FACTORED COVARIANCE MODELING FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Eryu Wang	extsuperscript{1}, Kong Aik Lee	extsuperscript{2}, Bin Ma	extsuperscript{2}, Haizhou Li	extsuperscript{3}, Wu Guo	extsuperscript{1}, Lirong Dai	extsuperscript{1}

	extsuperscript{1}iFly Speech Laboratory, University of Science and Technology of China, China
	extsuperscript{2}Institute for Infocomm Research (I2R), A*STAR, Singapore

ABSTRACT

Gaussian mixture models (GMMs) are commonly used to model the spectral distribution of speech signals for text-independent speaker verification. Mean vectors of the GMM, used in conjunction with support vector machine (SVM), have shown to be effective in characterizing speaker information. In addition to the mean vectors, covariance matrices capture the correlation between spectral features, which also represent some salient information about speaker identity. This paper investigates the use of local correlation between different dimensions of acoustic vector by using factor analysis and linear Gaussian model. Log-Euclidean inner product kernel is used to measure the similarity between two speech utterances in the form of covariance matrices. Experiments carried on NIST 2006 speaker verification tasks shows promising results.

Index Terms—covariance modeling, factor analysis, log-Euclidean, Gaussian mixture model, support vector machine

1. INTRODUCTION

The aim of speaker verification is to evaluate a claimed identity by comparing the claimant’s speech against the corresponding speaker model. In state-of-the-art text-independent speaker verification system, two widely used speaker modeling techniques are Gaussian mixture model (GMM) [1] and support vector machine (SVM) [2]. A GMM models the speaker characteristic as a multimodal distribution of acoustic features. On the other hand, SVM models directly the boundary separating a target speaker from the impostor population.

Recent research has shown a paradigm shift from the traditional vector-based speaker model to the so-called super-vector model. A super-vector is formed by concatenating the mean vectors of a GMM [2], by which the GMM is projected to a high dimensional space. The GMM is estimated, for each speech utterance, based on the same universal background model (UBM) using an adaptation process referred to as the maximum a posteriori (MAP) estimation. By adapting only the mean vectors, while keeping the weights and covariance matrices remain the same as the UBM, this speaker-dependent GMM can therefore be uniquely represented by the mean super-vector. Taking the super-vectors as inputs to SVM, we can take advantages of both generative and discriminative modeling pertaining to the GMM and SVM, respectively.

For computational simplicity, covariance matrices of the GMM are always assumed to be diagonal [1, 2, 3]. The intra-frame correlation is therefore inadequately modeled. In cepstral features, speaker information is generally coupled with the text content, though earlier study has shown that lower coefficients correspond more to text content while higher coefficients contains more speaker information [4]. The correlations between features could therefore provide an additional source of speaker information which has been shown useful in previous studies [8, 5]. The complementary aspect of the covariance matrices and mean vectors can also be understood from the fact that covariance matrices model a distribution by shape, while mean vectors model by location.

In this paper we advocate the use of spectral correlation captured with a factored-analyzed covariance structure, and investigate its effectiveness compared to and in complement with the mean super-vector. As commonly known, a GMM is a linear local model in which each component represents a local distribution. Applying factor analysis on the individual components of the GMM, we place each of the factor analyzers in different parts of the acoustic space. The factored covariance matrices could therefore characterize the local correlation of data through dimensional reduction, which can in turn handle the data scarcity problem. Another challenging issue concerning the use of the covariance matrices with SVM is to design a proper kernel metric. We introduce two metrics – the first operates directly on the factored covariance matrices, whereas the second operates on the loading matrices of the factor analyzers.

2. FACTORED-ANALYZED COVARIANCE MODELING

Gaussian mixture models (GMMs) are commonly used to model the spectral distribution of speech signals. The probability model is

\[
p(x) = \sum_{i=1}^{M} \omega_i N(x; \mu_i, \Sigma_i),
\]

where the function \( N(\cdot) \) denotes the Gaussian density, \( \omega_i \) is the mixture weight, \( \mu_i \) is the mean vector and \( \Sigma_i \) the covariance matrix of the \( i \)th Gaussian component. The dimensionality \( D \) of the feature vector is usually very large prohibiting the use of full covariance matrices as one rarely has enough data to reliably estimate full covariance matrices. One proven solution is to use interpolation [5]. Another option is to use factored covariance matrices [6], which we investigate in this paper for the purpose of speaker verification. Our aim is to capture the most significant correlation in the acoustic features with a small number of parameters. In the following, we start with a brief review of factor analysis for single Gaussian and extent the formulation to GMM.

