DEALING WITH ADDITIVE NOISE IN SPEAKER RECOGNITION SYSTEMS BASED ON I-VECTOR APPROACH

D. Matrouf, W. Ben Kheder, P-M. Bousquet, M. Ajili and J-F. Bonastre

LIA, University of Avignon, France

ABSTRACT

In the last years, the i-vector approach became the state-of-the-art in speaker recognition systems. As in previous approaches, i-vector -based systems suffer greatly in presence of additive noise, especially in low SNR cases. In this paper, we will describe a statistical framework allowing to estimate a clean i-vector given the noisy one or to integrate, directly, statistical knowledges about the noise and clean i-vectors in the scoring phase. The proposed procedure is essentially based on a method which enables to produce statistical knowledge about the noise effect in the i-vector domain. The work presented here is based on the hypothesis that the noise effect is Gaussian and additive in the i-vector space. To validate our approach, experiments were carried out on NIST 2008 data (det7). Significant improvement was observed compared to the baseline system and to the "muti-style" backend training technique.

Index Terms— i-vector, additive noise, speaker recognition

1. INTRODUCTION

Two kinds of noise can mainly affect the speaker verification performance : convolutive noise and additive noise. In the past years, the research attention was essentially focused on convolutive noise, in the context of i-vector use. The approaches in this category assume that the convolutive noise is Gaussian and additive in the i-vector space. Thus, the statistical knowledge about the convolutive noise are directly accounted for in the scoring phase, which lead to models such as PLDA [1] or two-covariance [2]. On the other hand, and despite its importance, additive noise has not received as much attention as convolutive noise from researchers in the context of i-vector approach. Dealing with additive noise generally falls into one of four categories: speech enhancement, feature compensation, robust modeling or score compensation. We will not discuss the latter here as it does not deal directly with additive noise.

At signal level, [3] proved that spectral and wavelet-based speech enhancement techniques do not perform consistently when used as a pre-processing block in a standard speaker recognition system even if the resultant speech quality increases. It was further shown in [4] that these algorithms might either enhance or degrade the recognition performance depending on the noise type and the SNR level. The speaker-related information has been proven to be vulnerable and hard to handle in this domain due to the natural complexity and redundancy in the speech signal which led to the development of other techniques based on different domains.

At feature level, [5] carried out an extensive comparison of several spectrum estimation methods under additive noise contamination and found that the best spectrum estimator was related to the noise type and level. Recent work [6,7], based on vector Taylor series (VTS) then developed using ”unscented transforms” [8] tried to model non-linear distortions in the cepstral domain based on a non-linear noise model in order to relate clean and noisy cepstral coefficients and help estimate a ”cleaned-up” version of i-vectors. Despite its efficiency, this model remains very rigid due to its complexity and not easily extensible. In such a technique, adding a normalization step or changing the parameters used could involve rewriting the whole technique. On another level, a set of stochastic techniques originally introduced for robust speech recognition such as RATZ [9], SPLICE [10], SSM [11] and TRAJMAP [12] have lately been investigated for speaker recognition [13]. In these techniques, the effect of noise is represented by additive terms in the mean vectors and covariance matrices of clean speech GMMs. Although some of these algorithms achieve very good results (SSM and TRAJMAP), a priori knowledge about the test environment is assumed and sterio training data is required.

