Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

LOCAL VARIABILITY MODELING FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Liping Chen[1], *Kong Aik Lee*[2], *Bin Ma*[2], *Wu Guo*[1], *Haizhou Li*[2], and *Li Rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China (USTC), China

[2]Institute for Infocomm Research
Agency for Science, Technology and Research (A*STAR), Singapore

clp2011@mail.ustc.edu.cn, kalee@i2r.a-star.edu.sg

ABSTRACT

Total variability model (TVM) was recently proposed for the compression of speech utterances to low dimensional vectors (i.e., the so-call identity vector or i-vector). Compared to the variable-length nature of speech utterances, i-vectors have fixed length and therefore could be used with simple classifiers for text-independent speaker verification task. This paper proposes the *local variability model* (LVM), the central idea of which is to capture the local variability associated with individual Gaussians in the acoustic space that are absent in the i-vector representation. We analyze the latent structure of both the *total* and *local* variability models and show that tying the latent variable across frames and mixtures leads to powerful methods for extracting information from variable sequences. Experimental results on NIST SRE'08 and SRE'10 datasets show that the proposed LVM is effective for speaker verification.

Index Terms—speaker recognition, factor analysis, session variability

1. INTRODUCTION

Over the past few years, many approaches based on the Gaussian mixture model (GMM) in a GMM-UBM framework [1] have been proposed for text-independent speaker verification task [2]. Inspired by the idea of joint factor analysis (JFA) [3], the total variability model [4] confines the speaker and channel variability within a low-dimensional subspace, leading to a fixed and reduced dimension representation for speech utterances, i.e., the so-called i-vector. Treating an i-vector as a compact representation of a speech utterance, channel compensation techniques, for instance, within-class covariance normalization [5], linear discriminant analysis (LDA) [6], and probabilistic LDA (PLDA) [7] can then be applied effectively on the low-dimensional i-vectors.

An i-vector could be seen as a reduced-dimension representation of a GMM mean supervector (obtained by concatenating the mean vectors in the GMM). Though dimension reduction could be performed on the supervector using deterministic techniques, e.g., principle component analysis (PCA) [6], the i-vector extraction is formulated in probabilistic terms based on a latent variable model. One obvious benefit is that, in addition to obtain the i-vector as the posterior mean of the latent variable, we could also compute the posterior covariance which quantifies the uncertainty of the estimate and fold in the information in subsequent modeling [8]. Among others, probabilistic PCA and factor analysis are two commonly used latent variable model [6] in speech applications. The total variability model, and the JFA alike, is an extension to the classical factor analysis with additional tying of latent variable across frames and mixtures. We shall further elaborate on this in Section 2.

In this paper, we analyze the tying scheme in the *total variability model* (TVM) and propose a different approach by changing the point of tying from the latent variable to the loading matrix across mixtures. We refer to the proposed model as the *local variability model* (LVM) the central idea of which is to capture the local variability factors associate with individual Gaussian in the acoustic space. The difficulty of the LVM lies at the estimation of the loading matrix. As the loading matrix is the same (tied) at each Gaussian, the derivation becomes slightly complicated compared to that of the TVM. In this regard, we derive the posterior inference and sort out the maximum likelihood estimate of the model parameters using the expectation-maximization (EM) algorithm. We also demonstrate the use of the proposed model for text-independent speaker verification task.

The rest of the paper is organized as follows. Section 2 presents a brief overview of the total variability model and the i-vector extraction process. Section 3 proposes the local variability model. In Section 4, we show the use of local variability vectors for speaker verification with PLDA. Section 5 shows some experiment results. Finally, Section 6 concludes the paper.

2. THE I-VECTOR PARADIGM

This section gives a brief overview of the state-of-the-art i-vector extraction procedure. In particular, we emphasize on the idea of latent variable tying across frames and mixtures so as to establish the connection to the local variability modeling proposed in this paper.

