UNSCENTED TRANSFORM FOR IVECTOR-BASED NOISY SPEAKER RECOGNITION

David Martínez[1], *Lukáš Burget*[2], *Themos Stafylakis*[3], *Yun Lei*[4], *Patrick Kenny*[3], *Eduardo Lleida*[1]

[1]Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2]Speech@FIT, Brno University of Technology, Czech Republic
[3]Centre de Recherche Informatique de Montreal (CRIM), Canada
[4]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

ABSTRACT

Recently, a new version of the iVector modelling has been proposed for noise robust speaker recognition, where the nonlinear function that relates clean and noisy cepstral coefficients is approximated by a first order vector Taylor series (VTS). In this paper, it is proposed to substitute the first order VTS by an unscented transform, where unlike VTS, the nonlinear function is not applied over the clean model parameters directly, but over a set of sampled points. The resulting points in the transformed space are then used to calculate the model parameters. For very low signal-to-noise ratio improvements in equal error rate of about 7% for a *clean* backend and of 14.50% for a *multistyle* backend are obtained.

Index Terms— Noise Robust Speaker Recognition, Unscented Transform, Vector Taylor Series, iVector

1. INTRODUCTION

Speaker recognition is one of the most important research fields in the speech technology industry. The main applications are found in banking, defense, forensics, video games, and also as front-end of other speech-related tasks like speech recognition. During the last decade, important technological advances have been achieved in this field. One important milestone was the development of the joint factor analysis (JFA) algorithm, a technique that makes possible to model simultaneously the inter- and intra-speaker variabilities of the features [1]. Currently, a new dimensionality reduction technique inspired by JFA is used, which allows representing a speech utterance by a low-dimension fixed length vector, or iVector, which is used for recognition [2]. The state-of-the-art recognizer is called probabilistic linear discriminant analysis (PLDA), and also allows modelling inter- and intra-speaker variability in the iVectors [3].

All these advances have brought a substantial improvement in performance and the researchers start to focus on other challenges. One important research direction is speaker recognition in noisy environments. This is not a new topic in speaker recognition [4, 5], but the interest currently lies in making the high-accuracy state-of-the-art JFA-based techniques robust to noise.

In [6], the authors present the PRISM evaluation set, a database to experiment speaker recognition systems under several noisy conditions with the aim of providing a common testbed to the community. They include language, channel, speech style, and vocal effort variabilities, also seen in NIST SRE evaluations, and other types not available on standard databases, like severe noise, and reverberation. In [7], a subset of this database is tested on different signal-to-noise ratios (SNR) and it is shown how the performance of a PLDA system modelling iVectors extracted from Mel-frequency cepstral coefficients (MFCC) is quickly degraded when the SNR decreases. It is observed that adding noisy data to the PLDA training gives relative improvements of up to 30% compared to the case where only clean data are used. The same behaviour is observed with prosodic features. By adding noisy data to train the iVector extractor no significant gains are obtained.

In [8], the authors propose a first order vector Taylor series (VTS) approximation [9] to extract noise-compensated iVectors. The approach is inspired by the VTS successfully applied in the field of automatic speech recognition (ASR) to compensate the models distorted by the nonlinear effects of noise in the cepstral domain [10, 11]. For the same PRISM subset as above, relative improvements of up to 80% compared to a state-of-the-art system with cepstral mean and variance normalization (CMVN) are observed for the speaker recognition problem, however the training process is very slow. To make it lighter, in [12] a simplified VTS (sVTS) version is proposed, where most of the improvement is kept, while the computational load is largely reduced.

In this work, the unscented transform (UT) is presented as a new approach to approximating the nonlinearity caused by noise in the cepstral domain in order to adapt the model parameters to noise. We compare UT to the first order VTS approximation. UT is a method to propagate the mean and covariance information through nonlinear tansformations [13]. It is more accurate, easier to implement, and in the same order of computational expense as the linearization used with VTS, and it has been already proven to be useful for noise robust ASR [14, 15]. As shown in the experimental part of the work, UT is especially useful for very low SNR, when the nonlinear distortion is stronger.

The rest of the paper is organized as follows: in section 2 a description of the iVector approach in noisy environments is given, together with the role of VTS and UT to approximate the nonlinear relationship between clean and noisy MFCC; in section 3 the experimental part of the work is shown; and in section 4 the conclusions are drawn.