On the model level, prior knowledge about the test environment is used in the form of a statistical model of the noise or a reliable estimate of the noise distribution. The parallel model combination (PMC) was first introduced in speech recognition technology [14] before being adapted to speaker recognition [15] by building a noisy model and using it to decode noisy test segments. The use of PMC inside modern speaker recognition i-vector systems is complex, as the noise has to be injected inside all the different models: UBM, i-vector extractor and scoring models. But in practice, the high computational expense, mainly in the scoring model, of such a procedure makes it unfeasible in practice. A robust backend training method called ”multi-style” [16] was proposed as a possible solution to account for the noise in the scoring phase. This method uses a large set of clean an noisy data affected with different noises and SNR levels to build a generic scoring model. The model obtained yields good performance in general, but is still suboptimal for a particular noise because of its generalization (the same system is used for all noises). Another problem with this approach is that it also assumes (theoretically) that test noise is in some way present in the training data, which is not always true. Finally, the use of deep neural networks (DNNs) has been investigated for robust speaker recognition before being successfully applied to speech recognition [17–20]. DNNs have been used either to improve the speaker model (like the ”d-vectors” model proposed in [21] and extracted from the last hidden layer of a DNN) or to improve the computation of the i-vectors statistics in noisy conditions [22]. But in spite of the extensive training time needed to build such models, no significant improvements were observed compared to the previously cited methods.

In [23] we have proposed an i-vector ”denoising” procedure, that we called i-MAP, to deal with additive noise. The advantage of the proposed approach is that we can use a regular clean backend since the resultant i-vectors are assumed to be noise free. In order to build this system, a number of assumptions were made over the clean i-vectors and the noise distribution in the i-vector space. In the work proposed, an estimation of the clean i-vector corresponding to a noisy one is obtained as the maximum a posteriori given the noisy i-vector, the pdf of noise and the pdf of clean i-vectors (in the i-vector space). In this paper we will present a more general statistical framework in which we will estimate the a posteriori pdf of the clean i-vector (Gaussian with full covariance), given the noisy i-vector, the pdf of noise and clean i-vectors. In the scoring phase, using the two-covariance method, instead of using a single estimate of the clean i-vector, we will integrate with respect to the obtained a posteriori pdf.

2. A POSTERIORI CLEAN I-VECTOR PDF ESTIMATION

We assume that both clean i-vectors and noise are normally distributed in the i-vector space. The first hypothesis is justified by the factor analysis model used to extract the i-vectors [24] which assumes a normal distribution for the resulting i-vectors. Regarding the noise, Gaussian distribution modeling seems to be suitable . Even though, the noise is theoretically known to be non-additive in the i-vector space, an additive noise model seems to give encouraging results. It shows an improvement by up to 60% in the recognition performance compared to the baseline system and by nearly 30% compared to the ”multi-style” scoring regime. In addition, the approach is extensible to a mixture of Gaussians to model the noise in i-vector space. The robustness of the proposed technique comes from the fact that it does not use information (pdf) about the distortion caused by the noise but also information (noise i-vector pdf) about the target i-vector (clean i-vector pdf).

Formally, given a noisy i-vector $\mathbf{w}_1$, let’s define two random variables X and Y corresponding respectively to the clean and noisy i-vectors. We define the noise random variable N by:

$$N = Y - X \tag{1}$$

We consider that clean i-vectors X are normally distributed as described in [24], and assume that noise (N) can also be represented by a normal distribution in the i-vector space. We can then define the corresponding probability distribution functions $f(X)$ and $f(N)$ as :

$$f(X) = \mathcal{N}(\mu_X, \Sigma_X) \tag{2}$$

$$f(N) = \mathcal{N}(\mu_N, \Sigma_N) \tag{3}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a normal distribution with mean μ and full covariance matrix Σ.

Referring to (1), (2) and (3) we can express $f(\mathbf{w}_1|X)$ for a given $\mathbf{w}_1$ as:

$$f(\mathbf{w}_1|X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_N|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{w}_1 - X - \mu_N)^t \Sigma_N^{-1}(\mathbf{w}_1 - X - \mu_N)} \tag{4}$$