2.1. I-vector extraction

The purpose of i-vector extraction is to represent variable-length

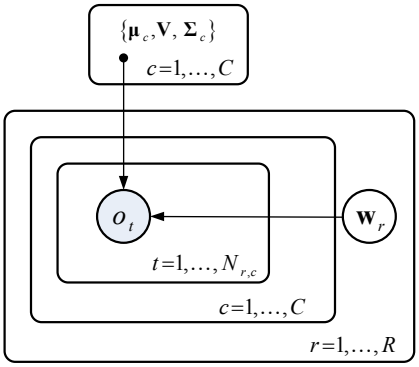

Figure 1: Probabilistic graphical model illustrating the total variability model (TVM). $N_{r,c}$ is the number of frames from session r that is assigned to the c-th Gaussian component, where C is the number of Gaussian components and R is the number of utterances.

utterances with fixed-length and low-dimensional vectors for the classifiers that follows. The fundamental assumption is that the feature vector sequence of an utterance, $\mathcal{O}_r$, was generated from a session-specific GMM. Furthermore, the mean supervector (i.e., obtained by stacking the means from all mixtures) of each session, $\mathbf{m}_r$, is constrained to lie in a low dimensional subspace $\mathbf{T}$ with origin $\boldsymbol{\mu}$, as follows

$$\mathbf{m}_r = \boldsymbol{\mu} + \mathbf{T}\mathbf{w}_r \,. \tag{1}$$

The matrix $\mathbf{T}$, referred to as the total variability matrix, models the speaker and session variations learned from a training set. An i-vector is then taken as the posterior mean of the latent variable $\mathbf{w}_r$, representing both the speaker and session information of an utterance [4]. Here, r is the session index, which gives a separate latent variable for individual utterance. Notice that the rank of the matrix $\mathbf{T}$, and therefore the dimension of the i-vectors, is usually taken to be a small fraction of the supervector. The central idea of total variability modeling is to find a subspace that best describes the speaker and channel variability within the supervector space. In a sense, the extraction of i-vector can be regarded as performing factor analysis on supervector for the purpose of dimension reduction.

2.2. Total variability model

Figure 1 shows the total variability model in the form of a probabilistic graphical model. Here, C denotes the number of Gaussian components, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ denote the mean vector and covariance matrix of the c-th Gaussian, respectively. We decompose the total variability matrix $\mathbf{T} = \left[\mathbf{T}_1^{\mathrm{T}},\ \mathbf{T}_2^{\mathrm{T}},\ \ldots,\ \mathbf{T}_C^{\mathrm{T}}\right]^{\mathrm{T}}$ to its component matrices, one associated with each Gaussian [9]. In Fig. 1, the observations are the acoustic feature vectors o_t represented with shaded circle. The rectangular box surrounding the circle, with the value $N_{r,c}$ at its bottom right corner, indicates that there are $N_{r,c}$ number of observed vectors from the c-th Gaussian for the r-th session:

$$p(o_t \mid c) = \mathcal{N}\left(o_t \mid \boldsymbol{\mu}_c + \mathbf{T}_c\mathbf{w}_r, \boldsymbol{\Sigma}_c\right) \text{ for } t = 1,2,...,N_{r,c} \,. \tag{2}$$

The outer box indicates that the same operation is repeated for all Gaussian components, for $c = 1,2,...,C$. Each Gaussian component accounts for $N_{r,c}$ number of observed vectors $\left\{o_1,\ldots,o_{N_{r,c}}\right\}$, collectively represented as $\mathcal{O}_{r,c}$, the union of which gives rise to the observed sequence $\mathcal{O}_r = \bigcup_{\forall c} \mathcal{O}_{r,c}$.

One important feature of the total variability model (and the joint factor analysis alike) is tying of the observed distributions conditioned on the same latent variable across frames and mixtures. For the current case, the tying appears at two places. Firstly, the latent variable $\mathbf{w}_r$ is tied across observations o_t, for $t = 1,...,N_{r,c}$, pertaining to a Gaussian. Secondly, the same latent variable $\mathbf{w}_r$ is tied across the C Gaussian components. In Fig. 1, the tying of variable is reflected by placing $\mathbf{w}_r$ outside the two rectangular boxes which essentially indicates that the same latent variable (un-shaded circle) is tied across mixtures and across frames of a given speech segment $\mathcal{O}_r$. In mathematical notation, this is reflected by dropping the mixture index c on the latent variable $\mathbf{w}_r$.

