Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification

Najim Dehak1,2, Réda Dehak3, Patrick Kenny4, Niko Brummer4, Pierre Ouellet4, Pierre Dumouchel1,2

1Centre de recherche informatique de Montréal (CRIM), Montréal, Canada
2École de Technologie Supérieure (ETS), Montréal, Canada
3Laboratoire de Recherche et de Développement de l’EPITA (LRDE), Paris, France
4Agnitio, Stellenbosch, South Africa
{najim.dehak,patrick.kenny,pierre.ouellet,pierre.dumouchel}@crim.ca
reda.dehak@lrde.epita.fr,nbrummer@agnitio.es

Abstract
This paper presents a new speaker verification system architecture based on Joint Factor Analysis (JFA) as feature extractor. In this modeling, the JFA is used to define a new low-dimensional space named the total variability factor space, instead of both channel and speaker variability spaces for the classical JFA. The main contribution in this approach, is the use of the cosine kernel in the new total factor space to design two different systems: the first system is Support Vector Machines based, and the second one uses directly this kernel as a decision score. This last scoring method makes the process faster and less computation complex compared to others classical methods. We tested several intersession compensation methods in total factors, and we found that the combination of Linear Discriminate Analysis and Within Class Covariance Normalization achieved the best performance. We achieved a remarkable results using fast scoring method based only on cosine kernel especially for male trials, we yield an EER of 1.12% and MinDCF of 0.0094 on the English trials of the NIST 2008 SRE dataset.

Index Terms: Total variability space, cosine kernel, fast scoring, support vector machines.

1. Introduction
The Joint Factor Analysis (JFA) [1] approach has become the state of the art in the field of speaker verification during the last three years. This modeling proposes powerful tools for addressing the problem of speaker and channel variability in Gaussian Mixture Models (GMM) [2] framework. Recently [3], we proposed a new technique for combining the JFA and Support Vector Machines (SVM) for speaker verification. In this modeling the SVMs were applied for the total variability factor vectors which are obtained using the JFA model. The best results were obtained when the cosine kernel was applied in this new space [4]. We also proposed several techniques for compensating for channel effects in the total factor space.

In this paper we propose a new fast scoring method based on the cosine kernel applied on the total variability factors without using the SVM approach. We used the same channel compensation technique as proposed in [3]. The results obtained with this scoring are compared to those obtained with SVM-JFA and classical JFA scorings.

The outline of the paper is as follows. Section 2 describes the joint factor analysis model. In section 3, we present the SVM-JFA approach based on the cosine kernel. Section 4 introduces the fast scoring technique. The comparison between different results is presented in section 5. Section 6 concludes the paper.

2. Joint Factor Analysis
Joint factor analysis is a model used to address the problem of speaker and session variability in GMMs. In this model, each speaker is represented by the means, covariance, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined in some continuous feature space of dimension F. The GMM for a target speaker is obtained by adapting the UBM mean parameters (UBM). In JFA [1], the basic assumption is that a speaker- and channel-dependent supervector \( M \) can be decomposed into a sum of two supervectors: a speaker supervector \( s \) and a channel supervector \( c \):

\[
M = s + c
\]  

where \( s \) and \( c \) are normally distributed.

In [1], Kenny et al. described how the speaker-dependent supervector and channel-dependent supervector can be represented in low-dimensional spaces. The first term in the right-hand side of (1) is modeled by assuming that if \( s \) is the speaker supervector for a randomly chosen speaker, then

\[
s = m + Dz + Vy
\]  

where \( m \) is the speaker- and channel-independent supervector (UBM), \( D \) is a diagonal matrix, \( V \) is a rectangular matrix of low rank, and \( y \) and \( z \) are independent random vectors having standard normal distributions. In other words, \( s \) is assumed to be normally distributed with mean \( m \) and covariance matrix \( VV^* + D^2 \). The components of \( y \) and \( z \) are respectively the speaker and common factors.

The channel-dependent supervector \( c \), which represents channel effects in an utterance, is assumed to be distributed according to

\[
c = UX
\]  

where \( U \) is a rectangular matrix of low rank, and \( x \) has standard normal distribution. This is equivalent to saying that \( c \) is normally distributed with zero mean and covariance \( UU^* \). The components of \( x \) are the channel factors.