Using the Bayesian rule, we can write $f(X/\mathbf{w}_1)$ as :

$$f(X/\mathbf{w}_1) = \frac{f(\mathbf{w}_1/X)f(X)}{f(\mathbf{w}_1)} \tag{5}$$

In [25] it is shown that the product of two normal laws is proportional to a third normal law (section 5.6). The resulting constant of proportionality exactly cancels out with the term $f(\mathbf{w}_1)$ (of equation 5) ensuring that the result is a valid probability density function. Hence $f(X/\mathbf{w}_1)$ is a Gaussian with mean $\mu_{\mathbf{w}_1}$ and covariance matrix $\Sigma_{\mathbf{w}_1}$:

$$\mu_{\mathbf{w}_1} = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1}(\Sigma_N^{-1}(\mathbf{w}_1 - \mu_N) + \Sigma_X^{-1}\mu_X) \tag{6}$$

$$\Sigma_{\mathbf{w}_1} = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1} \tag{7}$$

3. ESTIMATION OF $F(X)$ AND $F(N)$

The clean i-vectors distribution $f(X)$ and the noise distribution $f(N)$ are the two most important components in this denoising procedure. $f(X)$ has the advantage of being noise-independent, so it could be estimated once and for all over a large set of clean i-vectors in an off-line step before performing any compensation.

On the other hand, $f(N)$ makes the system able to adapt to the noise present in the signal and compensate its effect more effectively. It is estimated for each different test noise and it requires the existence of clean i-vectors and the noisy versions corresponding to the same segments. First, for the clean part and once the train files are fixed, the corresponding clean i-vectors (X) are extracted. Then, for a given noisy test segment, the noise is extracted from the signal (using a VAD and selecting the low-energy frames) then added to the clean train audio files. Finally, the corresponding noisy i-vectors (Y) are estimated and (1) is used to compute N then $f(N)$.

4. EXTENDED TWO-COVARIANCE SCORING

4.1. Two-covariance scoring

This model is a particular case of the Probabilistic Linear Discriminant Analysis (PLDA) described in [1]. It can be seen as a scoring method and a convolutive noise compensation technique. It consists of a simple linear-Gaussian generative model in which an i-vector $\mathbf{w}$ of a speaker s can be decomposed in:

$$\mathbf{w} = y_s + \varepsilon \tag{8}$$

where the speaker model y_s is a vector of the same dimensionality as an i-vector, ε is Gaussian noise and :

$$P(y_s) = \mathcal{N}(\mu, B) \tag{9}$$

$$P(\mathbf{w}|y_s) = \mathcal{N}(y_s, W) \tag{10}$$

$\mathcal{N}$ denotes the normal distribution, μ represents the overall mean of the training data set, B and W are the between- and within-speaker covariance matrices defined as :

$$B = \sum_{s=1}^{S} \frac{n_s}{n}(y_s - \mu)(y_s - \mu)^t$$

$$W = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(\mathbf{w}_i^s - y_s)(\mathbf{w}_i^s - y_s)^t$$

where n_s is the number of utterances for speaker s, n is the total number of utterances, $\mathbf{w}_i$ are the i-vectors of sessions of speaker s, y_s is the mean of all the i-vectors of speaker s and μ represents the overall mean of the training data set. Under assumptions (8, (9 and (10 the two-covariance score can be expressed as:

$$score(\mathbf{w}_1, \mathbf{w}_2) = \frac{\int \mathcal{N}(\mathbf{w}_1|y, W)\mathcal{N}(w_2|y, W)\mathcal{N}(y|\mu, B)dy}{\prod_{i=1,2} \int \mathcal{N}(w_i|y, W)\mathcal{N}(y|\mu, B)dy} \tag{11}$$

the explicit solution of (11) is given in [2]:

$$score\left(\mathbf{w}_1, \mathbf{w}_2\right) = \widetilde{\mathbf{w}}_1^t \mathcal{P} \widetilde{\mathbf{w}}_2 + \frac{1}{2}\widetilde{\mathbf{w}}_1^t \mathcal{Q} \widetilde{\mathbf{w}}_1 + \frac{1}{2}\widetilde{\mathbf{w}}_2^t \mathcal{Q} \widetilde{\mathbf{w}}_2 \tag{12}$$