2. UNSCENTED TRANSFORM AND VTS IN AN IVECTOR-BASED SYSTEM

2.1. Standard iVector System

In the standard iVector extraction process, it is assumed that the input features, in our case MFCCs, follow a Gaussian mixture model (GMM) distribution in which the mean vector of each Gaussian is assumed to be utterance-specific. Thus the MFCCs of utterance *i*, $\mathbf{x}^{(i)}$, are evenually modelled as

$$\mathbf{x}^{(i)} \sim \sum_k \pi_k \mathcal{N}(\mu_{x_k} + \mathbf{T}_k \omega^{(i)}, \mathbf{\Sigma}_{x_k}), \quad (1)$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

being π_k, μ_{x_k}, and $\mathbf{\Sigma}_{x_k}$, the weight, mean, and covariance, respectively, of Gaussian k of a pre-trained GMM, the universal background model (UBM), $\mathbf{T}_k$ a low-rank matrix spanning a subspace referred to as total variability subspace that describes intersession variability in the space of GMM mean parameters, and $\omega^{(i)}$ a segment-specific low-dimension latent variable with standard normal distributed prior.

The training of this model is performed via maximum likelihood (ML) in two parts. Firstly, the UBM is pre-trained using the expectation-maximization (EM) algorithm, and π_k, μ_{x_k}, and $\mathbf{\Sigma}_{x_k}$ are obtained for all the Gaussians. Secondly, the sufficient statistics are computed as defined in [2] using fixed Gaussian alignments given by the UBM, and they are used for the training of the $\mathbf{T}_k$ matrices, which is also performed with the EM algorithm [2].

The iVector of utterance *i* is defined as the maximum a posteriori (MAP) point estimate of $\omega^{(i)}$. The posterior probability distribution of $\omega^{(i)}$ is Gaussian with mean, $\langle\omega^{(i)}\rangle$, and covariance, $\mathbf{L}^{(i)}$, and thus the iVector is equal to $\langle\omega^{(i)}\rangle$. The expressions to compute it are

$$\langle\omega^{(i)}\rangle = \mathbf{L}^{(i)} \sum_k \tilde{\mathbf{T}}_k^T \tilde{\mathbf{f}}_k^{(i)} \tag{2}$$

$$\mathbf{L}^{(i)} = (I + \sum_k N_{xk}^{(i)} \tilde{\mathbf{T}}_k^T \tilde{\mathbf{T}}_k)^{-1} \tag{3}$$

where $\mathbf{\Sigma}_{xk} = \mathbf{P}_{xk}\mathbf{P}_{xk}^T$, with $\mathbf{P}_{xk}$ lower triangular by Cholesky decomposition, $\tilde{\mathbf{T}}_k = \mathbf{P}_{xk}^{-1}\mathbf{T}_k$, and

$$N_{xk}^{(i)} = \sum_t \gamma_{xt}^{(i)}(k), \qquad \tilde{\mathbf{f}}_k^{(i)} = \mathbf{P}_{xk}^{-1} \sum_t \gamma_{xt}^{(i)}(k)(\mathbf{x}_t^{(i)} - \mu_{xk}) \tag{4}$$

are the zeroth and *whitened* first order sufficient statistics precollected using the UBM as proposed in [16]. The first order statistic *whitening* ($\mu_{xk}^{(i)}$ subtraction and multiplication by $\mathbf{P}_{xk}^{-1}$) not only leads to a more efficient implementation, but it also plays an important role in the sVTS approach described in section 2.3.

2.2. VTS-Based iVector System for Noisy Environments

According to the model of the environment presented in [9], a clean MFCC vector affected by additive and convolutional noise is distorted as

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + g(\mathbf{n} - \mathbf{x} - \mathbf{h}), \tag{5}$$

where $\mathbf{y}$, $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{n}$ are the cepstral vectors of the noisy speech, clean speech, channel, and additive noise, respectively, and g is the nonlinear function defined as

$$g = \mathbf{C}\ln(1 + exp(\mathbf{C}^\dagger(\mathbf{n} - \mathbf{x} - \mathbf{h}))), \tag{6}$$

with $\mathbf{C}$ and $\mathbf{C}^\dagger$ the discrete cosine transform matrix and its pseudo-inverse, respectively. The corresponding relationship in the model space for the UBM means [11], assuming that both types of noise follow a Gaussian distribution, is approximated by a first order VTS expansion at ($\mu_{\mathbf{x_k}0}$, $\mu_{\mathbf{h}0}$, $\mu_{\mathbf{n}0}$),

