Front-End Factor Analysis For Speaker Verification

Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet,

***Abstract*—This paper presents an extension of our previous work which proposes a new speaker representation for speaker verification. In this modeling, a new low dimensional speaker- and channel-dependent space is defined using a simple factor analysis. This space is named the total variability space because it models both speaker and channel variabilities. Two speaker verification systems are proposed which use this new representation. The first system is a Support-Vector- Machine-based system that uses the cosine kernel to estimate the similarity between the input data. The second system directly uses the cosine similarity as the final decision score. We tested three channel compensation techniques in the total variability space, which are: Within-Class CovarianceNormalization (WCCN), Linear Discriminate Analysis (LDA), and Nuisance Attribute Projection (NAP). We found that the best results are obtained when LDA is followed by WCCN. We achieved an EER of 1.12% and MinDCF of 0.0094 using the cosine distance scoring on the male English trials of the core condition of the NIST 2008 Speaker Recognition Evaluation dataset. We also obtained 4% absolute EER improvement for both-gender trials on the 10sec-10sec condition compared to the classical joint factor analysis scoring.**

***Index Terms*—Joint Factor Analysis, Total Variability Space, Support Vector Machines, Cosine Distance Scoring.**

I. Introduction

Over recent years, Joint Factor Analysis (JFA) [1], [2], [3] has demonstrated state-of-the-art performance for text-independent speaker detection tasks in the NIST speaker recognition evaluations (SREs). JFA proposes powerful tools to model the inter-speaker variability and to compensate for channel/session variability in the context of Gaussian Mixture Models (GMM)[4].

At the same time the application of Support Vector Machines (SVM) in GMM supervector space [5] yields interesting results, especially when Nuisance Attribute Projection (NAP) is applied to deal with channel effects. In this approach, the kernel used is based on a linear approximation of the Kullback-Leibler (KL) distance between two GMMs. The speaker GMM mean supervectors were obtained by adapting the Universal Background Model (UBM) mean supervector to speaker frames using Maximum A Posteriori (MAP) adaptation [4].

Manuscript received October 15; revised March 3. This work was carried out when the first author was with Centre de recherche informatique de Montréal and École de technologie supérieure, Montréal, Canada. It was supported in part by the Natural Science and Engineering Research Council of Canada and in part by the Canadian Heritage Fund for New Media Research Networks.

N. Dehak is with the Spoken Language Systems group at Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA (e-mail: najim@csail.mit.edu)

P. Kenny and P. Ouellet are with Centre de recherche informatique de Montréal (CRIM), Montréal, Canada (e-mail: patrick.kenny@crim.ca)

R. Dehak is with Laboratoire de Recherche et de Développement de l'EPITA, Paris, France (e-mail: reda.dehak@lrde.epita.fr)

P. Dumouchel is with Centre de recherche informatique de Montréal (CRIM), Montréal, Canada and with École de technologie supérieure (ÉTS), Montréal, Canada (pierre.dumouchel@etsmtl.ca)

In [6], [7], we proposed a new way of combining JFA and Support Vector Machines (SVM) for speaker verification. It consists in directly using the speaker factors estimated with JFA as input to the SVM. We tested several kernels and the best results were obtained using the cosine kernel [6] when Within-Class Covariance Normalization (WCCN) [8] is also used to compensate for residual channel effects in the speaker factor space.

Recently [6], we carried out an experiment which proves that channel factors estimated using JFA, which are supposed to model only channel effects, also contain information about speakers. Based on this, we proposed a new speaker verification system based on factor analysis as a feature extractor [9]. The factor analysis is used to define a new low-dimensional space named total variability space. In this new space, a given speech utterance is represented by a new vector named total factors (we also refer to this vector as "i-vector" in this paper). The channel compensation in this new approach is carried out in low-dimensional space, the total variability space, as opposed to the high-dimensional GMM supervector space [1] for classical JFA [3]. We have proposed two new systems based on this new speech representation. The first system is an SVM-based system which uses the cosine kernel to compute the similarity between the total factors. The second system directly uses the value of the cosine distance computed between the target speaker factors and test total factors as a decision score. In this scoring, we removed the SVM from the decision process. One important characteristic of this approach is that there is no speaker enrollment, unlike in other approaches like SVM and JFA, which makes the decision process faster and less complex. This paper presents more details about how these two new systems were built and shows how the channel compensation techniques are used in order to remove the nuisance direction from these new total factor vectors. The best results are obtained with the Linear Discriminant Analysis (LDA) and WCCN combination which uses the cosine kernel. The motivation for using LDA is to maximize the variance between speakers and minimize the intra-speaker variance, which is the important point in speaker verification.

The outline of the paper is as follows. We first describe the JFA approach in Section II. Section III presents the total variability space, the two new speaker verification systems and all proposed channel compensation techniques. The experiments and results are given in section IV. Section V concludes the paper.

[1] A supervector is composed by stacking the mean vectors from a GMM.

II. Joint Factor analysis

In JFA [1], [2], [3], a speaker utterance is represented by a supervector (M) that consists of additive components from a speaker and a channel/session subspace. Specifically, the speaker-dependent supervector is defined as

$$M = m + Vy + Ux + Dz, \tag{1}$$

where m is a speaker- and session-independent supervector (generally from a Universal Background Model (UBM)), V and D define a speaker subspace (eigenvoice matrix and diagonal residual, respectively), and U defines a session subspace (eigenchannel matrix). The vectors y, z and x are the speaker- and session-dependent factors in their respective subspaces and each is assumed to be a random variable with a Normal distribution $N(0, I)$. To apply JFA to speaker recognition consists of first estimating the subspaces (i.e., V, D, U) from appropriately labelled development corpora and then estimating the speaker and session factors (i.e., x, y, z) for a given new target utterance. The speaker-dependent supervector is given by $s = m + Vy + Dz$. Scoring is done by computing the likelihood of the test utterance feature vectors against a session-compensated speaker model ($M - Ux$). A comparison among several JFA scorings is given in [10].

III. Front-end Factor Analysis

In this section, we present two new speaker verification systems which use factor analysis as a feature extractor. The first system is based on Support Vector Machines and the second one uses the cosine distance value directly as a final decision score.