where :

$$\begin{aligned}
\widetilde{\mathbf{w}}_i &= \mathbf{w}_i - \mu \\
\mathcal{P} &= \mathbf{W}^{-1}\left(2\mathbf{W}^{-1} + \mathbf{B}\right)^{-1}\mathbf{W}^{-1} \\
\mathcal{Q} &= \mathbf{W}^{-1}\left(2\mathbf{W}^{-1} + \mathbf{B}^{-1}\right)^{-1}\mathbf{W}^{-1} \\
&\quad - \mathbf{W}^{-1}\left(\mathbf{W}^{-1} + \mathbf{B}^{-1}\right)^{-1}\mathbf{W}^{-1}
\end{aligned}$$

In i-vector -based speaker recognition systems [24], length normalization was shown to improve the overall performance [26]. In our case, it is important to mention that all the noisy and clean i-vectors used were initially length-normalized. $\mu \approx 0$ after conditioning so it is possible to ignore it. We will do that in the following.

4.2. Two-covariance score with the a posteriori pdf

In this section, we assume that the target i-vector is clean and the test one is noisy. In section 2 we have shown that the a posteriori pdf of the clean test i-vector given the noisy one is Gaussian. We denote this pdf as $\mathcal{N}\left(\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right)$ (see equations 6 and 7). In equation 12, we will replace $\mathbf{w}_1$ by $\mathcal{N}\left(\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right)$ The score becomes[1]:

$$\int_{\mathbf{w}} \left(\mathbf{w}^t \mathcal{P}\mathbf{w}_2 + \frac{1}{2}\mathbf{w}^t \mathcal{Q}\mathbf{w} + \frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w}$$

$$\begin{aligned}
&= \int_{\mathbf{w}} \left(\mathbf{w}^t \mathcal{P}\mathbf{w}_2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w} \\
&+ \int_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^t \mathcal{Q}\mathbf{w}\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w} \\
&+ \int_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w}
\end{aligned}$$

First term:

$$\begin{aligned}
&\int_{\mathbf{w}} \left(\mathbf{w}^t \mathcal{P}\mathbf{w}_2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w} \\
=& \left(\int_{\mathbf{w}} \mathbf{w}^t \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w}\right) \mathcal{P}\mathbf{w}_2 \\
=& \mathbf{E}^t_{\mathcal{N}\left(\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right)}\left[\mathbf{w}\right] \mathcal{P}\mathbf{w}_2 = \mu_{\mathbf{w}_1}^t \mathcal{P}\mathbf{w}_2
\end{aligned}$$

Third term:

$\int_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w} = \frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2$

Second term:

$$\begin{aligned}
&\int_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^t \mathcal{Q}\mathbf{w}\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}_1}, \mathbf{\Sigma}_{\mathbf{w}_1}\right) d\mathbf{w} \\
=& \frac{1}{2}\int_{\mathbf{w}} \frac{\mathbf{w}^t \mathcal{Q}\mathbf{w}}{(2\pi)^{\frac{p}{2}} \left|\mathbf{\Sigma}_{\mathbf{w}_1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{w}-\mu_{\mathbf{w}_1}\right)^t \mathbf{\Sigma}_{\mathbf{w}_1}^{-1}\left(\mathbf{w}-\mu_{\mathbf{w}_1}\right)} d\mathbf{w}
\end{aligned}$$

where p is the i-vector space dimension. $\mathcal{Q}$ is symmetric and strictly positive-definite. Defining a new variable $\mathbf{z} = \mathcal{Q}^{\frac{1}{2}}\mathbf{w}$, the second term can be rewritten :