$$\begin{aligned}\mu_{y_k}^{(i)} \approx \mu_{x_k 0} + \mu_{h0}^{(i)} + g(\mu_{n0}^{(i)} - \mu_{x_k 0} - \mu_{h0}^{(i)}) \\ + \mathbf{G}_k^{(i)}(\mu_{x_k} - \mu_{x_k 0}) + \mathbf{G}_k^{(i)}(\mu_h^{(i)} - \mu_{h0}^{(i)}) \\ + \mathbf{F}_k^{(i)}(\mu_n^{(i)} - \mu_{n0}^{(i)}),\end{aligned} \tag{7}$$

where $\mathbf{G}_k$ is the Jacobian of g with respect to $\mathbf{x_k}$, and with respect to $\mathbf{h}$, and $\mathbf{F}_k$ with respect to $\mathbf{n}$. They are defined as

$$\mathbf{G}_k^{(i)} = \mathbf{C} \cdot diag(\frac{1}{1 + exp(\mathbf{C}^\dagger(\mu_{n0}^{(i)} - \mu_{x_k} - \mu_{h0}^{(i)}))}) \cdot \mathbf{C}^\dagger, \tag{8}$$

$$\mathbf{F}_k^{(i)} = \mathcal{I} - \mathbf{G}_k^{(i)}. \tag{9}$$

To compute the means of the noise-adapted UBM, $\mu_{\mathbf{y_k}\mathbf{0}}$, the VTS is evaluated at ($\mu_{\mathbf{x_k}} = \mu_{\mathbf{x_k}\mathbf{0}}$, $\mu_{\mathbf{h}} = \mu_{\mathbf{h}\mathbf{0}}$, $\mu_{\mathbf{n}} = \mu_{\mathbf{n}\mathbf{0}}$),

$$\mu_{y_k 0}^{(i)} \approx \mu_{x_k 0} + \mu_{h0}^{(i)} + g(\mu_{n0}^{(i)} - \mu_{x_k 0} - \mu_{h0}^{(i)}) \tag{10}$$

The relationship of the UBM covariances [11], following the same reasoning as for the mean, is

$$\mathbf{\Sigma}_{y_k} \approx \mathbf{G}_k^{(i)} \mathbf{\Sigma}_{x_k} \mathbf{G}_k^{(i)T} + \mathbf{F}_k^{(i)} \mathbf{\Sigma}_n^{(i)} \mathbf{F}_k^{(i)T}, \tag{11}$$

where $\mathbf{\Sigma}_n^{(i)}$ is the additive noise covariance matrix, and $\mathbf{\Sigma}_h^{(i)}$ is set to zero since the channel is considered to be fixed. Finally, the mean and covariance of the model for the noisy MFCC first derivative (Δ) are calculated with the continuous-time approximation also used in [11]. That is,

$$\mu_{\Delta y_k}^{(i)} \approx \mathbf{G}_k^{(i)} \mu_{\Delta x_k}^{(i)} \tag{12}$$

$$\mathbf{\Sigma}_{\Delta y_k}^{(i)} \approx \mathbf{G}_k^{(i)} \mathbf{\Sigma}_{\Delta x_k} \mathbf{G}_k^{(i)T} + \mathbf{F}_k^{(i)} \mathbf{\Sigma}_{\Delta n}^{(i)} \mathbf{F}_k^{(i)T}, \tag{13}$$

and identically for the MFCC second derivative (Δ^2), substituting Δ by Δ^2.

One important role of the VTS approximation is to make the EM objective function of the noise-adapted UBM differentiable, so closed form update formulae of the model parameters are obtained. As per [8] the objective function becomes

$$\begin{aligned}Q = \sum_i \sum_t \sum_k \gamma_{yt}^{(i)}(k)[-\frac{1}{2}\ln|\mathbf{\Sigma}_{y_k}^{(i)}| \\ -\frac{1}{2}(\mathbf{y}_t^{(i)} - \mu_{y_k 0}^{(i)})^T (\mathbf{\Sigma}_{y_k}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \mu_{y_k 0}^{(i)})],\end{aligned} \tag{14}$$

In order to include the total variability subspace in the model of the noisy MFCC of every utterance, $\mathbf{y}^{(i)}$, μ_{x_k} is substituted by $\mu_{x_k 0} + T_k\omega^{(i)}$ in (7), and also considering (11), it can be shown that

$$\mathbf{y}^{(i)} \sim \sum_k \pi_k \mathcal{N}(\mu_{y_k 0}^{(i)} + \mathbf{G}_k^{(i)} \mathbf{T}_k \omega^{(i)}, \mathbf{\Sigma}_{y_k}^{(i)}). \tag{15}$$

This model is trained using the EM algorithm and the equations are detailed in [8].

