Text-Independent Speaker Verification Using Variational Gaussian Mixture Model

Mohammad Hossein Moattar and Mohammad Mehdi Homayounpour

This paper concerns robust and reliable speaker model training for text-independent speaker verification. The baseline speaker modeling approach is the Gaussian mixture model (GMM). In text-independent speaker verification, the amount of speech data may be different for speakers. However, we still wish the modeling approach to perform equally well for all speakers. Besides, the modeling technique must be least vulnerable against unseen data. A traditional approach for GMM training is expectation maximization (EM) method, which is known for its overfitting problem and its weakness in handling insufficient training data. To tackle these problems, variational approximation is proposed. Variational approaches are known to be robust against overtraining and data insufficiency. We evaluated the proposed approach on two different databases, namely KING and TFarsdat. The experiments show that the proposed approach improves the performance on TFarsdat and KING databases by 0.56% and 4.81%, respectively. Also, the experiments show that the variationally optimized GMM is more robust against noise and the verification error rate in noisy environments for TFarsdat dataset decreases by 1.52%.

Keywords: Gaussian mixture model, expectation maximization, variational approximation, speaker verification.

Manuscript received Nov. 15, 2010; revised May 1, 2011; accepted May 16, 2011.

Mohammad Hossein Moattar (phone: +98 2164542722, email: moattar@aut.ac.ir) and Mohammad Mehdi Homayounpour (email: homayoun@aut.ac.ir) are with the Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran.

http://dx.doi.org/10.4218/etrij.11.0110.0684

I. Introduction

Speaker modeling is the main part of a speaker recognition system. The Gaussian mixture model (GMM) is the most common approach for speaker modeling in text-independent speaker recognition [1], [2]. A GMM is composed of a finite mixture of multivariate Gaussian components. The advantage of using a GMM as the likelihood function is that it is computationally inexpensive. Since a GMM is a generative approach for speaker modeling, it is vulnerable to data insufficiency and cannot adapt its complexity according to the data available.

For numerical and computational reasons, the covariance matrices of the GMM are usually diagonal, that is, variance vectors, which restricts the principal axes of the Gaussian ellipses in the direction of the coordinate axes. Estimating the parameters of a full-covariance GMM [3] requires, in general, much more training data and is computationally expensive.

Some previous works applied discriminative models such as artificial neural network (ANN) [4] or support vector machines (SVM) [5]. The main advantages of ANNs are their discriminative training power and flexible architecture that permits easy use of contextual information. Another potential advantage of ANNs is that feature extraction and speaker modeling can be combined into a single network, enabling joint optimization of the feature extractor and the speaker model [6]. The main disadvantage is that their structure has to be selected by trial and error procedures [7].

SVM is a popular method for speaker modeling specially in speaker verification. SVM classifiers are well suited to separate complex regions between two classes through an optimal, nonlinear decision boundary. SVM is a powerful discriminative classifier that has been adopted in speaker

recognition. It has been applied both with spectral [8], [9], prosodic [10], [11], and high-level features [12]. Currently, SVM is one of the most robust classifiers in speaker verification, and it has also been successfully combined with GMM to increase accuracy [10], [11]. One reason for the popularity of SVM is its good generalization performance to classify unseen data. The main problem of this model is its inappropriateness to handle the temporal structure of the speech.

Generally speaking, speaker modeling based on discriminative learning techniques can be tuned to obtain comparable performance to the state-of-the-art GMM, and in some specific conditions, they can outperform GMM [7]. Recent approaches have examined how to apply dynamic kernel SVMs to the speaker verification task [13], [14]. These approaches have generally been found to outperform traditional GMM-based approaches [15]-[17] and a variety of dynamic kernels have been successfully applied for speaker verification [15], [18]-[21].

This paper mainly focuses on the GMM approach as the most conventionally used speaker modeling and recognition method. There are various methods for training the GMMs, each with a different optimization criterion. For parameter estimation in GMMs, the most commonly used method is the maximum likelihood (ML) method. The ML algorithm estimates the model parameters and maximizes a likelihood function. The best known algorithm that finds maximum likelihood estimates in parametric models for incomplete data is the expectation maximization (EM) algorithm [22]-[24]. EM is an iterative two-step algorithm. The E-step calculates the conditional expectation of the complete data log likelihood given the observed data and parameter estimates. Then, the M-step finds the parameter estimates that maximize the complete data log likelihood from the E-step.

However, ML training algorithm suffers from overfitting if the model complexity is too high [25]. In that case, the model fits very well to the training data but lacks the generalization ability so that it does not tolerate noise or other deviations from the expected training data. As a consequence, the model cannot be used for making inferences about the new data. Moreover, algorithms such as EM may not converge properly due to an unsuitable initialization.

Overfitting can be reduced by adding a penalty term to the ML objective function [26]. The basic idea is that the penalty term becomes larger as the model complexity grows. Another way to control the overfitting is cross validation [26]. The original training data is divided into subsets and the model is trained using all subsets except the one which is saved for the evaluation. This can be repeated with partitioning the data to different training and validation sets. When comparing the models, that structure is chosen which gives the best performance for the validation data. The cross validation method is not suitable when the training dataset is small in size.

An alternative to ML optimization is to adapt a speaker-independent world model or universal background model (UBM) with the speech data from a specific speaker [27]. The background model represents speaker-independent distribution of the feature vectors. When enrolling a new speaker to the system, the parameters of the background model are adapted to the feature distribution of the new speaker. The adapted model is then used as the model of that speaker.