$$\begin{aligned}
&\frac{1}{2}\left|J_{\mathbf{z}}\left(\mathbf{w}\right)\right|^{-1} \times \\
&\int_{\mathbf{z}} \frac{\mathbf{z}^t \mathbf{z}}{(2\pi)^{\frac{p}{2}} \left|\mathbf{\Sigma}_{\mathbf{w}_1}\right|^{\frac{1}{2}}} \times e^{-\frac{1}{2}\left(\mathbf{z}-\mathbf{z}_1\right)^t \mathcal{Q}^{-\frac{1}{2}} \mathbf{\Sigma}_{\mathbf{w}_1}^{-1} \mathcal{Q}^{-\frac{1}{2}}\left(\mathbf{z}-\mathbf{z}_1\right)} d\mathbf{z}
\end{aligned}$$

where $\mathbf{z}_1 = \mathcal{Q}^{\frac{1}{2}}\mu_{\mathbf{w}_1}$ and $\left|J_{\mathbf{z}}\left(\mathbf{w}\right)\right|$ is the determinant of Jacobian $\mathbf{z}$ with respect to $\mathbf{w}$ and equal to : $\left|\frac{\delta \mathbf{z}}{\delta \mathbf{w}}\right| = \left|\mathcal{Q}^{\frac{1}{2}}\right| = \left|\mathcal{Q}\right|^{\frac{1}{2}}$ and the term becomes equal to :

$$\begin{aligned}
&\frac{1}{2}\left(2\pi\right)^{-\frac{p}{2}}\left|\mathbf{\Sigma}_{\mathbf{w}_1}\right|^{-\frac{1}{2}}\left|\mathcal{Q}\right|^{-\frac{1}{2}}\int_{\mathbf{z}} \mathbf{z}^t \mathbf{z} \times e^{-\frac{1}{2}\left(\mathbf{z}-\mathbf{z}_1\right)^t \mathbf{\Lambda}\left(\mathbf{z}-\mathbf{z}_1\right)} d\mathbf{z} \\
=& \frac{1}{2}\left|\mathbf{\Sigma}_{\mathbf{w}_1}\right|^{-\frac{1}{2}}\left|\mathcal{Q}\right|^{-\frac{1}{2}}\left|\mathbf{\Lambda}\right|^{-\frac{1}{2}}\int_{\mathbf{z}} \mathbf{z}^t \mathbf{z} \times \mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right) d\mathbf{z} \\
=& \frac{1}{2}\int_{\mathbf{z}} \mathbf{z}^t \mathbf{z} \times \mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right) d\mathbf{z}
\end{aligned}$$

where $\mathbf{\Lambda} = \mathcal{Q}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{w}_1}^{-1}\mathcal{Q}^{-\frac{1}{2}}$

$$\begin{aligned}
&\int_{\mathbf{z}} \mathbf{z}^t \mathbf{z} \mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right) d\mathbf{z} \\
=& \int_{\mathbf{z}} \left(\mathbf{z}-\mathbf{z}_1\right)^t\left(\mathbf{z}-\mathbf{z}_1\right)\mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right) d\mathbf{z} \\
&+ \int_{\mathbf{z}} \mathbf{z}_1^t \mathbf{z}_1 \mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right) d\mathbf{z} \\
=& \mathbf{E}_{\mathcal{N}\left(\mathbf{z}|\mathbf{z}_1, \mathbf{\Lambda}^{-1}\right)}\left[\left(\mathbf{z}-\mathbf{z}_1\right)^2\right] + \mathbf{z}_1^t \mathbf{z}_1 \\
=& Tr\left(\mathbf{\Lambda}^{-1}\right) + \left\|\mathbf{z}_1\right\|^2
\end{aligned}$$

Hence, the second term is equal to :

$$\frac{1}{2}\left|\mathbf{\Sigma}_{\mathbf{w}_1}\right|^{-\frac{1}{2}}\left|\mathcal{Q}\right|^{-\frac{1}{2}}\left|\mathbf{\Lambda}\right|^{-\frac{1}{2}}\left(Tr\left(\mathbf{\Lambda}^{-1}\right) + \left\|\mathbf{z}_1\right\|^2\right) \tag{13}$$