2.1 Factor Analysis
Let $\mathbf{x}$ be a $D$-dimensional random variable assumed to be normally distributed with mean vector $\mathbf{m}$. Factor analysis assumes that an utterance $\mathcal{X}$ consisting of $T$ independent observations of $\mathbf{x}$, $\{x_1, x_2, \ldots, x_T\}$, is generated via the following generative process:

For $t = 1, 2, \ldots, T$

(a) Draw $z_t$ from $\mathcal{N}(0, \mathbf{1})$

(b) Draw $x_t$ given $z_t$ from $\mathcal{N}(\mathbf{m} + \mathbf{A}z_t, \mathbf{I})$

In the model, $z$ is a random variable having a dimensionality, $f$, smaller than that of $\mathbf{x}$, i.e., $f < D$. A sampled value of the observed variable is generated by sampling $z_t$ from a standard normal distribution as in (a), computing the $D$-dimensional mean vector $\mathbf{A}z_t + \mathbf{m}$, and draw a sample $x_t$ from the resulting density function as in (b). The $D \times f$ matrix $\mathbf{A}$ is known as the loading matrix. The elements of $z$ are known as the latent factors, as opposed to the observed variable $\mathbf{x}$.

Given the relationship between the latent and observed variables as defined above, the marginal probability of $\mathbf{x}$ is again normally distributed, as follows

$$p(\mathbf{x}) = \int p(\mathbf{x} | z)p(z) \, dz = \mathcal{N}(\mathbf{x} | \mathbf{m}, \mathbf{\Psi} + \mathbf{A}\mathbf{A}^T),$$

where the covariance matrix is given by

$$\Sigma = \mathbf{\Psi} + \mathbf{A}\mathbf{A}^T.$$  

The matrix $\mathbf{A}\mathbf{A}^T$ captures the correlations between the elements of $\mathbf{x}$, while the $D \times D$ diagonal matrix $\mathbf{\Psi}$ accounts for the independent variations in each elements of $\mathbf{x}$. The covariance matrix in the form as in (3) is known as the factored covariance matrix. It requires $D(f + 1)$ parameters. By having $f \ll D$, the number of parameters is much lesser compared to a full covariance matrix which requires $D(D + 1)/2$ parameters.

### 2.2 Mixture of Factor Analyzers

Previous studies have shown that speech features exhibits a multimodal distribution [5]. Applying factor analysis in (2) to the individual Gaussian components of (1), we arrive at the following mixture distribution:

$$p(\mathbf{x}) = \sum_{i=1}^{M} \omega_i \mathcal{N}(\mathbf{x} | \mathbf{m}_i, \mathbf{\Psi}_i + \mathbf{A}_i \mathbf{A}_i^T).$$

The motivation of factor analysis is to find a lower dimensional subspace capturing most significant correlation in the observed variable $\mathbf{x}$. By giving each mixture component a different mean, $\mathbf{m}_i$, we place each of the factor analyzers in a different part of the acoustic space. The matrices $\mathbf{A}_i \mathbf{A}_i^T$ could therefore characterize the local correlation of the data that might exhibit different covariance structures at different parts of the acoustic space.

The combined use of factor analysis with GMM was first reported in [7]. The expectation maximization (EM) algorithm for estimating the parameters of the model is summarized as follows. Let $i$ be the mixture index of the GMM, the E-step of the EM algorithm computes the following posterior statistics of the hidden variable $z$ for each mixture component and observation $x_t$:

$$E_i[z | x_t] = \frac{1}{1 + \mathbf{\Lambda}_i \mathbf{\Psi}_i^{-1} \mathbf{\Lambda}_i^{-1}} \mathbf{\Lambda}_i^{-1} \mathbf{\Psi}_i^{-1} [x_t - \mathbf{m}_i],$$

and

$$\gamma_{i}^{(l)} = \frac{\omega_i \mathcal{N}(x_t | \mathbf{m}_i, \mathbf{\Psi}_i + \mathbf{\Lambda}_i ^T)}{\sum_{i=1}^{M} \omega_i \mathcal{N}(x_t | \mathbf{m}_i, \mathbf{\Psi}_i + \mathbf{\Lambda}_i ^T)}$$

is the posterior probability of the $i$th mixture component given the observation $x_t$. For the M-step of the EM algorithm, let define the following quantities:

$$A_{x}^{(i)}(t) = x_t - \frac{1}{N^{(i)}} \sum_{t=1}^{T} \gamma_{i}^{(l)} x_n$$

$$A_{z}^{(i)}(t) = E[z | x_t] - \frac{1}{N^{(i)}} \sum_{t=1}^{T} \gamma_{i}^{(l)} E[z | x_n]$$

where $N^{(i)} = \sum_{t=1}^{T} \gamma_{i}^{(l)}$ is the occupancy count of the $i$th mixture. Using these notations, the M-step updates of the model parameters are as follows:

$$\tilde{\mathbf{\Lambda}}_i = \left( \sum_{t=1}^{T} \gamma_{i}^{(l)} (A_{x}^{(i)}(t)) (A_{x}^{(i)}(t))^T \right) \left( \sum_{t=1}^{T} \gamma_{i}^{(l)} \left[ \mathbf{\Lambda}_i^{-1} + (A_{z}^{(i)}(t)) (A_{z}^{(i)}(t))^T \right] \right)^{-1}$$

$$\tilde{\mathbf{m}}_i = \frac{1}{N^{(i)}} \sum_{t=1}^{T} \gamma_{i}^{(l)} (x_t - \tilde{\mathbf{\Lambda}}_i E[z | x_t])$$

$$\hat{\mathbf{\Psi}}_i = \frac{1}{N^{(i)}} \sum_{t=1}^{T} \gamma_{i}^{(l)} \left[ (A_{x}^{(i)}(t) - \tilde{\mathbf{\Lambda}}_i A_{z}^{(i)}(t))^T + (\tilde{\mathbf{\Lambda}}_i A_{z}^{(i)}(t))^T \right]$$

where $\mathbf{A}$ is the shorthand for $\mathbf{I} + \mathbf{A}_i \mathbf{\Sigma}_k^{-1} \mathbf{A}_i$. In the above equations, the index $(d,d)$ indicates the element in the $d$th row and $d$th column of the matrix, the single $(d)$ index indicates the $d$th element in a vector. The matrices $\mathbf{\Psi}_i$ are kept diagonal, while the off-diagonal correlation elements are modeled by the loading matrices $\mathbf{A}_i$. We repeat the E-step and M-step several iterations until we get a stable loading matrices, mean vectors and diagonal covariance matrices. For ML optimization, a proper initialization is important to avoid sub-optimal solutions. To this end, we use a global loading matrix as our seed loading matrix.

### 3. KERNEL METRICS

We introduce two kernel metrics. The first operates directly on the factored covariance matrices, whereas the second operates on the loading matrix.

#### 3.1 Log-Euclidean Inner Product

Applying factor analysis on GMM, full covariance matrices can be obtained as in (4). These covariance matrices reveal the local correlation of the spectral distribution. Assuming different mixtures are functionally independent, the covariance matrices could be concatenated to form a block diagonal matrix, as follows

$$\hat{\Sigma} = \text{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_M),$$

where $\text{diag}(\cdot, \cdot)$ represents a block diagonal matrix. The resulting $\hat{\Sigma}$ is then a $DM \times DM$ super-matrix of covariance. We call this covariance super-matrix analogous to the mean super-vector [2].

To measure the similarity between two utterances from different speakers in terms of the covariance super-matrices, we introduce the log-Euclidean inner product. For two GMMs, if they have the same weight and mean vectors, the log-Euclidean is a approximated form of Kullback-Leibler (KL) divergence [8].
Generally, the log-Euclidean distance is a proper measurement between two positive symmetric matrices in the informational geometric perspective [9]. Applying the distance measure on the covariance super-matrices, the log-Euclidean distance is given by
\[ d(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = \log_{\text{F}} \tilde{\Sigma}_1 - \log_{\text{F}} \tilde{\Sigma}_2 \] (13)
In the above equation, \( \| \cdot \|_{\text{F}} \) is the Frobenius norm, where \( \| \tilde{\Sigma} \|_{\text{F}} = \sqrt{\text{tr}(\tilde{\Sigma}^T \tilde{\Sigma})} \), and \( \text{tr}(\cdot) \) denotes the trace of the matrix. The log-Euclidean distance corresponds to the Euclidean distance in the logarithm domain. The logarithm of the matrix is given by
\[ \log \tilde{\Sigma} = \tilde{V}(\log \tilde{D})\tilde{V}^T, \] (14)
where \( \tilde{D} \) is the diagonal matrix whose diagonal elements are eigenvalues of \( \tilde{\Sigma} \) and \( \tilde{V} \) is the matrix of eigenvectors. Using the polarization identity, we convert the log-Euclidean distance to the log-Euclidean inner product
\[ \kappa_{\text{LE}}(Y,X) = \text{tr} \left( (\log \tilde{\Sigma}_Y)^T (\log \tilde{\Sigma}_X) \right) = \text{tr} \left( \log \tilde{\Sigma}_Y \log \tilde{\Sigma}_X \right) , \] (15)
where \( \langle \cdot \rangle_T \) is the Frobenius inner product, which is the trace of the product of the two matrices.