There are various adaptation methods such as maximum a posteriori (MAP) [27] or maximum likelihood linear regression (MLLR) [28]-[30]. Selection of the proper method depends on the amount of available training data [31], [32]. For enough enrollment utterances MAP approach is the most popular, while for scarce enrollment speech MLLR method has shown to be more effective.

A fundamental question is how to choose the optimal model complexity which minimizes the overfitting. In the case of using mixture models, we may ask for the optimal number of the mixture components. This question may be crucial in speaker modeling and recognition tasks such as text-independent speaker verification. In such systems, the duration of available training utterances for each speaker may be different. This may be due to the various speaking rates of speakers, various amount of usable speech of each speaker, or different number of recording sessions. A reliable modeling and training framework should have the possibility of tuning the model complexity to avoid overfitting. Other than general guidelines and experimentation, there is no objective measure to determine the right number of components in the GMM speaker model a priori.

This paper proposes a variational approximation approach [33], [34] for GMM speaker model training and optimization to achieve a better and more reliable performance in text-independent speaker verification. This method is called the variational expectation maximization (VEM) algorithm. The proposed approach aims to solve the overfitting issue of the traditional expectation maximization GMM optimization technique and resolve the data insufficiency problem in conditions where the available training data for the speakers is scarce or the duration of available speech data for speakers is different.

Using variational approach to GMMs helps to determine the optimal number of components in a systematic manner [35]. Variational techniques offer a framework for parameter estimation and model selection. Similar to the MAP technique [27], [36], variational methods consider parameter posterior probability distributions but unlike MAP, they are not point

estimation methods, but the whole model probability is evaluated. Variational estimation provides information about the model quality while training it.

The rest of this paper is organized as follows. Section II introduces the variational approximation framework. Section III explains the proposed variational GMM approach. The experimental setup, a brief explanation of the evaluation measures and databases, and the experimental results are presented in section IV. Finally, section V concludes this paper.

II. Variational Approximation

Variational methods are known as deterministic approximation optimization schemes in contrast with stochastic approximations such as sampling methods [37]. Given a set of observed variables x, hidden (unobserved) variables z, and parameters θ, Bayesian learning aims at optimizing the following marginal likelihood for x:

$$\ln p(x \mid z,\theta) = \ln \int\int p(x \mid z,\theta) dz d\theta. \quad (1)$$

From the Bayes rule, we have $p(x|z,\theta) = p(x,\theta,z)/p(z,\theta|x)$. By considering the log of both sides, it is possible to write $\ln p(x|z,\theta)=\ln p(x,\theta,z)-\ln p(z,\theta|x)$. Instead of integrating θ and z, w.r.t their true unknown pdf, an approximation called variational posterior, denoted as $q(z,\theta|x)$, is used. Taking expectation w.r.t. $q(z,\theta|x)$, we obtain

$$\begin{aligned} \ln p(x \mid z,\theta) &= \int q(z,\theta \mid x) \ln p(x,\theta,z) dz d\theta \\ &\quad - \int q(z,\theta \mid x) \ln p(z,\theta \mid x) dz d\theta. \end{aligned} \quad (2)$$

According to the variational framework, the above mentioned log likelihood can then be expressed as

$$\begin{aligned} \ln p(x \mid z,\theta) &= \int q(z,\theta \mid x) \ln\left(\frac{p(x,\theta,z)}{q(z,\theta \mid x)}\right) dz d\theta \\ &\quad - \int q(z,\theta \mid x) \ln\left(\frac{p(z,\theta \mid x)}{q(z,\theta \mid x)}\right) dz d\theta \\ &= \int q(z,\theta \mid x) \ln\left(\frac{p(x,\theta,z)}{q(z,\theta \mid x)}\right) dz d\theta \\ &\quad + KL(q(z,\theta \mid x) \mid p(z,\theta \mid x)) \\ &= L(q(z,\theta \mid x)) + KL(q(z,\theta \mid x) \mid p(z,\theta \mid x)), \end{aligned} \quad (3)$$

where $KL(q(z,\theta|x)|p(z,\theta|x))$ denotes the Kullback-Leibler (KL) distance between the variational posterior and the true posterior. The term $L(q(z,\theta|x))$ is often considered as negative free energy. Because the KL distance is always positive, $L(q)$ represents a lower bound on $\ln p(x|z,\theta)$, which means that $\ln p(x|z,\theta) \geq L(q)$. Variational learning aims at maximizing the lower bound, $L(q(z,\theta|x))$, using an EM-like algorithm [33].

Variational algorithms guarantee to provide a lower bound on the approximation error [38]-[40]. The models which are previously trained by variational learning include mixtures of Gaussians [33], [34], hidden Markov model (HMM) [41], [42], mixtures of factor analyzers [43], linear models [44], and mixtures of products of Dirichlet and multinomial distributions [45]-[47]. Also, variational algorithms have been applied in different applications including independent component analysis [48], audio clip classification [49], speech emotion recognition [50], audio indexing [51], joint factor analysis [52], speech recognition [53], voice activity detection [54], and speech enhancement [55]. Next section concerns the variational optimization of GMM model.