A. Total variability

The classical JFA 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 represented by the eigenchannel matrix U. The approach that we propose is based on defining only a single space, instead of two separate spaces. This new space, which we refer to as the "total variability space", contains the speaker and channel variabilities simultaneously. It is defined by the total variability matrix that contains the eigenvectors with the largest eigenvalues of the total variability covariance matrix. In the new model, we make no distinction between the speaker effects and the channel effects in GMM supervector space. This new approach is motivated by the experiments carried out in [6], which show that the channel factors of the JFA which normally model only channel effects also contain information about the speaker. Given an utterance, the new speaker- and channel-dependent GMM supervector defined by Equation 1 is rewritten as follows:

$$M = m + Tw, \tag{2}$$

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 a standard normal distribution $N(0, I)$. The components of the vector w are the total factors. We refer to these new vectors as identity vectors or *i-vectors* for short. In this modeling, 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 exactly the same as learning the eigenvoice V matrix (see [11]), except for one important difference: in eigenvoice training, all the recordings of a given speaker are considered to belong to the same person; in the case of the total variability matrix however, a given speaker's entire set of utterances are regarded as having been produced by different speakers (we pretend that every utterance from a given speaker is produced by different speakers). The new model that we propose can be seen as a simple factor analysis that allows us to project a speech utterance onto the low-dimensional total variability space.

The total factor w is a hidden variable, which can be defined by its posterior distribution conditioned to the Baum-Welch statistics for a given utterance. This posterior distribution is a Gaussian distribution (see [11], Proposition 1) and the mean of this distribution corresponds exactly to our i-vector. The Baum-Welch statistics used to extract the i-vector are extracted using the UBM. Suppose we have a sequence of L frames $\{y_1, y_2, ..., y_L\}$ and an UBM Ω composed of C mixture components defined in some feature space of dimension F. The Baum-Welch statistics statistics needed to estimate the i-vector for a given speech utterance u are obtained by

$$N_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega) \tag{3}$$

$$F_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega)\, y_t, \tag{4}$$

where $c = 1, ..., C$ is the Gaussian index and $\mathrm{P}(c|x_t, \Omega)$ corresponds to the posterior probability of mixture component c generating the vector y_t. In order to estimate the i-vector, we also need to compute the centralized first-order Baum-Welch statistics based on the UBM mean mixture components.

$$\tilde{F}_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega)\,(y_t - m_c), \tag{5}$$

where m_c is the mean of UBM mixture component c. The i-vector for a given utterance can be obtained using the following equation:

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} . T^t \Sigma^{-1} \tilde{F}(u). \tag{6}$$

We define $N(u)$ as a diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c I$ $(c = 1, ..., C)$. $\tilde{F}(u)$ is a supervector of dimension $C \times F1$ obtained by concatenating all first-order Baum-welch statistics $\tilde{F}_c$ for a given utterance u. Σ is a diagonal covariance matrix of dimension $CF \times CF$ estimated during factor analysis training (see [11]) and it models the residual variability not captured by the total variability matrix T.

B. Support Vector Machines

Support vector machines are supervised binary classifiers. Proposed by Vapnik [12], they are based on the idea of finding, from a set of supervised learning examples $X = \{(x_1, y_1), (x_2, y_2), ..., (x_M, y_M)\}$, the best linear separator H for distinguishing between the positive examples $(y_i = +1)$ and negative examples $(y_i = -1)$. The linear separator is defined by the following function f:

$$\begin{aligned} f: \mathbb{R}^N &\rightarrow \mathbb{R} \\ x &\mapsto f(x) = w^t x + b, \end{aligned} \tag{7}$$

where x is an input vector and (w, b) are the SVM parameters chosen during the training. The classification of a new example x is based on the sign of the function $f(x)$:

$$h(x) = \text{sign}\left(f(x) = w^t x + b\right). \tag{8}$$

When the kernel function is used, The optimal separator is given by the following formula:

$$h(x) = \sum_{i=1}^{M} \alpha_i^* y_i k(x, x_i) + w_0^*, \tag{9}$$

where α_i^* and w_0^* are the SVM parameters set during the training step.

1) Cosine Kernel: In our previous experiments with SVM applied in the speaker factor space [7], we found that the best results were obtained with the cosine kernel. In the same way, we use the cosine kernel between two i-vectors w_1 and w_2. This kernel is defined by the following equation:

$$k(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{\|w_1\| \, \|w_2\|}. \tag{10}$$

Note that the cosine kernel consists in normalizing the linear kernel by the norm of both i-vectors. It considers only the angle between the two i-vectors and not their magnitudes. It is believed that non-speaker information (such as session and channel) affects the i-vector magnitudes so removing magnitude greatly improves the robustness of the i-vector system.

C. Cosine distance scoring

In this section, we propose a new scoring technique which directly uses the value of the cosine kernel between the target speaker i-vector w_{target} and the test i-vector w_{test} as a decision score:

$$\text{score}\left(w_{\text{target}}, w_{\text{test}}\right) = \frac{\left\langle w_{\text{target}}, w_{\text{test}} \right\rangle}{\left\| w_{\text{target}} \right\| \left\| w_{\text{test}} \right\|} \gtrless \theta \tag{11}$$

The value of this kernel is then compared to the threshold θ in order to take the final decision. The advantage of this scoring is that no target speaker enrollment is required, unlike for support vector machines and classical joint factor analysis, where the target speaker-dependent supervector needs to be estimated in an enrollement step[3]. Note that both target and test i-vectors are estimated exactly in the same manner (there is no extra process between estimating target and test i-vectors), so the i-vectors can be seen as new speaker recognition features. In this new modeling, the factor analysis plays the role of feature extractor rather than modeling speaker and channel effects [3] (this is the reason for the title of this paper). The use of the cosine kernel as a decision score for speaker verification makes the process faster and less complex than other JFA scoring methods [10].