3. Support Vector Machines
A Support Vector Machine (SVM) is a classifier used to find a separator between two classes. The main idea of this classifier is
to project the input vectors onto high-dimensional space called **feature space** in order to obtain linear separability. This projection is carried out using a mapping function. In practice, SVMs use kernel functions to perform the scalar product computation in the feature space. These functions allow us to compute the scalar product directly in the feature space without defining the mapping function.

3.1. Total Variability

Classical joint factor analysis modeling based on speaker and channel factors consists in defining two distinct spaces: the speaker space defined by the eigenvoice matrix \( V \) and the channel space defined by the eigenchannel matrix \( U \). The approach that we propose is based on defining only one space, instead of two separate spaces. This new space, which we refer to as the total variability space, simultaneously contains the speaker and channel variabilities. It is defined by the total variability matrix that contains the eigenvectors corresponding to the largest eigenvalues of the total variability covariance matrix. In the new model, we make no distinction between speaker effects and channel effects in GMM supervector space [1]. Given an utterance, the new speaker- and channel-dependent GMM supervector \( M \) defined the equation 1 is rewritten as follows:

\[
M = m + Tw
\]  

where \( m \) is the speaker- and channel-independent supervector (which can be taken to be the UBM supervector), \( T \) is a rectangular matrix of low rank and \( w \) is a random vector having standard normal distribution \( N(0, I) \). The components of the vector \( w \) are the total variability factors. In other words, \( M \) is assumed to be normally distributed with mean vector \( m \) and covariance matrix \( TT^T \).

The process of training the total variability matrix \( T \) is equivalent to learning the eigenvalue \( V \) matrix [1], except for one important difference: in eigenvoice training, all the recordings of a given speaker are considered to belong to the same person; however, in the case of the total variability matrix, a given speaker’s entire set of utterances are regarded as having been produced by different speakers. The new model that we propose can be seen as a principal component analysis and minimizing within class variance. These axes can be represented one class, LDA attempts to define new spatial axes that minimize the intra-class variance caused by channel effects, and to maximize the variance between speakers. The third and last approach is the Nuisance Attribute Projection (NAP), presented in [5]. This technique proposed a channel space definition based on the eigenvectors of the within class covariance matrix. The total factor vectors are projected onto the orthogonal complementary channel space, which is the speaker space.

3.1.1. Within Class Covariance Normalization

Within class covariance normalization is presented in detail in [6] and is successfully applied in speaker factor space [4]. It consists in computing the within class covariance matrix in the total factor space using a set of background impostors. The computation of this matrix is given by:

\[
W = \frac{1}{S} \sum_{i=1}^{S} \sum_{j=1}^{n} (w_i^j - \overline{w}_i)(w_i^j - \overline{w})^T
\]  

where \( \overline{w}_i = \frac{1}{n_s} \sum_{k=1}^{n_s} w_i^k \) is the mean of the speaker factor vectors of each speaker, \( S \) is the number of speakers and \( n_s \) is the number of utterances for each speaker \( s \).

3.3.2. Linear Discriminant Analysis

Linear discriminant analysis is a technique for dimensionality reduction that is widely used in the field of pattern recognition. The idea behind this approach is to seek new orthogonal axes to better discriminate between different classes. The axes found must satisfy the requirement of maximizing between-class variance and minimizing within class variance. These axes can be defined using projection matrix \( A \) comprised of the best eigenvectors (those with largest eigenvalues) of the general eigenvalues equation:

\[
S_b \cdot v = \lambda S_w \cdot v
\]  

where \( \lambda \) is the diagonal matrix of eigenvalues. The matrices \( S_b \) and \( S_w \) correspond respectively to the between class and within class covariance matrices. These are calculated as follows:

\[
S_b = \sum_{i=1}^{S} (w_i - \overline{w})(w_i - \overline{w})^T
\]  

\[
S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \overline{w}_s)(w_i^s - \overline{w}_s)^T
\]
where $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} n_s w_s$ is the mean of all total factor vectors for speaker $s$. $S$ is the number of speakers and $n_s$ is the number of utterances for speaker $s$. In the case of speaker factor vectors, the mean vector of all the speakers’ population $\bar{w}$ is equal to the null vector since, in JFA, the speaker factors have a standard normal distribution $w \sim N(0,I)$ with zero mean and identity covariance matrix. The total factor vectors are subjected to the projection matrix $A$ obtained by LDA. The new cosine kernel between two total factor vectors $w_1$ and $w_2$ can be rewritten as:

$$k(w_1, w_2) = \frac{(A^t w_1)^t (A^t w_2)}{\sqrt{(A^t w_1)^t (A^t w_1)} \sqrt{(A^t w_2)^t (A^t w_2)}}$$

(11)

The motivation for using LDA is that it allows us to define a new projection matrix aimed at minimizing the intra-class variance and maximizing the variance between speakers, which is the key requirement in speaker verification.