III. Variational Gaussian Mixture Model

Let us denote the observations by $X=\{x_1,\ldots,x_N\}$. GMM is a linear superposition of K Gaussian distribution in the form of $p(x) = \sum_{k=1}^{K} \pi_k Normal(x \mid \mu_k, \Sigma_k)$ in which μ_k and Σ_k are the parameters of the k-th Gaussian component and π_k are the mixing coefficients. Mixing coefficients follow the properties that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k$. Let us consider a K-dimensional binary random variable Z in which $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. Also, suppose that the marginal distribution over Z is specified in terms of the mixing coefficients, $p(z_k = 1) = \pi_k$. Then $p(Z) = \prod_{k=1}^{K} \pi_k^{z_k}$, and the conditional distribution of x given random variables Z is

$$p(x \mid Z) = \prod_{k=1}^{K} Normal(x \mid \mu_k, \Sigma_k)^{z_k}. \quad (4)$$

Following the above definition, there is a random variable z_k for each x_k. Random variables Z are latent variables because their value is not known beforehand and are different from parameters in that their number grows with the size of the observations, while the number of parameters does not depend on the size of training dataset. We can write down the conditional distribution of the observed data vectors given the latent variables and the parameters of the components as

$$p(X \mid Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} Normal(x_n \mid \mu_k, \Lambda_k^{-1})^{z_{nk}}, \quad (5)$$

where $\mu=\{\mu_k\}$ and $\Lambda=\{\Lambda_k\}$ are the mean vector and precision matrix of the corresponding Gaussian component, respectively. Literature proposes conjugate priors over the parameters for simplicity [35]. Therefore, Dirichlet distribution is defined as the prior over mixing coefficients π_k, and Gaussian-Wishart distribution is considered as the prior over the mean and

precision of each component:

$$p(\pi) = Dir(\pi \mid \alpha_0) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0 - 1}, \quad (6)$$

$$p(\mu, \Lambda) = p(\mu \mid \Lambda)p(\Lambda) = \prod_{k=1}^{K} Normal(\mu_k \mid m_0, (\beta_0\Lambda_k)^{-1}) Wishart(\Lambda_k \mid W_0, \upsilon_0), \quad (7)$$

where α, β, m, W, and υ are the parameters of the corresponding distributions and $C(\alpha_0)$ is the normalization constant in the Dirichlet distribution. We can consider a variational distribution which factorizes between the latent variables and the parameters:

$$q(z, \pi, \mu, \Lambda) = q(z)q(\pi, \mu, \Lambda). \quad (8)$$

Let us denote the model parameters by $\Theta=\{\pi,\mu,\Lambda\}$. Using the variational framework, the objective function to be maximized is the variational lower bound defined as

$$L(q(Z,\Theta \mid X)) = \int\int q(Z,\Theta \mid X)\log\left(\frac{p(X,Z \mid \Theta)}{q(Z,\Theta \mid X)}\right)d\Theta dZ = \int\int q(Z)q(\Theta)\log\left(\frac{p(X,Z \mid \Theta)}{q(Z)q(\Theta)}\right)d\Theta dZ. \quad (9)$$

The functional form of the factors $q(Z)$ and $q(\pi,\mu,\Lambda)$ can be obtained via VEM. VEM is an iterative method that consists of two steps, namely: i) variational expectation (VE) and ii) variational maximization (VM). In the VE-step, the posterior over latent variables are computed by solving $\partial L(q)/\partial q(Z)$. This results in the following equation [14]:

$$\log(q^*(Z)) \propto \int q(\Theta)\ln p(X,Z \mid \Theta)d\Theta. \quad (10)$$

Then the log of the optimized $q(Z)$ is given by [14]:

$$\ln(q^*(Z)) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + Const, \quad (11)$$

where

$$\ln\rho_{nk} = Expectation[\ln\pi_k] + \frac{1}{2}Expectation[\ln|\Lambda_k|] - \frac{d}{2}\ln(2\pi) - \frac{1}{2}Expectation_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]. \quad (12)$$

In (11), *Const* is a constant value and can be removed by normalizing ρ_{nk} values as

$$r_{nk} = \rho_{nk}\Big/\sum_{j=1}^{K}\rho_{nj}, \qquad \ln(q^*(Z)) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln r_{nk}, \quad (13)$$

where the quantities r_{nk} play the role of responsibilities and sum to unity: $\sum_k r_{nk} = 1$.

In the VM-step, the posterior distribution over the parameters is computed by solving $\partial L(q)/\partial q(\Theta)$ and therefore we have

$$\ln(q^*(\Theta)) \propto \ln(p(\Theta))\int q(Z)\ln p(X,Z \mid \Theta)dZ. \quad (14)$$

The parameters posterior is computed in two stages. First, using the result of the previous step (VE-step), the parameters are updated as

$$\bar{\pi}_k = \sum_{n=1}^{N} r_{nk}, \qquad \bar{\mu}_k = \frac{1}{\bar{\pi}_k}\sum_{n=1}^{N} r_{nk}x_n, \qquad \bar{\Sigma}_k = \frac{1}{\bar{\pi}_k}\sum_{n=1}^{N} r_{nk}(x_n-\bar{\mu}_k)(x_n-\bar{\mu}_k)^T. \quad (15)$$

The above values are analogous to the quantities evaluated in the ML-EM algorithm for GMM. Then, the posterior parameters are updated. Using the independence property of π_k against μ_k and Λ_k, the variational posterior distribution of parameters can be stated as

$$q(\pi, \mu, \Lambda) = q(\pi)\prod_{k=1}^{K} q(\mu_k, \Lambda_k). \quad (16)$$

The optimum for each log posterior distribution $\ln(q(\pi))$ and $\ln(q(\mu_k, \Lambda_k))$ can be given as follows [10]:

$$\ln q^*(\pi) = (\alpha_0 - 1)\sum_{k=1}^{K}\ln\pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln\pi_k + Const, \quad (17)$$

$$q^*(\mu_k, \Lambda_k) = Normal(\mu_k \mid m_k, (\beta_k\Lambda_k)^{-1}) Wishart(\Lambda_k \mid W_k, \upsilon_k). \quad (18)$$