D. Intersession compensation

In this new modeling based on total variability space, we propose carrying out channel compensation in the total factor space rather than in the GMM supervector space, as is the case in classical JFA modeling. The advantage of applying channel compensation in the total factor space is the low dimension of these vectors, as compared to GMM supervectors; this results in a less expensive computation. We tested three channel compensation techniques in the total variability space for removing the nuisance effects. The first approach is Within-Class Covariance Normalization (WCCN) [8], which was successfully applied in the speaker factor space in [7]. This technique uses the inverse of the within-class covariance to normalize the cosine kernel. The second approach is Linear Discriminant Analysis (LDA). The motivation for using this technique is that, in the case where all utterances of a given speaker are assumed to represent one class, LDA attempts to define new special axes that minimize the intra-class variance caused by channel effects, and to maximize the variance between speakers. The advantage of the LDA approach is based on discriminative criteria designed to remove unwanted directions and to minimize the information removed about variance between speakers. Similar work was carried out for speaker verification based on a discriminative version of the nuisance attribute projection algorithm without any success [13]. The last approach is the Nuisance Attribute Projection (NAP) presented in [5]. This technique defines a channel space based on the eigenvectors having the largest eigenvalues of the within-class covariance computed in the i-vector background. The new i-vectors are then projected in the orthogonal complementary channel space, which is the speaker space.

1) Within-Class Covariance Normalization: WCCN was introduced by Andrew Hatch in [8]. This approach is applied in SVM modeling based on linear separation between target speaker and impostors using a one-versus-all decision. The idea behind WCCN is to minimize the expected error rate of false acceptances and false rejections during the SVM training step. In order to minimize the error rate, the author in [8] defines a set of upper bounds on the classification error metric.

The optimized solution of this problem is found by minimizing these upper bounds which, by the same token, minimizes the classification error. This optimization procedure allows us to alter the hard-margin separation formalism of the SVM. The resulting solution is given by a generalized linear kernel of the form

$$k(w_1, w_2) = w_1^t R w_2, \tag{12}$$

where R is a symmetric, positive semi-definite matrix. The optimal normalized kernel matrix is given by $R = W^{-1}$, where W is the within-class covariance matrix computed over all the impostors in the training background. We assume that all utterances of a given speaker belong to one class.

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_i^s - \overline{w_s})\ (w_i^s - \overline{w_s})^t, \tag{13}$$

where $\overline{w_s} = \frac{1}{n_s}\sum_{i=1}^{n_s} w_i^s$ is the mean of i-vectors of each speaker, S is the number of speakers and n_s is number of utterances of speaker s. In order to preserve the inner-product form of the cosine kernel, a feature-mapping function φ can be defined as follows:

$$\varphi(w) = B^t w, \tag{14}$$

where B is obtained through Cholesky decomposition of matrix $W^{-1} = BB^t$. In our approach, the WCCN algorithm is applied to the cosine kernel. The new version of this kernel is given by the following equations:

$$k(w_1, w_2) = \frac{(B^t w_1)^t\ (B^t w_2)}{\sqrt{(B^t w_1)^t\ (B^t w_1)}\ \sqrt{(B^t w_2)^t\ (B^t w_2)}}. \tag{15}$$

The WCCN algorithm uses the within-class covariance matrix to normalize the cosine kernel functions in order to compensate for intersession variability, while guaranteeing conservation of directions in space, in contrast with other approaches such as NAP [5] and LDA [13].

2) Linear Discriminant Analysis: LDA 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 intra-class variance. In our modeling, each class is made up of all the recordings of a single speaker. The LDA optimization problem can be defined according to the following ratio:

$$J(v) = \frac{v^t S_b v}{v^t S_w v}. \tag{16}$$

This ratio is often referred to as the Rayleigh coefficient for space direction v. It represents the amount of information ratio of the between-class variance S_b and within-class variance S_w which is equivalent to Equation 13, given space direction v. These are calculated as follows:

$$S_b = \sum_{s=1}^{S}(w_s - \overline{w})\ (w_s - \overline{w})^t \tag{17}$$

$$S_w = \sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_i^s - \overline{w_s})\ (w_i^s - \overline{w_s})^t, \tag{18}$$

where $\overline{w_s} = \frac{1}{n_s}\sum_{i=1}^{n_s} w_i^s$ is the mean of i-vectors for each speaker, S is the number of speakers and n_s is the number of utterances for each speaker s. In the case of i-vectors, the speaker population mean vector $\overline{w}$ is equal to the null vector since, in FA, these i-vectors have a standard normal distribution $w \sim N(0, I)$, which has a zero mean vector. The purpose of LDA is to maximize the Rayleigh coefficient. This maximization is used to define a projection matrix A composed by the best eigenvectors (those with highest eigenvalues) of the general eigenvalue equation:

$$S_b\, v = \lambda S_w\, v, \tag{19}$$

where λ is the diagonal matrix of eigenvalues. The i-vectors are then submitted to the projection matrix A obtained from LDA. The new cosine kernel between two i-vectors w_1 and w_2 can be rewritten as:

$$k(w_1, w_2) = \frac{(A^t w_1)^t\ (A^t w_2)}{\sqrt{(A^t w_1)^t\ (A^t w_1)}\ \sqrt{(A^t w_2)^t\ (A^t w_2)}} \tag{20}$$

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 nuisance direction. 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

$$P = I - RR^t, \tag{21}$$

where R is a rectangular matrix of low rank whose columns are the k eigenvectors having the best eigenvalues of the same within-class covariance matrix (or channel covariance) given in Equation 13.

These eigenvectors define the channel space. The cosine kernel based on the NAP matrix is given as follows:

$$k(w_1, w_2) = \frac{(Pw_1)^t\ (Pw_2)}{\sqrt{(Pw_1)^t\ (Pw_1)}\ \sqrt{(Pw_2)^t\ (Pw_2)}} \tag{22}$$

where w_1 and w_2 are two total i-vectors.

IV. Experiments

A. Databases

All experiments were carried out on the core condition of both NIST 2006 speaker recognition evaluation (SRE) as development dataset and 2008 SRE as test data. The 2006 evaluation set contains 350 males, 461 females, and 51,448 test utterances. For each target speaker model, a five-minute telephone conversation recording is available containing roughly two minutes of speech for a given speaker. The core condition of the NIST 2008 SRE contains both similar telephone conversation data to 2006 SRE and new interview data. Our experiments are based only on telephone data for both training and testing. The core condition of the 2008 SRE is named short2-short3. It contains 1140 females, 648 males

and 37050 files. We also carried out experiments in short2-10sec and 10sec-10sec conditions of the NIST 2008 SRE. In the first condition, we have one telephone conversation to enroll the target model and ten seconds of telephone speech to verify the identity of the speaker. This condition comprises 1140 females, 648 males and 21907 test files. The second condition is characterized by having a 10-second telephone speech segment for enrolling the target speaker and also a 10-second speech segment for testing. it composed also by 1140 females, 648 males and 21907 test files.