### 3.2 Linear kernel with rank normalization

As shown in Section 2.2, the loading matrices represent the correlation information of the full covariance matrices. A loading matrix defines a subspace in which each column stands for a single direction in speaker distribution space. This subspace determines the spectral distribution of speaker. We can use this subspace instead of covariance matrices. Linear kernel with rank normalization [10] can be used here to measure the similarity between two loading matrices. Assuming different mixtures are functionally independent, we arrive at the following linear kernel:
\[ \kappa_{\text{LINEAR}}(Y,X) = \sum_{i=1}^{M} \langle \Lambda_i Y, \Lambda_i X \rangle_{\text{F}}, \] (16)
In (16), \( \Lambda'_i \) is the loading matrix of the \( i \)th mixture after rank normalization, in which each element of the loading matrix is replaced by its rank in a background corpus. Let \( \lambda \) be an element in a loading matrix, rank normalization replaces \( \lambda \) with its normalized rank \( \lambda' \), as follows
\[ \lambda' = \frac{\| b \in B : b < \lambda \|}{| B |} \] (17)
where \( B \) is the background set, and \( | \cdot | \) denotes the cardinality of a set.

### 3.3 Support vector machine

Let \( \kappa \) denote any of the kernel metric in (15) or (16). Given \( X \) as input, the discriminant function of an SVM can be expressed as
\[ f(X) = \sum_{i=1}^{N} \alpha_i y_i \kappa(X_i, X) + \beta \] (18)
where \( N \) is the number of support vectors, \( \alpha_i \) are the weights assigned to the \( i \)th support vector with its label given by \( y_i \in \{-1, +1\} \) and \( \beta \) is the basis parameter.

For ease of implementation using standard SVM packages (libSVM or SVMTorch), the matrices can be first reshaped and concatenated into a column vector. Channel compensation is then performed on the column vectors via nuisance attribute projection (NAP). SVM training can then be implemented using standard SVM packages with a linear kernel.

### 4. Experiments and Discussions

The experiments were carried out on the NIST 2006 speaker evaluation task. The core task consists of 810 target speakers, each enrolled with one side of a five-minute conversations, which roughly contained two minutes of speech. There are 3616 genuine and 47452 impostor trials, where test utterances are scored against target speakers of the same gender. All speech utterances were first pre-processed to remove silence and converted into sequences of 52-dimensional feature vector, each consisting of Mel frequency cepstral coefficients (MFCC) appended with deltas and double deltas. Relative spectral (RASTA) filtering and utterance-level mean and variance normalization were performed. We use two well-known metrics in evaluating the performance of the speaker verification system – equal error rate (EER) and the minimum detection cost function (MinDCF). The EER corresponds to the decision that gives equal false acceptance rate (FAR) and false rejection rate (FRR). The MinDCF is defined as the minimum value of the function \( 0.1 \times \text{FRR} + 0.99 \times \text{FAR} \).

The speaker verification system was designed to be gender-dependent. The gender-dependent UBMs were trained using data drawn from NIST SRE04 data. The same data was used to form the SVM background. The commonly available SVMTorch was used for training of SVMs. HTK toolkit was used for training the UBMs. NAP was used for channel compensation. The corranks for NAP was set to 60 and NIST SRE04 was used as NAP development data. Finally, t-norm was performed at the score level. The t-norm cohorts were derived from NIST SRE05 data.

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>MinDCF(×100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU</td>
<td>3.77</td>
<td>1.89</td>
</tr>
<tr>
<td>LE</td>
<td>7.83</td>
<td>3.32</td>
</tr>
<tr>
<td>MAT</td>
<td>8.11</td>
<td>3.53</td>
</tr>
<tr>
<td>Fusion(MU+LE)</td>
<td>3.42</td>
<td>1.67</td>
</tr>
<tr>
<td>Fusion(MU+MAT)</td>
<td>3.55</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Table 1. EER and MinDCF obtained using the LE kernel with different number of Gaussian mixtures. The number of latent factors is fixed at 20.