The notion of tying the latent variable across frames is based on the assumption that the channel and speaker being constant throughout a given speech segment (e.g., spoken by the same person using the same handset). Similar idea was used in joint factor analysis [3] and the local variability model proposed in this paper. This is different from that of the mixture of factor analyzers [10], and the method proposed in [11], where each frame has its own latent variable. These various assumptions determine the likelihood function used in optimizing the subspace parameter. For the case of total variability model (TVM), the likelihood function is given by

$$l_{\mathrm{TVM}}(\theta) = \prod_{r=1}^{R} \int \left(\prod_{c=1}^{C} \prod_{t=1}^{N_{r,c}} \mathcal{N}\left(o_t \mid \boldsymbol{\mu}_c + \mathbf{T}_c\mathbf{w}_r, \boldsymbol{\Sigma}_c\right) \right) \mathcal{N}\left(\mathbf{w}_r \mid 0, \mathbf{I}\right) d\mathbf{w}_r \,. \tag{3}$$

The first thing to note in (3) is that the latent variable $\mathbf{w}_r \sim \mathcal{N}(0, \mathbf{I})$ is assumed to follows standard normal prior. The same variable $\mathbf{w}_r$ is tied across frames and mixtures for a single session r while separate variables are used for the R speech segments or sessions available for training. Secondly, θ represents the set of model parameters $\left\{\boldsymbol{\mu}_c, \mathbf{T}_c, \boldsymbol{\Sigma}_c; c = 1,2,\ldots,C\right\}$, where the mean vectors and covariance matrices are generally taken as those of the UBM. Though it is possible to update the mean vectors and covariance matrices, they are usually fixed. Essentially, the loading matrices $\mathbf{T}_c$, for $c = 1,2,\ldots,C$, are the remaining parameters to be optimized. Concatenating these matrices one after another in a column wise manner, we form the so-called total variability matrix.

In (3), we assume that the alignment of frames to Gaussian components is known. In practice, this information is given by the zero-order and first-order statistics [3] extracted using the UBM. Given a speech utterance $\mathcal{O}_r$, we treat individual frames o_t as if they were generated from individual Gaussian distributions. An i-vector is then taken as the posterior mean of the latent variable $\mathbf{w}_r$. In particular, the formula for computing an i-vector is:

$$\boldsymbol{\phi}_r = E\left\{\mathbf{w}_r \mid \mathcal{O}_r\right\} = \mathbf{L}_r^{-1}\left(\sum_{c=1}^{C} \mathbf{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c}\right), \tag{4}$$

where

$$\mathbf{L}_r^{-1} = \left(\mathbf{I} + \sum_{c=1}^{C} N_{r,c} \mathbf{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{T}_c \right)^{-1} \quad (5)$$

is the posterior covariance. In (4) and (5), $N_{r,c}$ and $\mathbf{F}_{r,c}$ are the occupancy count and centralized first-order statistics [3] for the c-th Gaussian. Notice that both the posterior mean and covariance are estimated by summing up the statistics from all C mixtures due to the tying of $\mathbf{w}_r$ across components.

Let F be the dimension of the acoustic space and J the dimension of the total variability space. The $CF \times J$ matrix $\mathbf{T}$ consists of C sub-matrices of $F \times J$. The latent variable is tied across mixtures since the speaker and channel effects could be assumed homogenous for all mixtures. From the perspective of mathematical manipulation, tying across mixtures allows the number of columns in the loading matrices $\mathbf{T}_c$ to be higher than the number of rows (i.e., the dimension F of the acoustic space), i.e., $\mathbf{T}_c$ is a landscape matrix. This is so because the dimension of the parameter space becomes C times larger after the tying, where the total variability matrix $\mathbf{T}$ now spans a subspace in the $C \times F$ dimensional supervector space.

3. LOCAL VARIABILITY MODEL

We propose two modifications to the TVM. Firstly, we remove the tying of latent variable across mixtures. Secondly, we replace the former by tying together the loading matrix in each mixture. We refer to the new model as the local variability model (LVM). The motivation, derivation, and parameter learning of the LVM are presented below.