In the NIST evaluation protocol [14], we can use all previous NIST evaluation data and also other corpora to train our systems. For this purpose, we used all the following datasets to estimate our system hyperparameters:

- Switchboard :Switchboard II, Phase 2 and 3. Switchboard II Cellular, Part 1 and 2.
- NIST2004 : NIST 2004 Speaker recognition evaluation.
- NIST2005 : NIST 2005 Speaker recognition evaluation.
- Fisher : Fisher English database Part 1 and 2.

B. Experimental Setup

Our experiments operate on cepstral features, extracted using a 25 ms Hamming window. Every 10 ms, 19 Mel Frequency Cepstral Coefficients (MFCC) together with log energy were calculated. This 20-dimensional feature vector was subjected to feature warping [15] using a 3 s sliding window. Delta and delta-delta coefficients were then calculated using a 5-frame window to produce 60-dimensional feature vectors.

We used gender-dependent UBMs containing 2048 Gaussians and two gender-dependent joint factor analysis configurations. The first JFA is made up of 300 speaker factors and 100 channel factors only. The second configuration is full: we added the diagonal matrix D in order to have speaker and common factors. When the diagonal matrix was estimated, we used a decoupled estimation of the eigenvoice matrix V and diagonal matrix D [3]. We used 400 total factors defined by the total variability matrix T.

The decision scores obtained with the JFA scoring were normalized using zt-norm. We used 300 t-norm models for female trials. We used around 1000 z-norm utterances for females. In our SVM system, we take 307 female models to carry out t-normalization and 1292 female SVM background impostor models to train the SVM.

Table I summarizes all corpora that are used to estimate the UBM, JFA hyperparameters, total variability matrix, LDA, NAP, WCCN, SVM background speakers training. The choice of training each component of our systems on a specific dataset is based on the results obtained on the development dataset, which is from the NIST 2006 SRE.

All the results are reported on the female part of the core condition of the NIST 2006 and 2008 SRE telephone data.

C. SVM-FA

1) Within-class Covariance Normalization: The experiments carried out in this section compare the results obtained with and without applying WCCN to the total variability factors. We also present results given by the JFA scoring, based on integration over channel factors [1], [2]. The results are given in Table II.

TABLE I
CORPORA USED TO ESTIMATE THE UBM, TOTAL VARIABILITY MATRIX (T), LDA AND WCCN

		Switchboard	NIST 2004	NIST 2005
UBM		X	X	X
Full JFA	V	X		X
	D		X	
	U	X	X	X
Small JFA	V	X	X	X
	U	X	X	X
JFA zt-norm		X	X	X
T		X	X	X
WCCN			X	X
NAP		X	X	X
LDA		X	X	X
SVM-impostor		X	X	
SVM-tnorm				X

TABLE II
WCCN PERFORMANCE IN THE TOTAL FACTOR SPACE. THE RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE PART OF THE NIST 2006 AND 2008 SRE CORE CONDITION.

	English trials		All trials	
NIST 2006 SRE	EER	DCF	EER	DCF
JFA: $s = m + Vy$	1.74%	0.012	3.84%	0.022
SVM-FA	3.29%	0.021	5.39%	0.031
SVM-FA: with WCCN	**1.87%**	**0.011**	**2.76%**	**0.017**
NIST 2008 SRE	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**3.68%**	**0.015**	**6.3%**	**0.032**
SVM-FA	5.33%	0.020	8.40%	0.040
SVM-FA: with WCCN	4.73%	0.018	7.32%	0.035

If we compare the results with and without WCCN, we find that its use helps to compensate for channel variability in the total factor space. This improvement was very marked in the NIST 2006 SRE, especially for the all-trials condition. We obtained an EER of 2.76%, which represents a 1% absolute improvement compared to the JFA scoring. However, when we compare the same performance for NIST 2008 SRE data, we can conclude that the classical JFA scoring based on integration over channel factors [1], [2] yields the best results. It can be explained by the fact that the WCCN estimated only on the NIST2004 and 2005 SRE dataset is not appropriate for channel compensation on the NIST 2008 SRE.

2) Linear Discriminant Analysis: This section presents the results obtained with linear discriminant analysis applied to the i-vectors in order to compensate for channel effects. We carried out several experiments using different LDA dimension reductions, in order to show the effectiveness of this technique in removing the unwanted nuisance directions. The results

given in table III were obtained for NIST 2006 SRE dataset.

TABLE III
THE LDA DIMENSIONALITY REDUCTION RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE ENGLISH TRIALS OF THE CORE CONDITION OF THE NIST 2006 SRE.

	EER	DCF
JFA : $s = m + Vy$	1.74%	0.012
No channel compensation	3.29%	0.021
WCCN	**1.87%**	**0.011**
LDA dim $= 400$	2.38%	0.013
LDA dim $= 350$	2.25%	0.013
LDA dim $= 300$	2.31%	0.013
LDA dim $= 250$	2.38%	0.011
LDA dim $= 200$	2.56%	0.013
LDA dim $= 150$	2.65%	0.013
LDA dim $= 100$	2.84%	0.013

These results show the effectiveness of using LDA to compensate for channel effects. A first important remark is that application of LDA to rotate space for minimizing the within-speaker variance, without any dimensionality reduction (dim $= 400$), improves performance in the case of the cosine kernel. If we try to minimize the DCF as requested in the NIST evaluation, the best results are obtained by reducing dimensionality to (dim $= 250$). When no channel compensation is applied, we obtain a DCF value of 0.021. Applying a dimensional reduction from size 400 to 250 significantly improves performance, as shown by the resulting DCF value of 0.011. However, if we compare the EER obtained using LDA with that obtained using WCCN, we find that the latter approach gives better results than the former. This observation motivated us to combine both techniques. We performed several experiments where, in a preliminary step, we applied LDA to remove nuisance directions; thereafter we used WCCN in the reduced space in order to normalize the new cosine kernel. During the training step, we began by training the LDA projection matrix on all data used for training the matrix T; then, we projected the same data in the reduced space in order to compute the within-class covariance matrix. Figure 1 shows the value of MinDCF versus the number of spatial dimensions defined by the LDA, in order to find the optimal dimension of the new space. These results were computed on the NIST 2006 SRE dataset.