Simplification:

$\left|\mathbf{\Lambda}\right|^{-\frac{1}{2}} = \left|\mathcal{Q}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{w}_1}^{-1}\mathcal{Q}^{-\frac{1}{2}}\right|^{-\frac{1}{2}} = \left|\mathbf{\Sigma}_{\mathbf{w}_1}^{-1}\mathcal{Q}^{-1}\right|^{-\frac{1}{2}} = \left|\mathbf{\Sigma}_{\mathbf{w}_1}\mathcal{Q}\right|^{\frac{1}{2}}$

$\mathbf{\Lambda}^{-1} = \left(\mathcal{Q}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{w}_1}^{-1}\mathcal{Q}^{-\frac{1}{2}}\right)^{-1} = \mathcal{Q}^{\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{w}_1}\mathcal{Q}^{\frac{1}{2}}$

thus :

$Tr\left(\mathbf{\Lambda}^{-1}\right) = Tr\left(\mathcal{Q}^{\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{w}_1}\mathcal{Q}^{\frac{1}{2}}\right) = Tr\left(\mathbf{\Sigma}_{\mathbf{w}_1}\mathcal{Q}\right)$

and :

$\left\|\mathbf{z}_1\right\|^2 = \left\|\mathcal{Q}^{\frac{1}{2}}\mu_{\mathbf{w}_1}\right\|^2 = \mu_{\mathbf{w}_1}^t\mathcal{Q}^{\frac{1}{2}}\mathcal{Q}^{\frac{1}{2}}\mu_{\mathbf{w}_1} = \mu_{\mathbf{w}_1}^t\mathcal{Q}\mu_{\mathbf{w}_1}$

The final formula of the second term is :

$$\frac{1}{2}\left(Tr\left(\mathbf{\Sigma}_{\mathbf{w}_1}\mathcal{Q}\right) + \mu_{\mathbf{w}_1}^t\mathcal{Q}\mu_{\mathbf{w}_1}\right)$$

[1] for simplicity reasons, we assume here that only one of the two i-vectors to be compared is noisy ($\mathbf{w}_1$). The extension to the case of two noisy i-vectors is trivial.

Using the three terms, the final score is :

$$\mu_{\mathbf{w}_1}^t \mathcal{P}\mathbf{w}_2 + \frac{1}{2}\left[Tr\left(\boldsymbol{\Sigma}_{\mathbf{w}_1}\mathcal{Q}\right) + \mu_{\mathbf{w}_1}^t \mathcal{Q}\mu_{\mathbf{w}_1}\right] + \frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2 \quad (14)$$

where $Tr\left(.\right)$ is the trace operator. In our experiments, we have ignored the term related to the a posteriori covariance $\boldsymbol{\Sigma}_{\mathbf{w}_1}$, the score used is:

$$\mu_{\mathbf{w}_1}^t \mathcal{P}\mathbf{w}_2 + \frac{1}{2}\left[\mu_{\mathbf{w}_1}^t \mathcal{Q}\mu_{\mathbf{w}_1}\right] + \frac{1}{2}\mathbf{w}_2^t \mathcal{Q}\mathbf{w}_2 \quad (15)$$

where $\mu_{\mathbf{w}_1}$ is given by Equation 6.

5. EXPERIMENTAL PROTOCOL

In this paper, we have presented a general framework to deal with additive noise in i-vector -based speaker recognition systems. In [23], we have proposed a method which is a particular case (theoretically) of what we have presented in this paper. The experimental results presented here are performed by assuming $\boldsymbol{\Sigma}_{\mathbf{w}_1} = 0$. In our future work, we will present experiments taking into account the mean and the covariance of the a posteriori pdf.