3.1. Local variability model and local variability vectors

The motivation of the proposed local variability model is to extract local variability factors for each component considering the fact that individual Gaussian components of a UBM are associated with specific phonetic contents. This is achieved by assigning one latent variable dedicated to each mixture. Figure 2 shows the proposed local variability model (LVM) in the form of graphical model. One major difference from the TVM (c.f. Fig. 1) is that the circle representing the latent variable is now located inside the second rectangular box. By this we assign separate latent variable $\mathbf{w}_{r,c}$ to the C components, where c = 1, 2,…, C. This essentially removes the tying across the mixtures leading to the following likelihood function for the LVM:

$$l_{\mathrm{LVM}}(\theta) = \prod_{r=1}^{R} \prod_{c=1}^{C} \int \left(\prod_{t=1}^{N_{r,c}} \mathcal{N}\left(o_t \mid \boldsymbol{\mu}_c + \mathbf{V}\mathbf{w}_{r,c}, \boldsymbol{\Sigma}_c\right) \right) \mathcal{N}\left(\mathbf{w}_{r,c} \mid 0, \mathbf{I}\right) d\mathbf{w}_{r,c} \quad (6)$$

The latent variable $\mathbf{w}_{r,c}$ is now marginalized separately for each mixture as opposed to the marginalization over the product across mixtures in (3).

Recognizing the fact that each component has its own latent variable, the posterior distributions can be estimated separately. In this regard, it can be shown that the posterior of the latent variables are normally distributed with mean vector

$$\tau_{r,c} = E\left\{\mathbf{w}_{r,c} \mid \mathcal{O}_r\right\} = \mathbf{L}_{r,c}^{-1} \mathbf{V}^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c}, \quad (7)$$

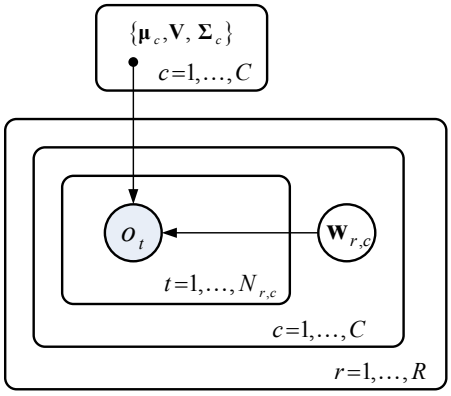

Figure 2: Probabilistic graphical model illustrating the local variability model (LVM). Gaussian components are associated with separate latent variables so as to capture local phonetic variability factors at each component.

and covariance matrix

$$\mathbf{L}_{r,c}^{-1} = \left(\mathbf{I} + N_{r,c} \mathbf{V}^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{V}\right)^{-1}, \quad (8)$$

for $c = 1, 2, ..., C$. Similar to that in (3), we assume that the latent variables $\mathbf{w}_{r,c}$ follow a standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and the frame alignment is known. Different from that in (4), we obtain C posterior distributions, as opposed to only one, given an observation sequence $\mathcal{O}_r$. Comparing (4) to (7), it can be seen that the summation of statistics across mixtures in (4) is due to the tying of latent variable in the TVM which does not exist in the LVM.

The crux of the LVM lies at the tying of loading matrices $\mathbf{T}_c = \mathbf{V}$. By sharing a common loading matrix $\mathbf{V}$ across mixtures, the set of latent variables $\{\mathbf{w}_{r,c}\}_{c=1}^{C}$ share the same set of axes $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_J\}$, i.e., they lie in the same subspace spanned by the columns $\mathbf{v}_j$ of the matrix $\mathbf{V}$. This is of particular interest as the posterior means $\tau_{r,c}$ of the latent variables would lie in the same subspace. The posterior means represent the localized characteristic of the acoustic observations falling into each Gaussian. As they are projected on the same set of axes, we could consider them one dimension at a time. Let the length of the posterior mean vector τ_c be J, which is determined by the number of columns in $\mathbf{V}$. We define the local variability vectors as

$$\rho_j = \left[\tau_1(j), \tau_2(j), \ldots, \tau_C(j)\right]^{\mathrm{T}} \text{ for } j = 1, 2, \ldots, J, \quad (9)$$

where $\tau_c(j)$ denotes the j-th element of the posterior mean vector τ_c of the c-th mixture. Notice that we have dropped the session index r for simplicity. For a given speech utterance, the posterior estimation gives rise to J local variability vectors ρ_j, each with a dimensionality of C. In a sense, the local variability vector ρ_j represents the acoustic information captured from all the C Gaussian components with a projection to a common axis $\mathbf{v}_j$. As individual Gaussian components are associated with different phonetic events, a local variability vector ρ_j would therefore represent all phonetic events at one specific dimension.

3.2. Tied-mixture loading matrix estimation

The difficulty of the LVM lies at the estimation of the loading matrix $\mathbf{V}$. The parameter tying makes the derivation, as shown below, slightly complicated compared to that of the TVM.