The best MinDCF achieved using the combination of LDA and WCCN is 0.010 for English trials and 0.016 for all trials. These results were obtained with a new space dimension of dim $= 200$. Table IV compares these results with those obtained with JFA scoring, WCCN alone and LDA alone on the NIST 2006 and 2008 SRE datasets. We first note that applying WCCN in the LDA-projected space helps to improve performance as compared to LDA alone. If we compare the performance of the LDA and WCCN combination with that obtained with JFA scoring and WCCN alone, we find that this combination achieves the best MinDCF in the English and all-trials conditions of both the NIST 2006 and 2008 SRE datasets. We can see that this combination also yields the best

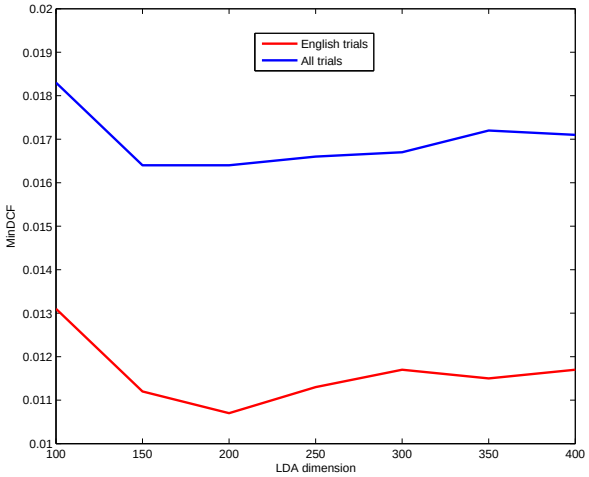

Fig. 1. MinDCF on the NIST 2006 SRE dataset of the SVM-FA system based on LDA technique.

EER in the all-trials condition of both datasets.

TABLE IV
COMPARISON OF RESULTS BETWEEN JFA SCORING AND SEVERAL SVM-FA CHANNEL COMPENSATION TECHNIQUES BASED ON LDA. THE RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE PART OF THE CORE CONDITION OF THE NIST 2006 AND 2008 SRE.

	English trials		All trials	
NIST 2006 SRE	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**1.74%**	0.012	3.84%	0.022
WCCN	1.87%	0.011	2.76%	0.017
LDA (250)	2.38%	0.011	3.31%	0.018
LDA (200) + WCCN	2.05%	**0.010**	**2.72%**	**0.016**
NIST 2008 SRE	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**3.68%**	0.015	6.3%	0.032
WCCN	4.73%	0.018	7.32%	0.035
LDA (200) + WCCN	3.95%	**0.014**	**6.09%**	**0.032**

Figures 2 and 3 show respectively the impact of projecting the i-vectors of five female speakers using the two-dimensional LDA projection matrix only and the impact of LDA followed by WCCN. Two remarks are in order for both figures. First, the application of WCCN in the projected two-dimensional space helps to reduce channel effects by minimizing the intra-speaker variability. Secondly, there is marked dilatation of the i-vectors for each speaker from the origin of the space which can be not compensated for by using the LDA and WCCN combination. This dilatation can be removed by the cosine kernel (normalizing by the length). This behavior explains the extraordinary results obtained with the cosine kernel in i-vector space and also in speaker factor space [6], [7].

3) Nuisance attribute projection: The same study as LDA before was carried out in order to show the performance of the NAP technique for compensating for channel effects. We begin by presenting the results obtained using NAP based on several corank numbers which represent the number of removed

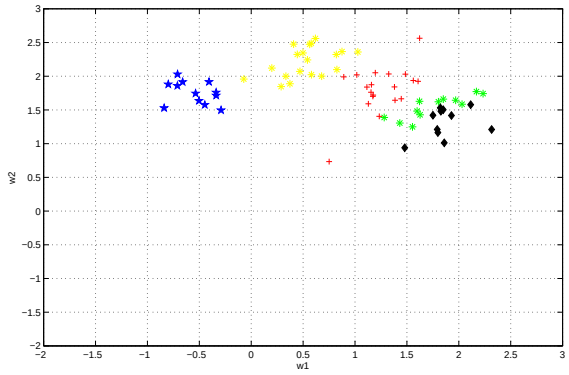

Fig. 2. i-vectors of five speakers after two dimensions LDA projection (w_1, w_2).

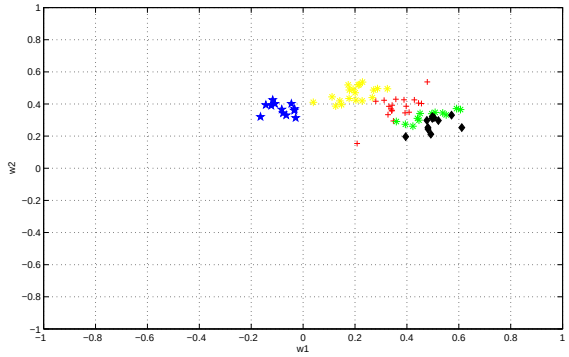

Fig. 3. i-vectors of five speakers after two dimensions LDA and WCCN projection (w_1, w_2).

dimensions. Table V gives the results of these experiments on the female trials of the core condition of the NIST 2006 SRE.

TABLE V
THE RESULTS OBTAINED WITH SEVERAL NAP CORANKS. THESE RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE ENGLISH TRIALS OF THE CORE CONDITION OF THE NIST 2006 SRE.