Also, in the VM-step, the parameters are updated as

$$\alpha_k = \alpha_0 + \bar{\pi}_k, \quad \beta_k = \beta_0 + \bar{\pi}_k, \quad m_k = \frac{1}{\beta_k}(\beta_0 m_0 + \bar{\pi}_k\bar{\mu}_k), \quad W_k^{-1} = W_0^{-1} + \bar{\pi}_k\bar{\Sigma}_k + \frac{\beta_0\bar{\pi}_k}{\beta_0 + \bar{\pi}_k}(\bar{\mu}_k - m_0)(\bar{\mu}_k - m_0)^T, \quad \upsilon_k = \upsilon_0 + \bar{\pi}_k + 1. \quad (19)$$

The described variational EM learning algorithm proceeds by iterating between VE-step (13) and (15) and VM-step (19). To perform the VM-step, we need to calculate r_{nk}. These parameters are calculated with normalizing ρ_{nk} values using (13). To calculate ρ_{nk} values, we need to calculate the following expectation w.r.t. the variational distribution of the parameters:

$$Expectation_{\mu_k,\Lambda_k}[(E_n-\mu_k)^T\Lambda_k(E_n-\mu_k)] = d\beta_k^{-1}+\upsilon_k(E_n-m_k)^T W_k(E_n-m_k). \quad (20)$$

The final values of the posterior parameters are the results of the variational EM algorithm and can be used for making inference about new data values. Variational GMM is not affected by initial model choice because the model prunes extra degrees of freedom.

IV. Experiments

1. Evaluation Metrics

The performance measures which are used in this paper are the same as the metrics used in the 2010 NIST Speaker Recognition Evaluation plan [56]. Our primary speaker verification performance measures are the false alarm rate (FAR) and the miss detection rate (MDR). A false alarm occurs when an imposter speaker is accepted. Miss detection occurs when a genuine speaker is not recognized and rejected by the system. Taking inspiration from [56], we use a single cost for measuring speaker verification performance. For this purpose, the cost function is defined as a weighted sum of miss detection and false alarm probabilities:

$$C_{\mathrm{DET}} = P_{\mathrm{Miss|Target}} + \left(P_{\mathrm{FalseAlarm|NonTarget}} * (1-P_{\mathrm{Target}})\right), \quad (21)$$

where $P_{\mathrm{Miss|Target}}$ and $P_{\mathrm{FalseAlarm|Target}}$ are the MDR and FAR, respectively. Also P_{Target} is the a priori probability of the specified target speaker. The values of the above measures are between zero and 1.

The traditional equal error rate (EER) is also employed as another evaluation metric in our experiments. This metric is usually reported in percent. An accurate speaker verification approach tries to minimize the aforementioned measures.

2. Evaluation Databases

In the speaker verification experiments, we need three sets of speech data for each speaker. The first one which is used for training should be long enough to facilitate model training. The second set of speech files will be used as development data for decision threshold determination. The third set is an amount of short duration speech segments which are used as the test data in evaluations. In the following experiments, the duration of training speech may be different, but in all experiments, the duration of development and test utterances is equal to 3 s.

Our first evaluation database is the telephony Farsi speech database called TFarsdat [57]. This database includes 64 speakers and two speech files from each speaker which are collected in two different sessions. The first session includes read speech. This file is considerably longer than the second session and can be used as training and development speech. The other speech file which includes spontaneous speech is clipped to short duration utterances and is used as test set in our evaluations. A set of 50 speakers from this database is selected and used in our experiments.

The second speech database is the KING corpus [58], which contains recorded speech from 51 male speakers in two versions which differ in channel characteristics. For each speaker and channel, there are ten files corresponding to sessions of about 30 s to 60 s duration. KING is designed principally for closed set experiments in text-independent speaker identification or verification over telephone lines and high-quality microphone. Fifty speakers from this database are selected as the evaluation data. For each speaker, the telephone-quality utterances are used. In the evaluations, one session is used as training and development data, and the other sessions are used as test data.

3. Experimental Setup

In both evaluation databases, silence detection is performed on original speech signals using the approach proposed in [59]. Then, apart from the training speech, for each speaker, six 3 s utterances are extracted as development data and six 3 s utterances as test speech. In the experiments of this paper, the duration of the training data is 30 s, unless otherwise mentioned.

As speech features, 12-order mel-frequency cepstral coefficients (MFCCs) [60] are used. Features are extracted from 30-ms frames multiplied by a Hamming window. A 24-channel mel filter bank is used and the frames are shifted every 10 ms.

GMMs are constructed using 64 components with diagonal covariance matrix. The number of iterations in EM and also VEM optimization is limited to 50.

Score normalization is proved to be effective in speaker verification systems and helps to reduce the effect of session variations [2]. In the following experiments, we employed the well-known T-norm score normalization approach [61] which applies the mean and variance of the scores of the input utterance on all the registered speaker models to normalize the score of the utterance on the target speaker model. This normalization approach can be formulated as

$$Tnorm(Score(U|\,TargetModel)) = (Score(U|\,TargetModel)\text{-}MeanScore)/ScoresVar, \quad (22)$$

where $Score(U|\,TargetModel)$ is the likelihood of the input test utterance U on the speaker model $TargetModel$, and $MeanScore$ and $ScoresVar$ are respectively the mean and

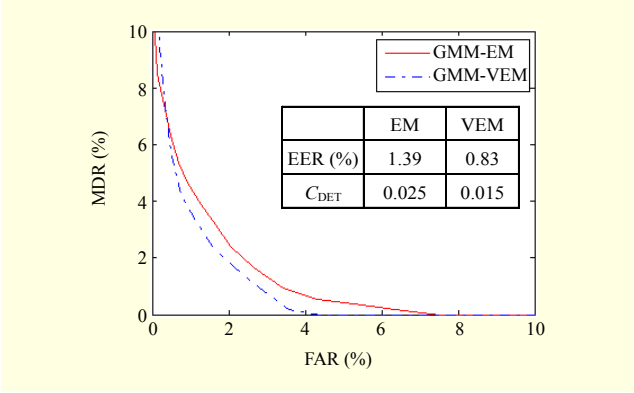


Fig. 1. MDR versus FAR of GMM-based speaker verification for EM and VEM optimization approaches on TFarsdat speech database. EER and C_{DET} of both approaches are summarized in Fig. 1.