As for most latent variable models, we resort to the EM algorithm [6] for estimating $\mathbf{V}$. In particular, we iteratively maximize the following auxiliary function:

$$Q = \sum_{r=1}^{R}\sum_{c=1}^{C} E\left\{ \mathbf{w}_{r,c}^{\mathrm{T}} \mathbf{V}^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c} - \frac{1}{2} N_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}} \mathbf{V}^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \mathbf{V} \mathbf{w}_{r,c} \right\}. \quad (10)$$

In the E-step, we estimate the posterior distributions of the latent variables as in (7) and (8). These are then used in (10) for estimating $\mathbf{V}$ in the M-step. To this end, we take the derivative of (10) with respect to $\mathbf{V}$, as follows

$$\frac{\partial Q}{\partial \mathbf{V}} = \sum_{r=1}^{R}\sum_{c=1}^{C} E\left\{ \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}} - N_{r,c} \boldsymbol{\Sigma}_c^{-1} \mathbf{V} \mathbf{w}_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}} \right\}.$$

Setting the derivative to zero and recognizing that the expectation is taken with respect to the posterior distribution of $\mathbf{w}_{r,c}$, we arrive at

$$\sum_{r=1}^{R}\sum_{c=1}^{C} \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c} E\left\{ \mathbf{w}_{r,c}^{\mathrm{T}} \right\} = \sum_{r=1}^{R}\sum_{c=1}^{C} N_{r,c} \boldsymbol{\Sigma}_c^{-1} \mathbf{V} E\left\{ \mathbf{w}_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}} \right\}. \quad (11)$$

To solve for $\mathbf{V}$, we first notice that the summation across sessions and mixtures on both sides of the equation leads to two $F \times J$ matrices of the same size. Individual elements of the matrix on the right have to correspond to those on the left for the two matrices to be equal.

Notice that the same matrix $\mathbf{V}$ is shared across mixtures in the right-hand-side of (11) due to the parameter tying. One straight forward way to factor $\mathbf{V}$ out from the summation is by assuming that the covariance matrices $\boldsymbol{\Sigma}_c$ are diagonal. With this assumption, the loading matrix $\mathbf{V}$ can be solved one row at a time, as follows:

$$\mathbf{v}_i = \mathbf{a}_i \left\{ \sum_{c=1}^{C} \left(\boldsymbol{\Sigma}_c^{-1} \right)_{ii} \sum_{r=1}^{R} N_{r,c} E\left\{ \mathbf{w}_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}} \right\} \right\}^{-1}, \quad (12)$$

where $\mathbf{a}_i$ is the i-th row of the matrix

$$\mathbf{A} = \sum_{r=1}^{R}\sum_{c=1}^{C} \boldsymbol{\Sigma}_c^{-1} \mathbf{F}_{r,c} \mathbf{w}_{r,c}^{\mathrm{T}}. \quad (13)$$

Since we assume that $\boldsymbol{\Sigma}_c$ is diagonal, $\left(\boldsymbol{\Sigma}_c^{-1} \right)_{i,i}$ denotes the diagonal elements of the precision matrices. Equations (12) and (13) constitute the M-step of the parameter optimization procedure for the proposed LVM. It is worth mentioning that an exact solution exists for $\mathbf{V}$ in (11) and the solution as given by (12) and (13) could be seen as a special case. We shall use the approximated solution for the preliminary investigation as reported in the current paper.

4. PLDA FOR LOCAL VARIABILITY VECTORS

Similar to i-vector, we need to get rid of the influence brought by the channel variability [12]. This is achieved with PLDA by introducing a speaker subspace to tease apart the contribution of the speaker factors from those of the channel factors [13].

Recall that there are J local variability vectors for each given segment of speech. In the current paper, we use a simple strategy whereby one PLDA is applied on each of the J streams of the local variability vectors. The output scores from this parallel bank of PLDAs are linearly combined to give the final score.