	EER	MinDCF
JFA : $s = m + Vy$	1.74%	0.012
No channel compensation	3.29%	0.021
WCCN	**1.87%**	**0.011**
NAP corank = 10	2.92%	0.017
NAP corank = 60	2.63%	0.014
NAP corank = 100	2.50%	0.013
NAP corank = 150	2.29%	0.011
NAP corank = 200	2.29%	0.011
NAP corank = 250	2.19%	0.013
NAP corank = 300	2.83%	0.014

These results prove that application of nuisance attribute projection to compensate for channel effects helps to improve the performance of SVM applied to the i-vector space. We decreased the MinDCF for the English trials from 0.021 when

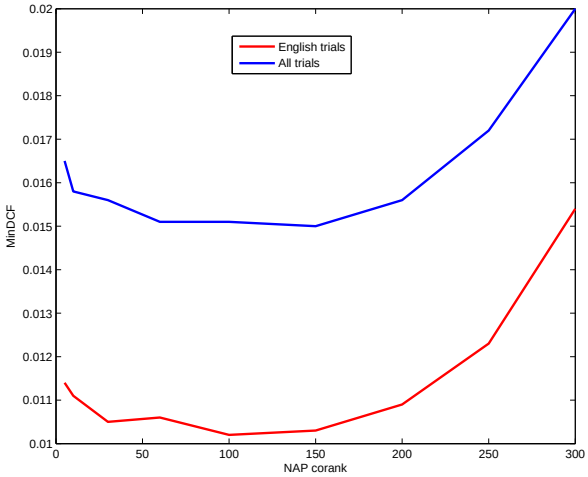

Fig. 4. MinDCF for the NIST 2006 SRE of the SVM-FA system based on the NAP technique.

no channel compensation was applied, to 0.011 when NAP corank is equal to 200. As was the case for LDA, we also found that the WCCN gave better results than NAP, which again persuaded us to combine NAP and WCCN. To train this new approach, we started by first training the nuisance attribute projection matrix in the same manner as before using all the data used in training the total variability matrix (see previous experimental setup section), then we computed the WCCN matrix in the new projected space. The MinDCF of this combination based on varying the number of the NAP corank is given in Figure 4.

The best MinDCF achieved using this combination, based on the NAP and WCCN, is 0.010 for English trials and 0.016 for all trials. These results were obtained with NAP corank equal to 150. Table VI compares these results with those obtained with JFA scoring and WCCN for both NIST 2006 and 2008 SRE datasets. The same remark as in LDA is applicable to the NAP case, which is that the combination of WCCN and NAP improves the performance compared to NAP applied alone. If we compare the performance of the NAP and WCCN combination with that obtained with JFA scoring and WCCN alone for both datasets, we find that this combination achieved the best MinDCF for both datasets. However, the best EER in both datasets are obtained with JFA scoring.

Table VII summarizes the results obtained using JFA scoring and SVM-FA based on WCCN, the LDA and WCCN combination, and NAP combined with WCCN. These results show that the LDA and WCCN combination gives the best DCF (**0.014**) in English trials and also the best EER in all trials; however, the NAP and WCCN combination yielded the best DCF in all trials.

4) Results for both genders: In this section, we present the results for both genders obtained by applying support vector machines in total factor space. We used exactly the same universal background model and factor analysis configuration (400 total factors) as in the last two previous experiments. The only difference lies in the amount of data used to train

TABLE VI
COMPARISON OF RESULTS BETWEEN JFA SCORING AND SEVERAL SVM-FA CHANNEL COMPENSATION TECHNIQUES BASED ON NAP. THE RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE PART OF THE CORE CONDITION OF THE NIST 2006 AND 2008 SRE.

	English trials		All trials	
NIST 2006 SRE	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**1.74%**	0.012	3.84%	0.022
WCCN	1.87%	0.011	2.76%	0.017
NAP (150)	2.29%	0.011	3.38%	0.017
NAP (150) + WCCN	1.83%	**0.010**	**2.66%**	**0.015**
NIST 2008 SRE	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**3.68%**	0.015	**6.3%**	0.032
WCCN	4.73%	0.018	7.32%	0.035
NAP (150) + WCCN	4.73%	**0.015**	6.70%	**0.030**

TABLE VII
SUMMARY OF RESULTS OBTAINED WITH JFA SCORING AND SEVERAL SVM-FA CHANNEL COMPENSATION TECHNIQUES. THE RESULTS ARE GIVEN FOR EER AND MINDCF ON THE FEMALE PART OF THE CORE CONDITION OF THE NIST 2008 SRE.

	English trials		All trials	
	EER	DCF	EER	DCF
JFA : $s = m + Vy$	**3.68%**	0.015	6.3%	0.032
WCCN	4.73%	0.018	7.32%	0.035
LDA (200)+WCCN	3.95%	**0.014**	**6.09%**	0.032
NAP (150)+WCCN	4.73%	0.015	6.70%	**0.030**

the total variability matrix T for both genders. We added the Fisher English database Part 1 and 2 to the previous used data, namely LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004-2005 SRE datasets, in order to capture a greater extent of variability. Note that Fisher corpora are only used to train the total variability matrix and not JFA parameters. The reason is that, in JFA training [3], we used only speakers that have minimum five recordings. However, in Fisher dataset, there are very few speakers that have five recordings and more. The most of these speakers have maximum three recordings. Adding Fisher to train the JFA parameters proves to be not useful. This is not the case in our Total variability matrix training because we used speakers that have minimum two recordings (Fisher dataset contains a lot of a speakers that have minimum two recordings). We applied LDA and NAP, in combination with WCCN, to compensate for channel effects. We used the same female impostors to estimate the SVM model and to carry out the score normalization as described in previous experiments. For Male gender, We used 1007 impostors to train the SVM. These impostors are taken from the same dataset as the UBM training except for the NIST 2005 SRE dataset. We applied t-norm score normalization based on 204 impostors taken from the NIST 2005 SRE dataset. The experiments were carried out on the telephone data of the core condition of the NIST 2008 SRE dataset. Table VIII compares results between SVM-FA and JFA scoring based on both configurations (with and without common factors).

TABLE VIII
COMPARISON OF RESULTS BETWEEN JFA SCORING AND SEVERAL SVM-FA CHANNEL COMPENSATION TECHNIQUES. THE RESULTS ARE GIVEN FOR EER AND MINDCF ON BOTH GENDERS OF THE CORE CONDITION OF THE NIST 2008 SRE DATASET.