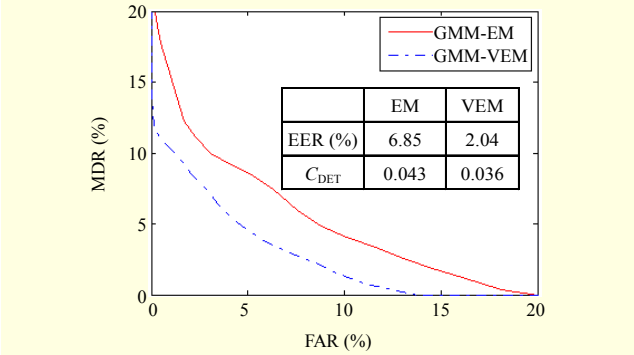


Fig. 2. MDR versus FAR of GMM-based speaker verification for EM and VEM optimization approaches on KING speech database. EER and C_{DET} of both approaches are summarized in Fig. 2.

variance of the likelihood scores of U on all registered speaker models except for *TargetModel*.

4. Speaker Verification Performance

In our first experiments, the performance of EM and VEM approaches is evaluated on the TFarsdat speech database. In these experiments, the number of iterations and the amount of speech data for all 50 speakers are the same. Figure 1 illustrates the detection error tradeoff (DET) of the mentioned approaches in which the MDR is plotted against the FAR Figure 1 also depicts the EER and the C_{DET} of these two approaches on the evaluations datasets.

Figure 1 shows that the performance of the proposed GMM-VEM approach is higher than the performance of traditionally used GMM-EM speaker modeling approach. As illustrated in Fig. 2, similar result is yielded on KING speech database.

Figure 2 illustrates that the improvement ratio on the KING database is higher than the TFarsdat dataset which is due to the higher verification error on the KING database. Apart from the verification performance, it is worth remembering that since the obtained models and even the likelihood computation algorithm have the same complexity in both approaches, the response time of the two methods is the same. Therefore, no significant overload is imposed on the verification framework when the EM approach is substituted with VEM.

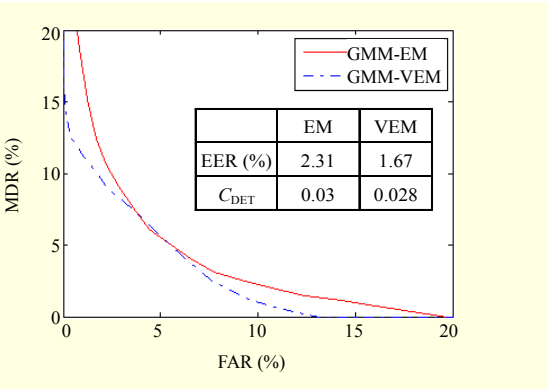


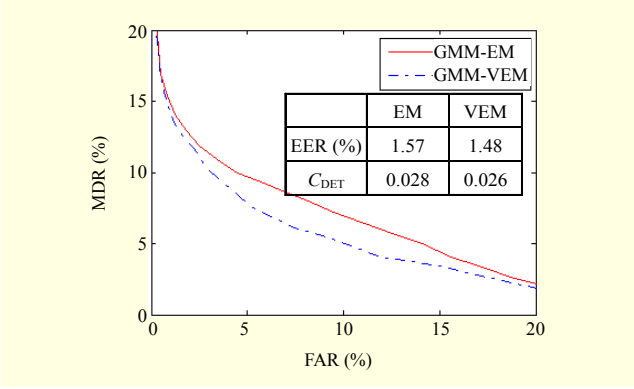
Fig. 3. MDR versus FAR of GMM-based speaker verification for EM and VEM optimization on TFarsdat speech database when training speech duration for each speaker is limited to 10 s. EER and C_{DET} of both approaches are summarized in Fig. 3.

5. Speaker Verification Performance in Adverse Conditions

The other experiments are conducted in adverse conditions which are probable in real world conditions and may occur in a specific application. Let's first consider a situation in which the amount of available training data is limited. In this condition, the available data may be very scarce to build a reliable speaker model.

To observe the impact of this data insufficiency on the proposed GMM-VEM approach and also the traditional GMM-EM approach, an experiment is conducted for a condition in which the duration of the training speech data for each speaker is limited to 10 s. Figure 3 shows the ROC curve of this evaluation on the TFarsdat database. Figure 3 shows the better performance of the GMM-VEM approach compared to the GMM-EM method in data insufficiency.

On the other hand, another experiment is conducted for the condition in which the available amount of speech data for speakers is different. This condition will happen in real world applications in which the number of registration sessions for speakers is different. A reliable modeling approach should be robust against this data variability. This robustness may originate from two characteristics of the model, namely robustness against data insufficiency and overfitting immunity.



	EM	VEM
EER (%)	1.57	1.48
C_{DET}	0.028	0.026

Fig. 4. MDR versus FAR of GMM-based speaker verification for EM and VEM optimization on TFarsdat speech database when training speech for half of speakers is 10 s and for other half is 30 s. EER and C_{DET} of both approaches are summarized in Fig. 4.