Let ρ_j be the j-th local variability vector extracted from a given segment of speech. We assume the following marginal densities for the local variability vectors:

$$p\left(\rho_j\right) = \mathcal{N}\left(\rho_j \,|\, \boldsymbol{\mu}_j, \mathbf{F}_j \mathbf{F}_j^{\mathrm{T}} + \mathbf{G}_j \mathbf{G}_j^{\mathrm{T}} + \boldsymbol{\Sigma}_j\right), \text{ for } j=1,2,\ldots,J. \quad (14)$$

In the above equation, the vector $\boldsymbol{\mu}_j$ denotes the global mean of the j-th stream of local variability vector, $\mathbf{F}_j$ and $\mathbf{G}_j$ are the speaker and channel loading matrices, while the covariance matrix $\boldsymbol{\Sigma}_j$ models the remaining variability not accounted for by the loading matrices. We refer to the set $\theta_j = \left\{\boldsymbol{\mu}_j, \mathbf{F}_j, \mathbf{G}_j, \boldsymbol{\Sigma}_j\right\}$ as the PLDA parameters which could be determined by fitting the model onto the each of the J streams of the local variability vectors extracted from a labeled training set. Details of the EM algorithm used in the current work could be found in [7, 14].

The task of speaker verification is to determine whether an enrollment segment and a test segment are from the same speaker or not [12]. This question gives rise to the following log-likelihood ratio:

$$l\left(\rho_j^{\mathrm{e}}, \rho_j^{\mathrm{t}}\right) = \log \frac{p\left(\rho_j^{\mathrm{t}}, \rho_j^{\mathrm{e}}\right)}{p\left(\rho_j^{\mathrm{t}}\right) p\left(\rho_j^{\mathrm{e}}\right)} \text{ for } j=1,2,\ldots,J, \quad (15)$$

where each of the likelihood terms in the numerator and denominator is evaluated using (14). Here, we use the superscripts $^{\mathrm{e}}$ and $^{\mathrm{t}}$ to denote enrollment and test, respectively. Detailed steps to evaluate the likelihood function can be found in [15]. For each trial, we obtain J number of scores. The performance of the system is partly determined by how to combine the scores for final decision. The simplest way is to average all scores as adopted in the current paper.

5. EXPERIMENTS

Experiments were carried out on the telephone trials of the *short2-short3* task of NIST SRE'08 and the *core-core* task of SRE'10. The nominal duration of the training and test segments was about two and a half minutes. The performance was evaluated based on the equal-error-rate (EER) and the detection cost function (DCF), $C_{\mathrm{DET}} = P_{\mathrm{tar}} P_{\mathrm{miss}}(\theta) + (1 - P_{\mathrm{tar}}) P_{\mathrm{fa}}(\theta)$ [16]. We consider the minimum DCF at two different operation points, namely, DCF08 and DCF10. The minimum DCF is found by sliding the threshold θ for different value of miss and false-alarm probabilities denoted as $P_{\mathrm{miss}}(\theta)$ and $P_{\mathrm{fa}}(\theta)$ respectively.

The acoustic features were 57-dimensional vectors of *mel frequency cepstral coefficients* (MFCC) with first and second derivatives appended. We trained gender-dependent UBMs of 512 Gaussians with NIST SRE'04 dataset. For the i-vector, the Gaussian components of the UBMs have full covariance matrices. For the

local variability model (LVM), we used diagonal covariance matrices for the reason as explained in Section 3.2. The modeling capacity of a UBM reduces when its covariance matrices are constrained to be diagonal [17]. In other words, we should have increased the number of mixtures for the LVM. Nevertheless, we kept the size of the UBM to be the same for both so as not to favor the LVM in the performance comparison.

For i-vector extraction, we trained the total variability matrix **T** with $J = 400$ columns using the telephone data from NIST SRE'04, 05 and 06. As such, the i-vector has a dimensionality of 400. The same dataset was used to train the PLDA model with an eigenvoice matrix of rank 200 and a full covariance matrix. For the LVM, the tied-mixture loading matrix **V** with $J = 57$ columns (which is the same as dimensionality of the acoustic features) was trained using the same dataset as used for training the total variability matrix **T**. This configuration resulted in 57 local variability vectors of 512 dimensions for any given speech segment. The local variability vectors are then modeled with a parallel bank of 57 PLDA models with $\mathbf{F}_j$ of rank 200 and full covariance matrices $\mathbf{\Sigma}_j$. Length normalization is exerted on the local vectors and i-vectors before PLDA modeling [18].