Our experiments operate on 19 Mel-Frequency Cepstral Coefficients (plus energy) augmented with 19 first (Δ) and 11 second ($\Delta\Delta$) derivatives. A mean and variance normalization (MVN) technique is applied on the MFCC features estimated using the speech portion of the audio file. The low-energy frames (corresponding mainly to silence) are removed.

A gender-dependent 512 diagonal component UBM (male model) and a total variability matrix of low rank 400 are estimated using 15660 utterances corresponding to 1147 speakers (using NIST SRE 2004, 2005, 2006 and Switchboard data). The LIA_SpkDet package of the LIA_RAL/ALIZE [27] toolkit is used for the estimation of the total variability matrix and the i-vector extraction. The algorithms used are described in [28]. Finally a two-covariance-based scoring [2] is applied. The equal-error rate (EER) over the NIST SRE 2008 male test data on the "short2/short3" task under the "det7" conditions [29] will be used as a reference to monitor the performance improvement compared to the baseline system in noisy conditions.

We use noise samples from the free sound repository FreeSound.org [30] as background noises. The open-source toolkit FaNT [31] was used to add these noises to the full waveforms generating new noisy audio files for each noise / SNR level. For each noisy condition, 500 train speech segments having SNR levels greater than $25dB$ and speech durations of nearly 2 minutes are used to estimate noise distributions in the i-vector space.

Table 1 and Table 2 show the performance of the baseline system and the proposed method. In all noisy experiments, enrollment data are clean while each test segment is affected by a certain noise on a fixed SNR level. In the "mixed noises" condition, test data are divided into 3 equal parts affected respectively by air-cooling, car-driving and crowd noise at a certain SNR level (indicated by the table column). We can see in Table 1 that the EER rate increases rapidly when SNR decreases. Comparing Table 1 and Table 2, we can see that the proposed method brings a relative improvement greater than 50%. A very important feature of this method is that when there is no noise the performance using the proposed approach does not degrade with respect to the baseline system (see column "clean").

Table 1. Baseline system performances (EER). The Target i-vectors are assumed to be clean.

		Test SNR		
	Clean	**0dB**	**5dB**	**10dB**
Air-cooling noise	1.59	26.85	15.21	9.51
Car driving noise	1.59	25.54	14.54	8.32
Crowd noise	1.59	24.24	13.94	7.77
Mixed noises	1.59	25.03	14.83	8.64

Table 2. Proposed system performances (EER). The Target i-vectors are assumed to be clean.

		Test SNR		
	Clean	**0dB**	**5dB**	**10dB**
Air-cooling noise	1.59	13.21	7.25	4.85
Car driving noise	1.59	12.05	6.65	3.78
Crowd noise	1.59	11.55	5.09	3.05
Mixed noises	1.59	12.25	6.84	4.32

6. CONCLUSION

In this paper we have described a statistical approach to deal with additive noise. This approach works in the i-vector space in which we assumed that clean and noisy i-vectors follow Gaussian pdf. In contrast to our previous work, the clean i-vector corresponding to the noisy one is considered to be a Gaussian pdf (called a posteriori pdf). In the scoring phase (with two-covariance), the score is estimated by integrating with respect to the obtained a posteriori pdf. The experimental results are very encouraging (greater than 50% relative gain). The calculation of an i-vector is mainly based on the a posteriori probabilities of Guassians given frames. In this work, these latter probabilities are calculated using a GMM-UBM trained on clean data frames. An interesting perspective is to adapt the UBM to the test noise to better estimate the a posteriori probabilities.

REFERENCES

[1] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[2] Niko Brümmer and Edward De Villiers, "The speaker partitioning problem.," in *Odyssey*, 2010, p. 34.

[3] A El-Solh, A Cuhadar, and RA Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on*. IEEE, 2007, pp. 235–239.

[4] Seyed Omid Sadjadi and John HL Hansen, "Assessment of single-channel speech enhancement techniques for

speaker identification under mismatched conditions.," in *INTERSPEECH*, 2010, pp. 2138–2141.