### 3.3.3. Nuisance attribute projection

The nuisance attribute projection algorithm is presented in [5]. It is based on finding an appropriate projection matrix intended to remove the channel components. The projection matrix carries out an orthogonal projection in the channel’s complementary space, which depends only on the speaker. The projection matrix is formulated as:

$$P = I - vv^t$$

(12)

where $v$ is rectangular matrix of low rank whose columns are the $k$ best eigenvectors of the same within class covariance matrix (channel covariance) given in equation 6. These eigenvectors define the channel space. The cosine kernel based on the NAP matrix is given as follows:

$$k(w_1, w_2) = \frac{(Pw_1)^t (Pw_2)}{\sqrt{(Pw_1)^t (Pw_1)} \sqrt{(Pw_2)^t (Pw_2)}}$$

(13)

where $w_1$ and $w_2$ are two total variability factor vectors.

### 4. Fast scoring

In this section, based on the results obtained with SVM in the total variability space using the cosine kernel, we propose to directly use the value of the cosine kernel between the target speaker total factors $w_{\text{target}}$ and the test total factors $w_{\text{test}}$ as decision score:

$$\text{score}(w_{\text{target}}, w_{\text{test}}) = \frac{\langle w_{\text{target}}, w_{\text{test}} \rangle}{\|w_{\text{target}}\| \|w_{\text{test}}\|} \theta$$

(14)

The value of this kernel is then compared to the threshold $\theta$ in order to take the final decision. The use of the cosine kernel as a decision score for speaker verification makes the process faster and less complex than other JFA scoring [7].

### 5. Experiments

#### 5.1. Experimental setup

Our experiments operate on cepstral features, extracted using a 25 ms Hamming window. 19 Mel Frequency Cepstral Coefficients together with log-energy were calculated every 10 ms. This 20-dimensional feature vector was subjected to feature warping [8] using a 3 s sliding window. Delta and double-delta coefficients were then calculated using a 5-frame window to produce 60-dimensional feature vectors. We used gender-dependent Universal Background Models (UBM) containing 2048 Gaussians. These UBMs were trained using LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004-2005 speaker recognition evaluation data.

For classical JFA, we used two gender-dependent factor analysis models comprised by 300 speaker factors, 100 channel factors, and common factors. We used decoupled estimation of the eigenvoice matrix $V$ and diagonal matrix $D$ [1]. The eigenvoice matrix $V$ was trained on all the UBM training data, except for the NIST 2004 SRE data. The $D$ matrix was trained on 2004 SRE data. The decision scores obtained with factor analysis were normalized using zt-norm normalization. We used 300 t-norm models and around 1000 z-norm utterances for each gender. All these impostors were taken from the same dataset used for UBM training.

In our SVM-JFA system, we used exactly the same UBM as the classical JFA described above. The total variability matrix $T$ was trained on LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; NIST 2004 and 2005 SRE; and Fisher English database Part 1 and 2. We used 400 total factor vectors. The within class covariance matrix was trained on NIST 2004 and 2005 SRE data. LDA and NAP projection matrices were trained on the same data as the total variability matrix training except for the Fisher English database.

We used around 250 t-norm impostor models taken from NIST 2005 SRE data and around 1200 impostor models taken from Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004 SRE data in order to train the SVM.

The fast scoring is based on the same total variability matrix and total factor vectors as the previous SVM-JFA system. In this modeling, the scores are normalized using the zt-norm technique based on the same t-norm model impostors as in the SVM-JFA system. Data from the preceding training SVM impostors are used as z-norm utterances.

#### 5.2. Results

All our experiments were carried out on the telephone data for the core condition of the NIST 2008 SRE dataset. In the next sections, we compared the results obtained with SVM-JFA and fast scoring approaches with those obtained with classical JFA scoring based on integration over channel factors [1].