Table I and Table II compare the performances of i-vector and the proposed LVM on NIST SRE'08 and NIST SRE'10, respectively. The i-vector PLDA system is used as the baseline. The results confirm that the local variability vectors extracted using the LVM are effective for speaker characterization even though there is still a considerable gap compared to the baseline i-vector PLDA system. There are two possible reasons for this. Firstly, it is expected that the diagonal covariance UBM used for LVM degrades the performance. Secondly, the modeling using parallel bank of PLDAs is obviously inadequate. As such, it is difficult at the current stage to reach any conclusive result as if the local variability vectors are sufficient in capturing local phonetic information dedicated to each Gaussian. One point for future work is to investigate the use of asymmetric bilinear model [13] for this purpose.

Also shown in the tables is the performance of another approach (denoted as LVM$^+$) whereby the local variability vectors were concatenated to form a supervector and the speaker and channel variability is model in the supervector space with a single PLDA. In this regard, the rank of both **F** and **G** used was 600. Comparing the results between the parallel PLDAs and supervector PLDA approaches, it seems that the later does not show significant advantage over the former especially on CC5 of SRE'10. This might give a hint that the local variability vectors from different streams are less correlated. As such, the asymmetric bilinear model (or tied PLDA) might be a better option.

Table I: Performance comparison of i-vector and local variability model (LVM) on DET6 of *short2-short3* task in NIST SRE'08.

	Male		
	EER (%)	minDCF08	minDCF10
i-vector	3.6617	0.2034	0.6660
LVM	5.9424	0.3251	0.9542
LVM$^+$	6.3045	0.3380	0.8295
	Female		
	EER (%)	minDCF08	minDCF10
i-vector	5.3716	0.2716	0.9967
LVM	8.2940	0.4370	0.9884
LVM$^+$	7.5206	0.3897	0.9873

Table II: Performance comparison of i-vector and local variability model (LVM) on CC5 of *core-core* tests in NIST SRE'10.

	Male		
	EER (%)	minDCF08	minDCF10
i-vector	3.2807	0.1224	0.3711
LVM	4.4325	0.2113	0.6006
LVM$^+$	5.3712	0.2917	0.8463
	Female		
	EER (%)	minDCF08	minDCF10
i-vector	2.8001	0.1402	0.3465
LVM	6.7150	0.2982	0.6648
LVM$^+$	7.9109	0.3775	0.7972

6. CONCLUSION

We have proposed the *local variability model* (LVM) pivoted on the idea of cross-mixture tying upon a common loading matrix. The proposed LVM was formulated as an extended form of the classical factor analysis similar to that used in the i-vector paradigm. The major difference lies at the tying scheme across mixtures, which could best be illustrated using the probabilistic graphical model as shown in the paper. We also derived the posterior inference and the EM steps for parameter learning.

In the LVM, the loading matrix is tied across mixtures so that the same set of basis are used to project local variability, observed in individual Gaussians, on to the same set of axes. In our current implementation, the local variability vectors extracted using the LVM are modeled separately for each of the principle directions. Experimental results confirm that the local variability vectors are effective for speaker characterization, though this approach does not lead to a better performance than the baseline i-vector. One major obstacle remains is the modeling of the local variability vectors for speaker verification. This will be a point for future research.

7. ACKNOWLEDGEMENTS

The work of Liping Chen was partially supported by the National Nature Science Foundation of China (Grant No. 61273264) and the National 973 program of China (Grant No. 2012CB326405).

8. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dumn, "Speaker verification using adapted Gaussian mixture model," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-Based speaker verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448-1460, May 2007.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2006.

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

[7] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision*, 2007.

[8] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterance of arbitrary duration," in *Proc. IEEE ICASSP*, 2013, pp. 7649 - 7653.

[9] P. Kenny, "A Small Footprint i-Vector Extractor," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, June 2012.

[10] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, 1996.

[11] T. Hasan and J. H. Hansen, "Acoustic Factor Analysis for Robust Speaker Verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 21, no. 4, pp. 842-853, Oct 2012.

[12] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, Jun. 2010.

[13] S. J. D. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[14] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification," in *Proc. INTERSPEECH*, 2012, paper 198.

[15] K. A. Lee, A. Larcher, C. H. You, B. Ma, H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection," in *Proc. INTERSPEECH*, 2013, pp. 3651 - 3655.

[16] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230-275, 2006.

[17] P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed plda in i-vector speaker verification," in *Proc. IEEE ICASSP*, 2011, pp. 4828 - 4831.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in Proc. Interspeech, pp. 249-252, Aug. 2011.