<table>
<thead>
<tr>
<th>Mixture number</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER (%)</td>
<td>9.07</td>
<td>8.27</td>
<td>7.83</td>
<td>8.36</td>
<td>11.29</td>
</tr>
<tr>
<td>MinDCF(×100)</td>
<td>3.71</td>
<td>3.52</td>
<td>3.32</td>
<td>3.43</td>
<td>4.36</td>
</tr>
</tbody>
</table>

Table 2: EER and MinDCF obtained using the LE kernel with different number of latent factors. The number of mixtures is fixed at eight.

<table>
<thead>
<tr>
<th>Factor number</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER (%)</td>
<td>11.53</td>
<td>7.83</td>
<td>7.82</td>
<td>7.78</td>
<td>7.83</td>
</tr>
<tr>
<td>MinDCF(×100)</td>
<td>4.54</td>
<td>3.32</td>
<td>3.28</td>
<td>3.25</td>
<td>3.26</td>
</tr>
</tbody>
</table>

Table 3. EER and MinDCF obtained using the linear kernel with different number of latent factors. The number of mixtures is fixed at eight.

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>MinDCF(×100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion(MU+LE)</td>
<td>3.42</td>
<td>1.67</td>
</tr>
<tr>
<td>Fusion(MU+MAT)</td>
<td>3.55</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Table 4. Comparison of EER and MinDCF for the covariance super-matrix (LE), loading matrix (MAT), mean super-vector (MU), and their fusion.
The first half of the experiment compares the performance in using the covariance super-matrix or the loading matrices as input to SVM. First, we tested out different configurations to optimize the factor analysis modeling. As shown in Table 1 for the log-Euclidean (LE) inner product whereby the covariance super-matrix is used, the best performance is obtained when the number of mixtures is set to eight. On the other hand, best performance in obtained with 20 latent factors and above, as shown in Table 2. We use this configuration in subsequent experiment.

Table 3 shows the results when the loading matrices are used as input to the SVM (with the linear kernel with rank normalization as described in Section 3.2). We arrive at the same conclusion as observed in Table 2. Twenty latent factors are enough to model the variation in the subspace. Nonetheless, using the loading matrices as input to SVM gives a slightly worse performance than the covariance super-matrix. This is likely caused by the variances $\Psi_i$, which has been separated from the loading matrices.

The second half of the experiments examines the performance of the proposed method to the state-of-the-art GMM mean super-vector. Different from the proposed method where full covariance matrices are used, the GMM mean super-vector is formed by concatenating the mean vectors of 512 mixtures, each associated with a diagonal covariance matrix. This configuration leads to mean super-vectors of dimensionality 26624, which is twice larger than the vectorized LE super-matrices of dimensionality 11024 and Linear ranked loading matrices of dimensionality 8320. The datasets used for UBM training, SVM background data, NAP and t-norm are all the same as before.

Table 4 shows the EER and MinDCF. Though the proposed factor analyzed covariance modeling coupled with the LE kernel and loading matrices with rank normalized linear kernel do not perform better than the mean super-vector (labeled as MU), their fusion gives considerable improvement as shown in Table 3. Fusion between mean super-vector and covariance super-matrix is better than the other fusion system. Here, linear logistic regression was used for the score fusion. For better fusion system, the fusion leads to 9.28% and 11.64% relative improvement in EER and MinDCF, respectively, compared to the mean super-vector alone which is also can be seen in Fig. 1. This shows the promise of the proposed method to the state-of-the-art GMM mean super-vector and LE method.

5. CONCLUSION

We have introduced a covariance super-matrix representation with factor analyzed method. The aim is to capture the intra-frame and local correlation. We showed that this kind of information can be modeled by factor analysis. We also used two representation term to stand for the correlation information – the log-Euclidean inner product (LE) for covariance super-matrices and the linear kernel with rank normalization (MAT) for loading matrix. Experiments show that the LE performs better than MAT. Fusion between state-of-the-art GMM mean super-vector and LE method, which represent two different sources of statistical information, gave 9.28% and 11.64% relative improvement in EER and MinDCF, respectively, compared to the mean super-vector alone.

6. REFERENCES