#### 5.3. SVM-JFA

We first start by comparing the results obtained with SVM-JFA and classical JFA scoring. Table 1 and 2 give comparison result between SVM-JFA and JFA scoring for both genders. In [3], we proved that both LDA and NAP techniques need to be combined with WCCN in order to obtain the best results. The new WCCN matrix is computed after projecting the total factors with LDA and NAP. We have also found that the best LDA dimension reduction is $\text{dim} = 200$ and the best NAP corank is 150.

We conclude from both tables that the combination of LDA and WCCN definitively gave the best performance compared to other channel compensation techniques. Generally, the SVM-JFA achieves better results than the full configuration for joint factor analysis (with speaker and common factors), especially in male trials. We obtain 1.23% absolute improvement in EER on
Table 1: Comparison of results from JFA scoring and several SVM-JFA channel compensation techniques. The results are given as EER and DCF on the female part of the core condition of the NIST 2008 SRE.

<table>
<thead>
<tr>
<th></th>
<th>English trials</th>
<th>All trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>DCF</td>
</tr>
<tr>
<td>JFA scoring</td>
<td>3.17%</td>
<td>0.0150</td>
</tr>
<tr>
<td>WCCN</td>
<td>4.42%</td>
<td>0.0169</td>
</tr>
<tr>
<td>LDA (200) + WCCN</td>
<td>3.68%</td>
<td>0.0150</td>
</tr>
<tr>
<td>NAP (150) + WCCN</td>
<td>3.95%</td>
<td>0.0157</td>
</tr>
</tbody>
</table>

Table 2: Comparison of results from JFA scoring and several SVM-JFA channel compensation techniques. The results are given as EER and DCF on the male part of the core condition of the NIST 2008 SRE.

<table>
<thead>
<tr>
<th></th>
<th>English trials</th>
<th>All trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>DCF</td>
</tr>
<tr>
<td>JFA scoring</td>
<td>2.64%</td>
<td>0.0111</td>
</tr>
<tr>
<td>WCCN</td>
<td>1.48%</td>
<td>0.0113</td>
</tr>
<tr>
<td>LDA (200) + WCCN</td>
<td>1.28%</td>
<td>0.0095</td>
</tr>
<tr>
<td>NAP (150) + WCCN</td>
<td>1.51%</td>
<td>0.0108</td>
</tr>
</tbody>
</table>

Table 3: Comparison of results from JFA scoring and fast scoring with several channel compensation techniques. The results are given as EER and DCF on the female part of the core condition of the NIST 2008 SRE.

<table>
<thead>
<tr>
<th></th>
<th>English trials</th>
<th>All trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>DCF</td>
</tr>
<tr>
<td>JFA scoring</td>
<td>3.17%</td>
<td>0.0150</td>
</tr>
<tr>
<td>WCCN</td>
<td>3.46%</td>
<td>0.0159</td>
</tr>
<tr>
<td>LDA (200) + WCCN</td>
<td>2.90%</td>
<td>0.0124</td>
</tr>
<tr>
<td>NAP (150) + WCCN</td>
<td>2.63%</td>
<td>0.0133</td>
</tr>
</tbody>
</table>

Table 4: Comparison of results from JFA scoring and fast scoring with several channel compensation techniques. The results are given as EER and DCF on the male part of the core condition of the NIST 2008 SRE.

<table>
<thead>
<tr>
<th></th>
<th>English trials</th>
<th>All trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>DCF</td>
</tr>
<tr>
<td>JFA scoring</td>
<td>2.64%</td>
<td>0.0111</td>
</tr>
<tr>
<td>WCCN</td>
<td>1.32%</td>
<td>0.0140</td>
</tr>
<tr>
<td>LDA (200) + WCCN</td>
<td>1.12%</td>
<td>0.0094</td>
</tr>
<tr>
<td>NAP (150) + WCCN</td>
<td>1.32%</td>
<td>0.0111</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we compare two scoring techniques, SVM and fast scoring. Both techniques are based on a cosine kernel applied in the total factor space, where vectors are extracted using a simple factor analysis. The best results are obtained using fast scoring when LDA and WCCN combination are applied in order to compensate for the channel effects. The use of the cosine kernel as a decision score make the decision process faster and less complex.

7. References