2.3. Simplified VTS

The major drawback of the VTS approach presented in previous section is the computational cost of the EM training algorithm for the total variability subspace $\mathbf{T}_k$ of (15). In particular, in the *M step* the computation of the Kronecker product and large matrix inversion given in equation (18) of [8] is several orders of magnitude more computationally and memory demanding than the calculations required for training the standard model of (1). The main differences between the two techniques are that in the VTS approach the UBM mean and covariance are utterance-dependent, and that the total variability subspace is adapted to noise differently for each utterance through the term $\mathbf{G}_k^{(i)}\mathbf{T}_k$ in (15).

In [12], a new approach is proposed that largely simplifies the equations and reduces the computational cost, the sVTS. In the sVTS, first, the UBM is adapted to each file as described in section 2.2. Then, the zeroth and *whitened* first order sufficient statistics of utterance *i* are collected over its noise-adapted UBM as

$$N_{yk}^{(i)} = \sum_t \gamma_{yt}^{(i)}(k), \qquad \tilde{\mathbf{f}}_{yk}^{(i)} = \mathbf{P}_{yk}^{(i)-1} \sum_t \gamma_{yt}^{(i)}(k)(\mathbf{y}_t^{(i)} - \mu_{y_k}^{(i)}), \tag{16}$$

where $\mathbf{\Sigma}_{y_k}^{(i)} = \mathbf{P}_{yk}^{(i)}\mathbf{P}_{yk}^{(i)T}$ by Cholesky decomposition. In this way the dependence on $\mu_{y_k}^{(i)}$, $\mathbf{\Sigma}_{y_k}^{(i)}$, and $\mathbf{G}_k^{(i)}$ completely disappears from the training, and the equations, and therefore the complexity, are reduced to the ones of the standard iVector training algorithm. Also for iVector extraction equations (2) and (3) can still be used, but replacing the sufficient statistics defined in (4) with those defined in (16). This transformation of the sufficient statistics moves the noise compensation operation to the domain of the sufficient statistics, while the former VTS approach introduced in [8] is a model domain compensation technique. In spite of the complexity reduction, the experiments made in [12] show that the sVTS preserve most of the improvements obtained with the VTS-based iVector model.

2.4. Simplified Unscented Transform

The UT is used to substitute the first order VTS in the model parameter adaptation. The goal is to obtain more accurate estimates of $\mu_{y_k}^{(i)}$ and $\mathbf{\Sigma}_{y_k}^{(i)}$ when the linear approximation is not good enough. The first UT method explained in [14] is followed here. Given the clean and noisy mean cepstral estimates, $\mu_{x_k 0}$ and $\mu_n^{(i)}$, an augmented signal $\hat{\mathbf{s}}_k^{(i)} = [\hat{\mathbf{x}}_k^T \;\; \hat{\mathbf{n}}^{(i)T}]^T$ is built by sampling as

$$\begin{aligned}
\hat{\mathbf{s}}_{k0}^{(i)} &= [\mu_{x_k 0}^T \;\; \mu_n^{(i)T}]^T \\
\hat{\mathbf{s}}_{kj}^{(i)} &= [\mu_{x_k 0}^T + (\sqrt{2D\mathbf{\Sigma}_{x_k}})_j \;\; \mu_n^{(i)T}]^T \\
\hat{\mathbf{s}}_{k(j+D)}^{(i)} &= [\mu_{x_k 0}^T - (\sqrt{2D\mathbf{\Sigma}_{x_k}})_j \;\; \mu_n^{(i)T}]^T \\
\hat{\mathbf{s}}_{k(j+2D)}^{(i)} &= [\mu_{x_k 0}^T \;\; \mu_n^{(i)T} + (\sqrt{2D\mathbf{\Sigma}_n^{(i)}})_j]^T \\
\hat{\mathbf{s}}_{k(j+3D)}^{(i)} &= [\mu_{x_k 0}^T \;\; \mu_n^{(i)T} - (\sqrt{2D\mathbf{\Sigma}_n^{(i)}})_j]^T
\end{aligned} \tag{17}$$

where D is the feature dimension, $j = 1...D$, therefore $\hat{\mathbf{s}}_k^{(i)}$ contains 4D+1 2D-dimension sampled vectors, and $(A)_j$ denotes the j*th* column of matrix A. Observe that the means and covariance matrices calculated from these samples match the actual means and covariances from which the samples were derived. Next, the sampled points are transformed using the nonlinear function