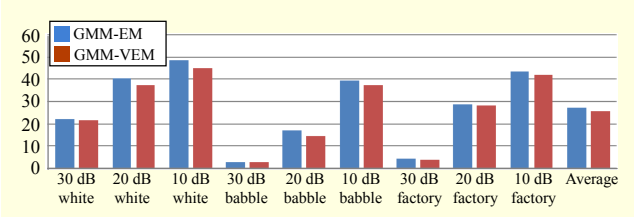


Fig. 5. Accuracy (denoted in EER (%)) of EM-GMM and VEM-GMM approaches in presence of different noises and SNR levels. Evaluation database is TFarsdat.

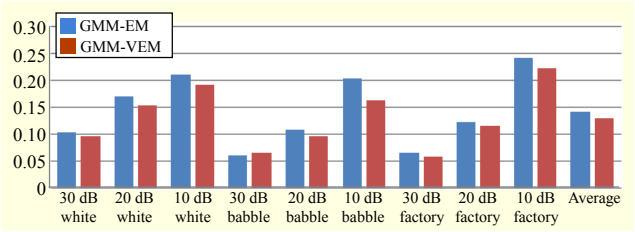


Fig. 6. Accuracy (denoted in detection cost (C_{DET})) of EM-GMM and VEM-GMM approaches in presence of different noises and SNR levels. Evaluation database is TFarsdat.

In the following experiment, the duration of training data for 25 of the speakers is 30 s, while for the other 25 speakers, the duration of the available training data is limited to 10 s. The results of this experiment is illustrated in Fig. 4 in terms of FAR and MDR.

Once again, Fig. 4 depicts the better robustness and higher performance of the proposed GMM-VEM approach in this abnormal operating condition. This can lead us to two conclusions. First, as concluded from the previous experiment (see Fig 3), the proposed optimization approach is robust against data insufficiency. Second, the proposed GMM-VEM modeling approach is relatively robust against overfitting which may occur in presence of different amount of training data. However, the second issue will also be justified in the next experiments from a different point of view.

6. Speaker Verification Performance in Noisy Conditions

To evaluate the proposed VEM optimization approach in noisy conditions and compare its performance with the GMM-EM approach, three different noise signals from the NOISEX-92 noise corpus [62], including White, Babble, and Factory noises, are synthetically added to the test utterances in different SNR levels, that is, 30 dB, 20 dB, and 10 dB. Figures 5 and 6 denote the verification accuracy of GMM-EM and GMM-VEM approaches in presence of noise. The accuracy measure in Fig. 5 is the EER (%) while the values in Fig. 6 denote the previously introduced detection cost (C_{DET}). The evaluation database is TFarsdat.

The average EER of the proposed approach in the presence of noise (averaged for all noises and all SNRs) is 25.6%, while this measure for the GMM-EM approach is 27.12%. The same improvement on the verification accuracy can be seen for average C_{DET} measure, which is 0.129 and 0.142 for the proposed method and the GMM-EM approach, respectively. Again, these achievements confirm that the proposed VEM approach not only provides better accuracy in clean speech condition but also improves the verification accuracy in presence of noise due to its resistance against overfitting.

7. Overfitting Investigation

Overfitting happens when the model is highly adapted to the training data and its generalization ability in confrontation with unseen data is decreased. This reduction in generalization ability can be depicted in the likelihood of the unseen data on the resulting model. General EM optimization approach is well known for its overfitting problem. To see if the proposed VEM approach can resolve this overfitting problem, experiments are conducted in this section. In these experiments, each of the two approaches are performed in increasing number of iterations and, after each iteration, the likelihood of a set of test utterances on the resulting model is computed. The contour of the average likelihood ratio will illustrate the change in the likelihood of unseen data with the increment of the iterations. To perform these experiments, a randomly chosen speaker from the TFarsdat database and its 6 test utterances are used. Figure 7 illustrates the contours of changes in the likelihood ratio with the increment in the number of iterations for the GMM-EM and GMM-VEM methods. For better illustration, the average log likelihood values are normalized.

As can be seen in Fig. 7, the likelihood of the test data on the GMM speaker model resulted from the EM optimization

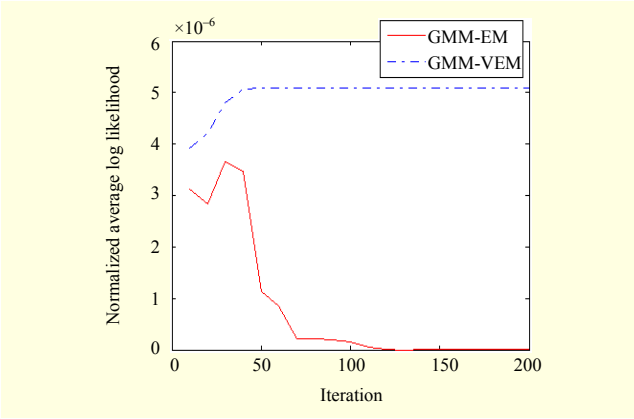


Fig. 7. Change in average likelihood ratio of unseen test utterances on the GMM model versus number of iterations of optimization algorithm using either EM or VEM approaches.

technique rises until about 40 iterations and reduces for more iterations. This is due to the model overfitting. On the other hand, as expected, the average likelihood of the test data on the VEM trained GMM speaker model increases with the number of iterations. However, in contrary to the GMM-EM approach, the VEM approach maintains its generalization power and the average likelihood ratio contour becomes relatively flat. These experiments show that the proposed GMM-VEM approach is robust against the overfitting problem.

The above results are the same as expected. Since the variational approach uses an approximate substitute for the conventional ML/EM optimization, it avoids fitting the training data. Also, it has lead to a more generalizable modeling approach which has achieved a more robust speaker verification performance in the previously discussed experiments.