[5] Cemal Hanilçi, Tomi Kinnunen, Rahim Saeidi, Jouni Pohjalainen, Paavo Alku, Figen Ertas, Johan Sandberg, and Maria Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," in *ICASSP*, 2012, pp. 4769–4772.

[6] Yun Lei, Lukas Burget, and Nicolas Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *ICASSP*, 2013, pp. 6788–6791.

[7] Yun Lei, Mitchell McLaren, Luciana Ferrer, and Nicolas Scheffer, "Simplified vts-based i-vector extraction in noise-robust speaker recognition," *ICASSP, Florence, Italy*, 2014.

[8] David Martınez, Lukáš Burget, Themos Stafylakis, Yun Lei, Patrick Kenny, and Eduardo Lleida, "Unscented transform for ivector-based noisy speaker recognition," *ICASSP, Florence, Italy*, 2014.

[9] Pedro J Moreno, Bhiksha Raj, Evandro Gouvea, and Richard M Stern, "Multivariate-gaussian-based cepstral normalization for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. IEEE, 1995, vol. 1, pp. 137–140.

[10] Li Deng, Alex Acero, Mike Plumpe, and Xuedong Huang, "Large-vocabulary speech recognition under adverse acoustic environments.," in *INTERSPEECH*, 2000, pp. 806–809.

[11] Mohamed Afify, Xiaodong Cui, and Yuqing Gao, "Stereo-based stochastic mapping for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1325–1334, 2009.

[12] Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Stereo-based stochastic noise compensation based on trajectory gmms," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4577–4580.

[13] Sourjya Sarkar and K Sreenivasa Rao, "Stochastic feature compensation methods for speaker verification in noisy environments," *Applied Soft Computing*, vol. 19, pp. 198–214, 2014.

[14] MJF Gales and Steve J Young, "Hmm recognition in noise using parallel model combination.," in *Eurospeech*, 1993, vol. 93, pp. 837–840.

[15] Olivier Bellot, Driss Matrouf, Teva Merlin, and Jean-François Bonastre, "Additive and convolutional noises compensation for speaker recognition.," in *INTERSPEECH*, 2000, pp. 799–802.

[16] Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *ICASSP*, 2012, pp. 4253–4256.

[17] Chao Weng, Dong Yu, Shinji Watanabe, and BH Juang, "Recurrent deep neural networks for robust speech recognition," *Proc. of ICASSP, Florence, Italy*, 2014.

[18] Jürgen T Geiger, Zixing Zhang, Felix Weninger, Björn Schuller, and Gerhard Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[19] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust asr.," in *INTERSPEECH*, 2012.

[20] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," .

[21] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[22] Mitchell McLaren, Yun Lei, Nicolas Scheffer, and Luciana Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[23] Waad Ben Kheder, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre, and Moez Ajili, "Robust speaker recognition using map estimation of additive noise in i-vectors space," in *Statistical Language and Speech Processing*, pp. 97–107. Springer, 2014.

[24] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[25] S.J.D. Prince, *Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012.

[26] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011, pp. 249–252.

[27] J-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, G. Pouchoulin, B. Fauve, and J. Mason, "Alize/spkdet: a state-of-the-art open source software for speaker recognition," in *ISCA/IEEE Speaker Odyssey, South Africa, January 2008*.

[28] Driss Matrouf, Nicolas Scheffer, Benoit GB Fauve, and Jean-François Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification.," in *INTERSPEECH*, 2007, pp. 1242–1245.

[29] "The NIST year 2008 speaker recognition evaluation plan," `http://www.itl.nist.gov/iad/mig//tests/sre/2008/`, 2008, [Online; accessed 15-May-2014].

[30] "Freesound.org," `http://www.freesound.org`.

[31] H. Guenter Hirsch, "FaNT - Filtering and Noise Adding Tool," `http://dnt.kr.hsnr.de/download.html`, [Online; accessed 15-May-2014].