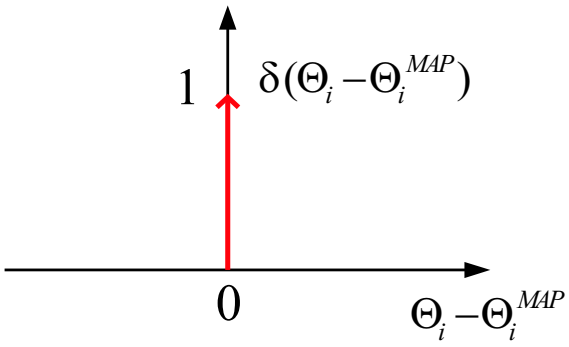

Figure 1: The Dirac delta function of $\mathbf{\Theta}_i - \mathbf{\Theta}_i^{MAP}$.

Here, we consider the case where the prior importance of each speaker is the same, i.e., $p(\hat{\mathcal{S}}_i) = 1/K$ for $i = 1, \cdots, K$. Then, the ASI task involves evaluating the log *speaker evidence* $p(\mathcal{O}|\hat{\mathcal{S}}_i)$ for each speaker.

To infer the speaker evidence using the Bayesian scenario, a parametric model $\mathbf{\Theta}_i$[1] for $\hat{\mathcal{S}}_i$ is introduced first. Then, considering $\mathbf{\Theta}_i$ as a latent variable, the speaker evidence can be expressed as a marginal likelihood as

$$p(\mathcal{O}|\hat{\mathcal{S}}_i) = \int p(\mathcal{O}|\mathbf{\Theta}_i) p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i) d\mathbf{\Theta}_i, \tag{2}$$

which is the expectation of $p(\mathcal{O}|\mathbf{\Theta}_i)$ over $p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i)$. For the case where the posterior distribution $p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i)$ is a Dirac delta function of $\mathbf{\Theta}_i - \mathbf{\Theta}_i^{MAP}$, as shown in Figure 1, $p(\mathcal{O}|\hat{\mathcal{S}}_i)$ is equivalent to $p(\mathcal{O}|\mathbf{\Theta}_i^{MAP})$. Thus, we may expect that if a more general, reasonable distribution rather than a delta function is applied to $p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i)$, the ASI performance obtained with the speaker evidence computed by Eq. (2) will be better than that obtained with the MAP-derived likelihood.

3. Implementation of Bayesian ASI using GMM and Laplace approximation

First, we briefly review Laplace approximation (LA) in Section 3.1, and then describe our implementation of Bayesian ASI in Section 3.2.

3.1. Laplace approximation

As shown in Figure 2, Laplace approximation (LA) finds a Gaussian approximation $q(\mathbf{z})$ for a probability density function $f(\mathbf{z})$ [12]. The first step of LA involves finding one of $f(\mathbf{z})$'s modes. Suppose the mode found is $\mathbf{z_0}$, the resultant Gaussian distribution is

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{z_0}, \mathbf{A}^{-1}), \tag{3}$$

where

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z_0}}. \tag{4}$$

3.2. Implementation of Bayesian ASI

We apply a GMM with G mixture components to parameterize speaker S_i's training data $\hat{\mathcal{S}}_i$; and denote its parameter set as $\mathbf{\Theta}_i = \{\mathbf{\Theta}_{i_1}, \cdots, \mathbf{\Theta}_{i_G}\}$, where $\mathbf{\Theta}_{i_g} = \{w_{i_g}, \mu_{i_g}, \mathbf{\Sigma}_{i_g}\}$ and w_{i_g}, μ_{i_g}, and $\mathbf{\Sigma}_{i_g}$ are, respectively, the weight, mean vector, and covariance matrix of the gth Gaussian component.

[1] $\mathbf{\Theta}_i$ denotes the parameter set of the model.

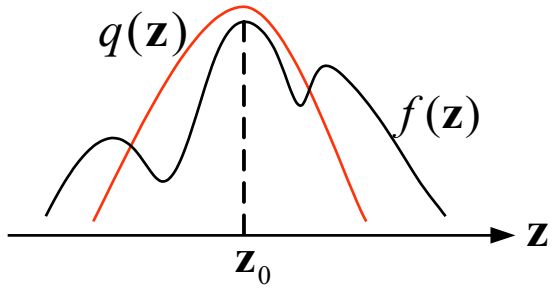

Figure 2: Laplace approximation finds a Gaussian approximation $q(\mathbf{z})$ for the probability density function $f(\mathbf{z})$.