	English trials		All trials	
Female gender	EER	DCF	EER	DCF
JFA: *s=m+Vy+Dz*	**3.17%**	0.015	6.15%	0.032
JFA: *s=m+Vy*	3.68%	0.015	6.38%	0.032
LDA (200) + WCCN	3.68%	**0.015**	**6.02%**	**0.031**
NAP (150) + WCCN	3.95%	0.015	6.36%	0.032
Male gender	EER	DCF	EER	DCF
JFA: *s=m+Vy+Dz*	2.64%	0.011	5.15%	0.027
LDA (200) + WCCN	**1.28%**	**0.009**	**4.57%**	**0.024**
NAP (150) + WCCN	1.51%	0.010	4.58%	0.024

Inspection of the tabulated results reveals that, in the case of the SVM-FA system, the LDA/WCCN combination achieves better performance than the NAP/WCCN combination. Adding more training data to the total variability factor space improves the performance of the SVM-FA system. The EER values for the NIST 2008 SRE English trials decreases from 3.95% (Table IV) to 3.68% (Table VIII) when LDA and WCCN are applied. Finally, the SVM-FA achieves better results than the full configuration of the joint factor analysis scoring (with speaker and common factors), especially in male trials. We obtain 1.23% absolute EER improvement For the English trials of the NIST 2008 SRE data. In female trials, the JFA achieves a better English trials EER (a value of 3.17% in EER for JFA scoring compared to 3.68% for the SVM-FA); however, the SVM-FA produced a better EER in all trials (6.02% in EER for SVM-FA compared to 6.15% in EER for JFA scoring). In conclusion, the application of SVM in the total factor space leads to remarkable results compared to those obtained with the full JFA configuration (with common factors), despite the absence of common factors in our new SVM-FA modeling. The results obtained with the cosine kernel applied to these new i-vectors show that there is a quite linear separation between speakers in that space. In the next section, we propose a new scoring based on the cosine kernel values as decision scores.

D. Cosine Distance scoring

Cosine distance scoring is based on the same total variability matrix and i-vectors as the previous SVM-FA system (where the Fisher data are used to train the total variability matrix T). In this modeling, the scores are normalized using the zt-norm technique based on the same t-norm model impostors as in the SVM-FA system. Impostors used for training SVM are used as z-norm utterances in this new system. We used the same LDA and WCCN combination matrix as the SVM-FA system.

The experiments were carried out on the short2-short3 (core condition), short2-10sec and 10sec-10sec conditions of the NIST 2008 SRE dataset. We used exactly the same cosine

distance scoring and channel compensation for all these conditions.

1) Short2-short3 condition: Table IX presents the results obtained with cosine distance, SVM-FA and JFA scorings for both genders on the core condition for telephone data of the NIST 2008 SRE dataset. We used the same channel compensation techniques as in the SVM-FA experiments.

TABLE IX
COMPARISON OF RESULTS FROM JFA, SVM-FA AND COSINE DISTANCE SCORING WITH LDA(DIM=200)+WCCN CHANNEL COMPENSATION TECHNIQUES. THE RESULTS ARE GIVEN AS EER AND DCF ON BOTH GENDER OF THE CORE CONDITION OF THE NIST 2008 SRE DATASET

		English trials		All trials	
		EER	DCF	EER	DCF
Female	JFA	3.17%	0.015	6.15%	0.032
	SVM-FA	3.68%	0.015	6.02%	**0.031**
	cosine	**2.90%**	**0.012**	**5.76%**	0.032
Male	JFA	2.64%	0.011	5.15%	0.027
	SVM-FA	1.28%	0.009	4.57%	0.024
	cosine	**1.12%**	**0.009**	**4.48%**	**0.024**

The results given in this table show that cosine distance scoring based on i-vectors definitively gave the best results in all conditions of the NIST evaluation compared to JFA scoring. If we compare these results with those obtained with the SVM-FA system, we find that cosine distance scoring achieves the best results, especially for female trials. Using cosine distance scoring, we obtained an EER of 2.90% and MinDCF of 0.0124 for English trials versus an EER of 3.68% and MinDCF of 0.0150 for the SVM-FA system. An explanation of these results may be that the background speakers used to train our SVM might not be adequate. Recently [16], the authors proposed a new SVM background speaker selection algorithm for speaker verification. Applying this technique in our modeling will probably improve the performance of the SVM. However, for simplicity, we keep using the cosine distance scoring rather than SVM. Figure 5 and 6 shows a DET curve comparison between classical JFA scoring, SVM-FA combination and cosine distance scoring on the core condition of the NIST 2008 SRE.

2) Short2-10sec condition: Table X presents the results obtained with cosine distance scoring, SVM-FA system and JFA scoring for both genders. The experiments are carried out on telephone data of the short2-10sec condition. In this condition, we have around 2 min of speech to enroll the speaker and 10 s for testing. We used the same channel compensation techniques as in the SVM-FA experiments.

This Table reveals that cosine distance scoring achieves better results than the full joint factor analysis configuration (with speaker and common factors), especially in female trials. We obtain around 2% absolute improvement in EER for the English trials. The cosine distance also gives in general better results than SVM-FA. However, the improvement is barely significant for male trials compared to the female trials.

3) 10sec-10sec condition: Table XI presents the results obtained with cosine distance scoring, full JFA scoring and

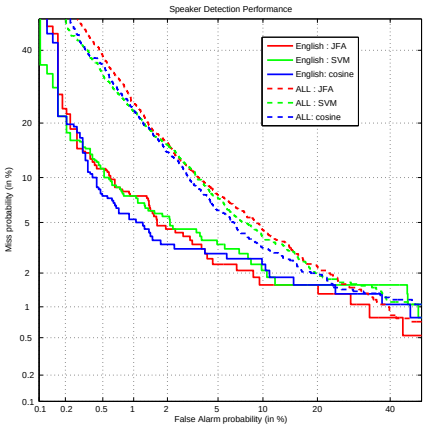

Fig. 5. Detcurves comparison between JFA scoring, SVM-FA and cosine distance. The results are given in English and all trials of female part of core condition of the NIST 2008 SRE.