V. Conclusion

This paper proposed a variational approach for GMM speaker model training and optimization to achieve a better and more reliable performance in text-independent speaker verification. The proposed approach aims to solve the overfitting issue of the traditional expectation maximization GMM optimization technique and resolve the data insufficiency problem in conditions where the available training data for the speakers is scarce or the duration of available speech data for speakers is different. The proposed approach is evaluated on two different speech databases and in adverse conditions. Also, the performance of the proposed approach in noisy environments and different SNR levels is compared with the traditional EM approach. The experiments show that the proposed approach outperforms the GMM-EM approach and can resolve the troublesome characteristic of EM approach, namely overfitting. The reason behind this improvement is that the proposed variational approach optimizes the likelihood function of the GMM model approximately and hence avoids fitting the training data maximally.

References

[1] D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Commun.*, vol. 17, no. 1-2, Aug. 1995, pp. 91-108.

[2] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Commun.*, vol. 52, no. 1, Jan. 2010, pp. 12-40.

[3] K.H. You and H.C. Wang, "Joint Estimation of Feature Transformation Parameters and Gaussian Mixture Model for Speaker Identification," *Speech Commun.*, vol. 28, no. 3, July 1999, pp. 227-241.

[4] K.R. Farrell, R. Mammone, and K. Assaleh, "Speaker Recognition Using Neural Networks and Conventional Classifiers," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 1, 1994, pp. 194-205.

[5] Y. Gu and T. Thomas, "A Text-Independent Speaker Verification System Using Support Vector Machines Classifier," *Proc. European Conf. Speech Commun. Technol.*, 2001, pp. 1765-1769.

[6] L. Heck et al., "Robustness to Telephone Handset Distortion in Speaker Recognition by Discriminative Feature Design," *Speech Commun.*, vol. 31, no. 2-3, 2000, pp. 181-192.

[7] F. Bimbot et al., "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 4, 2004, pp. 430-451.

[8] W. Campbell et al., "Support Vector Machines for Speaker and Language Recognition," *Comput. Speech Language*, vol. 20, no. 2-3, 2006, pp. 210-229.

[9] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, 2006, pp. 308-311.

[10] E. Shriberg et al., "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Commun.*, vol. 46, no. 3-4, 2005, pp. 455-472.

[11] L. Ferrer et al., "Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition," *Proc. ICASSP*, vol. 4, 2007, pp. 233-236.

[12] W. Campbell et al., "Phonetic Speaker Recognition with Support Vector Machines," *Adv. Neural Inf. Process. Syst.*, vol. 16, Cambridge, MA: MIT Press, 2004.

[13] W.M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," *Proc. ICASSP*, 2002.

[14] V. Wan and S. Renals, "Speaker Verification Using Sequence

Discriminant Support Vector Machines," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, 2005, pp. 203-210.

[15] W.M. Campbell et al., "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation," *Proc. ICASSP*, 2006.

[16] R. Dehak et al., "Linear and Non Linear GMM Supervector Machines for Speaker Verification," *Proc. ICSLP*, 2007.

[17] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-Transform Based Speaker Recognition," *Proc. Odyssey*, 2006.

[18] A. Stolcke et al., "MLLR Transforms as Features in Speaker Recognition," *Proc. Interspeech*, 2005.

[19] H. Yang et al., "Cluster Adaptive Training Weights as Features in SVM-Based Speaker Verification," *Proc. Interspeech*, 2007.

[20] C. LongWorth, *Kernel Methods for Text-Independent Speaker Verification*, Ph.D Thesis, Cambridge University Engineering Department, 2010.

[21] X. Dong, W. Zhaohui, and Y. Yingchun, "Exploiting Support Vector Machines in Hidden Markov Models for Speaker Verification," *Proc. ICSLP*, 2002, pp. 1329-1332.

[22] R.A. Redner and H.F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Rev.*, vol. 26, no. 2, 1984, pp. 195-239.

[23] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via EM Algorithm," *J. Royal Statist. Soc.*, vol. 39, no. 1, 1997, pp. 1-38.

[24] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, New York: Wiley, 1997.

[25] S. Pettersen, M. Johnsen, and C. Wellekens, "Variational Bayesian Learning of Speech GMMs for Feature Enhancement Based on Algonquin," *Proc. ICASSP*, vol. 4, 2007, pp. 905-908.

[26] P. Somervuo, "Speech Modeling Using Variational Bayesian Mixture of Gaussians," *Proc. ICSLP*, 2002, pp. 1245-1248.

[27] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, 2000, pp. 19-41.

[28] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, vol. 9, no. 2, 1995, pp. 171-185.

[29] Z. Karam and W. Campbell, "A New Kernel for SVM MLLR Based Speaker Recognition," *Proc. Interspeech*, 2007, pp. 290-293.

[30] A. Stolcke et al., "Speaker Recognition with Session Variability Normalization Based on MLLR Adaptation Transforms," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 7, 2007, pp. 1987-1998.

[31] M.W. Mak, R. Hsiao, and B. Mak, "A Comparison of Various Adaptation Methods for Speaker Verification with Limited Enrollment Data," *Proc. ICASSP*, vol. 1, 2006, pp. 929-932.

[32] J. Mariethoz and S. Bengio, "A Comparative Study of Adaptation Methods for Speaker Verification," *Proc. ICSLP*, 2002, pp. 581-584.

[33] H. Attias, "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," *Proc. 15th Conf. Uncertainty Artif. Intell.*, Stockholm, Sweden, 1999, pp. 21-30.