Suppose the test data samples are i.i.d., then, we obtain

$$p(\mathcal{O}|\hat{\mathcal{S}}_i) = \prod_{j=1}^{N} p(\mathbf{o}_j|\hat{\mathcal{S}}_i), \tag{5}$$

where

$$p(\mathbf{o}_j|\hat{\mathcal{S}}_i) = \int p(\mathbf{o}_j|\mathbf{\Theta}_i) p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i) d\mathbf{\Theta}_i. \tag{6}$$

However, the calculation of the sample-based speaker evidence in Eq. (6) is not analytically feasible even if LA is used to Gaussianize $p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i)$, unless the following assumptions are made:

A1. Independence between the posteriors of Gaussian components:

$$p(\mathbf{\Theta}_i|\hat{\mathcal{S}}_i) = \prod_{g=1}^{G} p(w_{i_g}, \mu_{i_g}, \mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i). \tag{7}$$

A2. Independence between the weight and the Gaussian parameters for a mixture component:

$$p(w_{i_g}, \mu_{i_g}, \mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i) = p(w_{i_g}|\hat{\mathcal{S}}_i) p(\mu_{i_g}, \mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i), \tag{8}$$

where

$$p(\mu_{i_g}, \mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i) = p(\mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i) p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}). \tag{9}$$

A3. $p(w_{i_g}|\hat{\mathcal{S}}_i)$ is a Dirac delta function of $w_{i_g} - w_{i_g}^c$ and $p(\mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i)$ is a Dirac delta function of $\mathbf{\Sigma}_{i_g} - \mathbf{\Sigma}_{i_g}^c$, where $w_{i_g}^c$ and $\mathbf{\Sigma}_{i_g}^c$ are fixed points that can be obtained by point estimation using the ML or MAP method.

A4.

$$p(\mu_{i_g}|\mathbf{\Sigma}_{i_g}^c) \equiv \mathcal{N}(\mu_{i_g}|\mu_{i_g}^{prior}, \mathbf{\Sigma}_{i_g}^c). \tag{10}$$

By applying A1 and A2 to Eq. (6), we have

$$\begin{aligned} p(\mathbf{o}_j|\hat{\mathcal{S}}_i) &= \sum_{g=1}^{G} \int w_{i_g} \mathcal{N}(\mathbf{o}_j|\mu_{i_g}, \mathbf{\Sigma}_{i_g}) \\ &\qquad p(w_{i_g}, \mu_{i_g}, \mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i) d\mathbf{\Theta}_{i_g} \\ &= \sum_{g=1}^{G} \int \mathcal{N}(\mathbf{o}_j|\mu_{i_g}, \mathbf{\Sigma}_{i_g}) p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}) p(\mathbf{\Sigma}_{i_g}|\hat{\mathcal{S}}_i) \\ &\qquad w_{i_g} p(w_{i_g}|\hat{\mathcal{S}}_i) d\mathbf{\Theta}_{i_g}. \end{aligned} \tag{11}$$

Then, by applying A3 to Eq. (11), we have

$$p(\mathbf{o}_j|\hat{\mathcal{S}}_i) = \sum_{g=1}^{G} w_{i_g}^c \int \mathcal{N}(\mathbf{o}_j|\mu_{i_g}, \mathbf{\Sigma}_{i_g}^c) p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c) d\mu_{i_g}. \quad (12)$$

Eq. (12) shows that only the uncertainty of Gaussian mean vectors over the speaker training data is considered. Since $\mu_{i_g}^{MAP}$ is a mode of $p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c)$, we use LA to approximate $p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c)$ and obtain

$$p(\mathbf{o}_j|\hat{\mathcal{S}}_i) \simeq \sum_{g=1}^{G} \int w_{i_g}^c \mathcal{N}(\mathbf{o}_j|\mu_{i_g}, \mathbf{\Sigma}_{i_g}^c) \mathcal{N}(\mu_{i_g}|\mu_{i_g}^{MAP}, \mathbf{B}_{i_g}^{-1}) d\mu_{i_g}, \quad (13)$$

where

$$\mathbf{B}_{i_g} = -\nabla\nabla \ln p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c)|_{\mu_{i_g}=\mu_{i_g}^{MAP}}. \quad (14)$$

Moreover, since the convolution of two independent Gaussians is another Gaussian, Eq. (13) can be rewritten as

$$p(\mathbf{o}_j|\hat{\mathcal{S}}_i) \simeq \sum_{g=1}^{G} w_{i_g}^c \mathcal{N}(\mathbf{o}_j|\mu_{i_g}^{MAP}, \mathbf{\Sigma}_{i_g}^c + \mathbf{B}_{i_g}^{-1}). \quad (15)$$