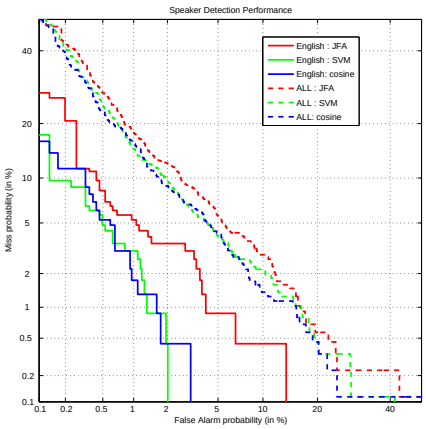

Fig. 6. Detcurves comparison between JFA scoring, SVM-FA and cosine distance. The results are given in English and all trials of male part of core condition of the NIST 2008 SRE.

SVM-FA for both genders on the 10sec-10sec condition for NIST 2008 SRE data. In this condition, we have only 10 seconds of speech to enroll the target speaker model and also 10 seconds for testing, which makes the recognition process more difficult. We used the same LDA and WCCN combination to compensate for channel effects as in the SVM-FA experiments.

The results given in this table show an absolute improvement of around 4% in the EER for both genders. The EER for the English trials goes from 16.01% to 12.19% for females and 15.20% to 11.09% for males. We also note a quite significant improvement in DCF. To our knowledge, these results are the best results ever obtained in the 10sec-10sec condition. It is not easy to explain these extraordinary results obtained with cosine distance scoring. A possible explanation is that in our modeling, we have few parameters to estimate: only 400 total factors compared to JFA, where common factors are also used. This means that we need fewer speech frames to estimate the i-vectors compared to the full JFA. However,

TABLE X
COMPARISON OF RESULTS FROM JFA, SVM-FA AND COSINE DISTANCE SCORING WITH LDA(DIM=200)+WCCN CHANNEL COMPENSATION TECHNIQUES. THE RESULTS ARE GIVEN AS EER AND DCF ON BOTH GENDERS OF SHORT2-10SEC CONDITION OF THE NIST 2008 SRE DATASET

		English trials		All trials	
		EER	DCF	EER	DCF
Female	JFA	7.89%	0.035	11.19%	0.064
	SVM-FA	7.57%	0.034	10.97%	0.052
	cosine	**5.91%**	**0.034**	**9.59%**	**0.050**
Male	JFA	5.36%	0.027	8.09%	0.038
	SVM-FA	5.24%	0.030	7.97%	0.038
	cosine	**5.18%**	**0.026**	**7.38%**	**0.036**

TABLE XI
COMPARISON OF RESULTS FROM JFA, SVM-FA AND COSINE DISTANCE SCORING WITH LDA+WCCN CHANNEL COMPENSATION TECHNIQUES. THE RESULTS ARE GIVEN AS EER AND DCF ON THE FEMALE TRIALS OF 10SEC-10SEC CONDITION OF THE NIST 2008 SRE DATASET

		English trials		All trials	
		EER	DCF	EER	DCF
Female	JFA	16.01%	0.064	17.99%	0.075
	SVM-FA	14.68%	0.062	17.85%	0.073
	cosine	**12.19%**	**0.057**	**16.59%**	**0.072**
Male	JFA	15.20%	0.057	15.45%	0.068
	SVM-FA	12.04%	0.058	14,81%	0.069
	cosine	**11.09%**	**0.047**	**14.44%**	**0.063**

The results obtained with small JFA configuration (without common factor) which are based only on only 400 factors too (300 speaker factor and 100 channel factor), are also worse than cosine distance scoring. As a conclusion, may be the good performances are related to the application of the cosine scoring on the total factor space.

V. CONCLUSION

This paper presented a new speaker verification system where factor analysis is used to define a new low-dimensional space that models both speaker and channel variabilities. We proposed two new scoring methods based on the cosine kernel in the new space. The first approach uses a discriminative method, SVM, and the second one uses the cosine distance values directly as decision scores. The latter approach makes the decision process less complex because there is no speaker enrollment as opposed to the classical methods. In this new modeling, each recording is represented using a low-dimensional vector named i-vector (for *identity vector*) extracted using a simple factor analysis. The main difference between the classical use of joint factor analysis for speaker verification and our approach is that we address the channel effects in this new low-dimensional i-vectors space rather than in the high-dimensional GMM mean supervector space. We tested three different techniques to compensate for the intersession problem: linear discriminant analysis, nuisance attribute projection and within-class covariance normalization. The best results were obtained with the combination of LDA and WCCN. The advantage of using LDA is the removal of nuisance directions and the maximization of the variance between the speakers, which is the key point in speaker verification. The results obtained with cosine distance scoring outperform those obtained with both SVM-FA and classical JFA scorings on several NIST evaluation conditions. However, the cosine scoring system seems to be more powerful and robust, especially on short duration conditions like 10sec-10sec of the NIST 2008 SRE dataset, where we achieved an absolute improvement of 4% on the EER compared to classical JFA. In future work, we will try to extend the total variability systems to the case of the microphone and interview data of the NIST 2008 SRE dataset.

VI. ACKNOWLEDGMENTS

We would like to thank the Center for Spoken Language Processing at Johns Hopkins University for their hospitality and the speaker recognition team for their collaboration. We would like to thank also the anonymous reviewers for their comments that helped to improve the content of this paper.

REFERENCES

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," *IEEE Transaction on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transaction on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transaction on Audio, Speech and Language*, vol. 16, no. 5, pp. 980–988, july 2008.

[4] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, 2006, vol. 1, pp. 97–100.

[6] N. Dehak, *Discriminative and Generative Approches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*, Ph.D. thesis, École de Technologie Supérieure, Montreal, 2009.

[7] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, and V. Hubeika, "Support Vector Machines and Joint Factor Analysis for Speaker Verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009.

[8] 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.

[9] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in *Interspeech*, Brigthon, 2009.

[10] O. Glembek, L. Burget, N. Brummer, and P. Kenny, "Comparaison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009.

[11] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, may 2005.

[12] V.N. Vapnik, *The Nature of Statistical Learning*, Springer, 1995.

[13] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminat NAP for SVM Speaker Recognition," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan 2008.
[14] "http://www.nist.gov/speech/tests/spk/index.htm," .
[15] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, 2001, pp. 213–218.
[16] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM Speaker Verification through Data-Driven Background Dataset Collection," in *IEEE-ICASSP*, Taipei, Taiwan, 2009.