[34] N. Nasios and A. Bors, "Variational Learning for Gaussian Mixture Models," *IEEE Trans. Systems, Man, Cybern.*, Part B, vol. 36, no. 4, 2006, pp. 849-862.

[35] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science, 2006.

[36] G. Box and G. Tiao, *Bayesian Inference in Statistical Models*, MA: Addison-Wesley, 1992.

[37] Y. Shen et al, "A Comparison of Variational and Markov Chain Monte Carlo Methods for Inference in Partially Observed Stochastic Dynamic Systems," *J. Signal Process. Syst.*, vol. 61, no. 1, 2008, pp. 51-59.

[38] M.I. Jordan et al., "An Introduction to Variational Methods for Graphical Models," *Learning in Graphical Models*, M.I. Jordan, Ed., Cambridge, MA: MIT Press, 1999, pp. 105-161.

[39] T.S. Jaakkola and M.I. Jordan, "Bayesian Parameter Estimation via Variational Methods," *Statistical Commputing*, vol. 10, no. 1, 2000, pp. 25-37.

[40] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. Thesis, University of Cambridge, UK, 2003.

[41] N. Ding and Z. Ou, "Variational Nonparametric Bayesian Hidden Markov Model," *Proc. ICASSP*, 2010, pp. 2098-2101.

[42] D. Su, X. Wu, and L. Xu, "GMM-HMM Acoustic Model Training by a Two Level Procedure with Gaussian Components Determined by Automatic Model Selection," *Proc. ICASSP*, 2010, pp. 4890-4893.

[43] Z. Ghahramani and M.J. Beal, "Variational Inference for Bayesian Mixtures of Factor Analyzers," *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, 2000, pp. 449-455.

[44] S.J. Roberts and W.D. Penny, "Variational Bayes for Generalized Autoregressive Models," *IEEE Trans. Signal Process*., vol. 50, no. 9, 2002, pp. 2245-2257.

[45] T. Minka and J. Lafferty, "Expectation-Propagation for the Generative Aspect Model," *Proc. 18th Conf. Uncertainty Artif. Intell*., Edmonton, AB, Canada, 2002, pp. 352-359.

[46] Y.W. Teh, K. Kurihara, and M. Welling, "Collapsed Variational Inference for HDP," *Adv. Neural Info. Process. Syst.*, vol. 20, 2008.

[47] D.M. Blei and M.I. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, 2005, pp. 121-144.

[48] R.A. Choudrey and S.J. Roberts, "Variational Mixture of Bayesian Independent Component Analyzers," *Neural Comput.*, vol. 15, no. 1, 2003, pp. 213-252.

[49] V.P. Sahu, H.K. Mishra, and C.C. Shekar, "Variational Bayes Adapted GMM Based Models for Audio Clip Classification," *Proc. Int. Conf. Pattern Recognition Mach. Intell.*, 2009, pp. 513-518.

[50] H.K. Mishra and C.C Sekhar, "Variational Gaussian Mixture

Models for Speech Emotion Recognition," *Proc. Int. Conf. Adv. Pattern Recognition*, 2009, pp. 183-186.
[51] F. Valente, *Variational Bayesian Methods for Audio Indexing*, PhD Dissertation, Eurecom, Sept. 2005.
[52] X. Zhao et al., "Variational Bayesian Joint Factor Analysis for Speaker Verification," *Proc. ICASSP*, 2009, pp. 4049-4052.
[53] S. Watanabe et al., "Variational Bayesian Estimation and Clustering for Speech Recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, 2004, pp. 365-381.
[54] D. Cournapeau et al., "Using Online Model Comparison in the Variational Bayes Framework for Online Unsupervised Voice Activity Detection," *Proc. ICASSP*, 2010, pp. 4462-4465.
[55] Q. Huang, J. Yang, and Y. Zhou, "Variational Bayesian Method for Speech Enhancement," *Neurocomput.*, vol. 70, no. 16-18, 2007, pp. 3063-3067.
[56] The NIST Year 2010 Speaker Recognition Evaluation Plan, December 23, 2009. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf, Accessed on 2010-10-22.
[57] M. Bijankhan et al., "TFarsdat, the Telephony Farsi Speech Database," *Proc. EuroSpeech*, 2003, pp. 1525-1528.
[58] J. Godfrey, D. Graff, and A. Martin, "Public Databases for Speaker Recognition and Verification," *Proc. ESCA Workshop Automatic Speaker Recognition*, 1994, pp. 39-42.
[59] M.H. Moattar and M.M. Homayounpour, "A Weighted Feature Voting Approach for Robust and Real-Time Voice Activity Detection," *ETRI J.*, vol. 33, no. 1, Feb. 2011, pp. 99-109.
[60] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
[61] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Process.*, vol. 10, no. 1-3, 2000, pp. 42-54.
[62] A.P. Varga et al., "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Technical Report, DRA Speech Research Unit, 1992.

Mohammad Hossein Moattar received the BSc in computer engineering from Azad University of Mashhad, Mashhad, Iran, in 2003. He received the MSc and PhD in computer engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2005 and 2011, respectively. His research interests include speech and signal rocessing and audio indexing and retrieval.

Mohammad Mehdi Homayounpour received the BSc in electronics engineering from Amirkabir University of Technology in 1986 and the MSc in telecommunications from Khajeh Nasireddin Toosi University, Tehran, Iran, in 1989. He received his PhD in electrical engineering from University of Paris 11 (Orsay), Paris, France, in 1995. Since 1995, he has been an associate professor of the Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Iran. His current research interests are natural language processing, speech and signal processing, and audio indexing. He is a member of the Computer, Information and Telecommunication, and Cryptology Societies of Iran.