By using the fact that

$$p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c) = \frac{p(\mu_{i_g}|\mathbf{\Sigma}_{i_g}^c) p(\hat{\mathcal{S}}_i|\mu_{i_g}, \mathbf{\Sigma}_{i_g}^c)}{p(\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c)}, \quad (16)$$

we obtain

$$\mathbf{B}_{i_g} = -\nabla\nabla \ln p(\mu_{i_g}|\mathbf{\Sigma}_{i_g}^c)|_{\mu_{i_g}=\mu_{i_g}^{MAP}} - \nabla\nabla \ln p(\hat{\mathcal{S}}_i|\mu_{i_g}, \mathbf{\Sigma}_{i_g}^c)|_{\mu_{i_g}=\mu_{i_g}^{MAP}}. \quad (17)$$

Then, by assuming that the training data samples are i.i.d. and applying A4 to Eq. (17), we obtain

$$\begin{aligned}\mathbf{B}_{i_g} &= -\nabla\nabla \ln \mathcal{N}(\mu_{i_g}|\mu_{i_g}^{prior}, \mathbf{\Sigma}_{i_g}^c)|_{\mu_{i_g}=\mu_{i_g}^{MAP}} \\ &\quad -\nabla\nabla \sum_{k=1}^{n_i} \ln \mathcal{N}(\mathbf{s}_i^k|\mu_{i_g}, \mathbf{\Sigma}_{i_g}^c)|_{\mu_{i_g}=\mu_{i_g}^{MAP}} \\ &= (1+n_i)(\mathbf{\Sigma}_{i_g}^c)^{-1}. \end{aligned} \quad (18)$$

From Eq. (18), we observe that when $n_i \to \infty$, $p(\mu_{i_g}|\hat{\mathcal{S}}_i, \mathbf{\Sigma}_{i_g}^c) \simeq \mathcal{N}(\mu_{i_g}|\mu_{i_g}^{MAP}, \mathbf{B}_{i_g}^{-1})$ becomes a delta function and the uncertainty vanishes. By applying Eq. (18) to Eq. (15), we finally obtain

$$p(\mathbf{o}_j|\hat{\mathcal{S}}_i) \simeq \sum_{g=1}^{G} w_{i_g}^c \mathcal{N}(\mathbf{o}_j|\mu_{i_g}^{MAP}, (1+\frac{1}{1+n_i})\mathbf{\Sigma}_{i_g}^c). \quad (19)$$

In summary, the proposed Bayesian approach performs ASI by

$$\begin{aligned} S_{output} &= \arg\max_{\hat{\mathcal{S}}_i} \ln p(\mathcal{O}|\hat{\mathcal{S}}_i) \\ &\simeq \arg\max_{\hat{\mathcal{S}}_i} \sum_{j=1}^{N} \ln p(\mathbf{o}_j|\hat{\mathcal{S}}_i), \end{aligned} \quad (20)$$

where the computation of $p(\mathbf{o}_j|\hat{\mathcal{S}}_i)$ is depicted in Eq. (19).

3.2.1. Estimation of the parameters

In order to calculate $p(\mathbf{o}_j|\hat{\mathcal{S}}_i)$ in Eq. (19), we need to estimate $w_{i_g}^c$, $\mu_{i_g}^{MAP}$, and $\mathbf{\Sigma}_{i_g}^c$ for $g = 1, \cdots, G$. In this paper, we apply the MAP estimation method to derive these parameters from the UBM using $\hat{\mathcal{S}}_i$ as the training data. When infinite training samples are available for each speaker, i.e., $n_i \to \infty$, it is clear from Eq. (19) that the Bayesian ASI approach in Eq. (20) is equivalent to the GMM-UBM approach. In other words, we can consider the GMM-UBM approach from a fully Bayesian perspective.

4. Experiments

4.1. Experiment setup

We conducted speaker identification experiments on the TIMIT acoustic-phonetic continuous speech corpus [14], in which each speaker has 10 utterances. The regions of silence at the beginning and the end of an utterance were removed according to the label files in the experiments. The average length of the utterances after silent region removal was 2.69 seconds. Each utterance was converted into a stream of 19-dimensional feature vectors, each consisting of 19 Mel-frequency cepstrum coefficients extracted using a 32-ms Hamming-windowed frame with 10-ms shifts [15].

We used the standard TIMIT test set, which contained 1680 utterances of 168 speakers, to train the UBM. The UBM was a GMM with 512 Gaussian components, each with a *diagonal* covariance matrix. The speaker identification experiments were conducted on the standard TIMIT training set, which contained 4620 utterances of 462 speakers.

We used the GMM-UBM approach as our baseline, where each speaker GMM was learned using MAP estimation with the UBM as the prior [5]. Only the Gaussian mean vectors were updated, while the weights and covariance matrices were set to the corresponding parameters of the UBM. The same strategy was applied in the proposed Bayesian approach. Therefore, for the speaker evidence calculation in Eq. (19), $\mu_{i_g}^{MAP}$ was learned using MAP estimation with the UBM as the prior, while $w_{i_g}^c$ and $\mathbf{\Sigma}_{i_g}^c$ were set to the corresponding weight and covariance matrix of the UBM.

4.2. Experiment results

We performed 10-fold cross validation to evaluate the system performance. In each fold, for each speaker, one utterance was used as the training data, and the remaining 9 utterances were used as the test data. To evaluate the effect of the training data size, we used the first 0.5, 1.0, 1.5, and 2.0 seconds of the training utterance and the whole utterance as the speaker training data.

The experiment results are shown in Table 1. The second and third columns show the identification accuracy of the GMM-UBM approach and the proposed Bayesian approach respectively; while the last column shows the significance of the performance difference between these two approaches evaluated by McNemar's Test [16]. From the table, it is clear that the Bayesian approach consistently outperforms the GMM-UBM approach when the training data size is small; however the difference becomes insignificant as the amount of training data increases. The results conform to our claim in Section 3.2.1 that the Bayesian approach is equivalent to the GMM-UBM approach when infinite training samples are available for each speaker. In addition, although there is no significant difference

Table 1: Speaker identification accuracy of the GMM-UBM approach and the proposed Bayesian approach on the TIMIT corpus under different configurations of training data sizes.

Training data size	GMM-UBM (%)	Bayesian (%)	Significance
0.5 sec	30.6558	30.6831	72.00 %
1.0 sec	50.1878	50.2046	67.30 %
1.5 sec	63.5377	63.5642	88.12 %
2.0 sec	70.8560	70.8537	0.00 %
1 utterance	74.4012	74.4204	53.46 %

between the two approaches when the training data size is two seconds, the Bayesian approach still outperforms the GMM-UBM approach when the whole training utterance is used to train the speaker model. In this case, the training data size is longer than two seconds and varies among the speakers. This might imply that the Bayesian approach can compensate better than the GMM-UBM approach in the case of unbalanced training data sizes for different speakers, although further investigation is necessary.

The results of the GMM-UBM approach and the proposed Bayesian approach in Table 1 were obtained based on adapting the Gaussian mean vectors from the UBM. We have also evaluated the case where the Gaussian covariance matrices are also updated, i.e., $\mathbf{\Sigma}_{ig}^{c}$ in Eq. (19) is substituted by $\mathbf{\Sigma}_{ig}^{MAP}$. For this case, the performance of both the GMM-UBM and Bayesian approaches is slightly better than that obtained by adapting the Gaussian mean vectors only. Since the results conform to many research results on applying the GMM-UBM approach in the speaker identification task, for simplicity, they are not reported in this paper.

5. Conclusion

In this paper, we have proposed a GMM-based Bayesian approach for speaker identification, where Laplace approximation is applied to the Bayesian integration. To evaluate the speaker score, the proposed approach considers the uncertainty of the model parameters rather than relies on learning a single model via point estimation. Theoretically, when infinite training data is available for each speaker, the proposed approach is equivalent to the GMM-UBM approach because uncertainty about the model parameters disappears; on the other hand, based on the proposed approach, we can consider GMM-UBM from a fully Bayesian perspective. The results of speaker identification experiments on the TIMIT corpus show that the proposed Bayesian approach consistently outperforms the GMM-UBM approach under very limited training data conditions, although the improvement is not very significant.

6. Acknowledgements

This work was supported in part by the National Science Council of Taiwan under Grant: NSC96-2628-E-001-024-MY3.

7. References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings Of The IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[4] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

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

[6] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Open-set speaker identification using adapted Gaussian mixture models," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1997–2000.

[7] A. Stergiou, A. Pnevmatikakis, and L. C. Polymenakos, "Enhancing the performance of a gmm-based speaker identification system in a multi-microphone setup," in *Proc. Interspeech*, Pittsburgh, Pennsylvania, Sep. 2009, pp. 1463–1466.

[8] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 4529–4532.

[9] Z. Lei, "UBM-based sequence kernel for speaker recognition," in *Proc. Interspeech*, Brighton, UK, Sep. 2009, pp. 1279–1282.

[10] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[11] S.-S. Cheng, H.-M. Wang, and H.-C. Fu, "BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 141–157, 2010.

[12] C. H. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[13] T. Ito, K. Hashimoto, Y. Nankaku, A. Lee, and K. Tokuda, "Speaker recognition based on variational Bayesian method," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1417–1420.

[14] TIMIT, *TIMIT acoustic-phonetic continuous speech corpus*, http://www.ldc.upenn.edu.

[15] J. Gonzalez-rodriguez, J. Fierrez-aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Hong Kong, China, Apr. 2003, pp. II–93–II–96.

[16] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Glasgow,UK, 1989, pp. 532–535.