$$(f(\hat{\mathbf{s}}_k^{(i)}))_j = (\hat{\mathbf{y}}_k^{(i)})_j = (\hat{\mathbf{x}}_k)_j + \mu_{h0}^{(i)} + g((\hat{\mathbf{n}}^{(i)})_j - (\hat{\mathbf{x}}_k)_j - \mu_{h0}^{(i)}) \tag{18}$$

to obtain the noisy version of the sampled points. The mean and covariance of the noise-adapted UBM are the mean and covariance of the 4D+1 D-dimension vectors $\hat{\mathbf{y}}_k^{(i)}$, respectively,

$$\hat{\mu}_{y_k}^{(i)} = \frac{\sum_{j=0}^{4D} (\hat{\mathbf{y}}_k^{(i)})_j}{4D+1}, \tag{19}$$

$$\hat{\mathbf{\Sigma}}_{y_k}^{(i)} = \frac{\sum_{j=0}^{4D} ((\hat{\mathbf{y}}_k^{(i)})_j - \hat{\mu}_{y_k}^{(i)})((\hat{\mathbf{y}}_k^{(i)})_j - \hat{\mu}_{y_k}^{(i)})^T}{4D+1}. \tag{20}$$

Likewise, the Jacobians $\mathbf{G}_k$ and $\mathbf{F}_k$, used in the update formulae of the noise parameters and in the continuous-time approximation of the Δ and Δ^2 model parameters, also depend on the sampled points and are calculated as

$$\hat{\mathbf{G}}_k^{(i)} = \frac{\sum_{j=0}^{4D} \mathbf{C} \cdot diag(\frac{1}{1+exp(\mathbf{C}^\dagger \cdot ((\hat{\mathbf{n}}^{(i)})_j - (\hat{\mathbf{x}}_k)_j - \mu_{h0}^{(i)}))}) \cdot \mathbf{C}^\dagger}{4D+1} \tag{21}$$

$$\hat{\mathbf{F}}_k^{(i)} = \mathcal{I} - \hat{\mathbf{G}}_k^{(i)} \tag{22}$$

Once the noise-adapted UBM mean and covariance, and the Jacobians are estimated, the rest of the training is exactly the same as for the VTS case. To avoid the computational complexity of the exact noise-compensated iVector extraction presented before, the simplified version is also used with the UT. Hence, this approach is named simplified UT (sUT). Note that the augmented signal contains information only of the cepstrum and not of the derivatives. These are derived through the Jacobian $\hat{\mathbf{G}}_k^{(i)}$ as per (12) and (13).

3. EXPERIMENTAL PART

Our features are 20 MFCC coefficients (with C0) including first and second derivatives, extracted in 25 ms long windows every 10 ms. A diagonal UBM with 512 components is trained with data coming from NIST SRE '04, '05, '06, and '08 evaluations. The 400-dimension iVector extractor is trained with data coming from NIST SRE '04, '05, '06, '08, Fisher, and Switchboard. A simplified PLDA (sPLDA) [17] with 200-dimension speaker factors is trained with the same dataset as the iVector extractor. Previously the iVectors are centered, whitened, and length-normalized [18]. Two training methods are tested for sPLDA, the *clean*, where only clean data are used, and the *multistyle*, where noisy data of 20, 15, and 8 dB are also included. The enrollment and test data is the same subset of the PRISM dataset used in [7, 8, 12]. It includes additive noise from different scenarios at three different SNRs of 20, 15, and 8 dB. Experiments are reported in terms of equal error rate (EER) and minimum of decision cost function (minDCF) as defined in [19], only on females. The SNR in enrollment and test is always the same.

In our approach, mean updates of the noise parametes **n** and **h** are obtained in the odd iterations of the EM algorithm, while the covariance update of **n** is obtained in even iterations. The reason to do it in this way is that the covariance update depends on the mean update. We have swept over several number of iterations for noise-adapted UBM training to find optimal performance. The results are obtained for the first iteration, in which only means are updated, and then every other iteration, in order to complete full updates of means and covariance.

In tables 1 and 2, the results of four different systems are compared for the *clean* sPLDA and the *multistyle* sPLDA. They are a system without noise compensation, a system with the same iVector configuration and CMVN, an sVTS system, and an sUT system. Some interesting conclusions can be found in the results. First, the *multistyle* sPLDA gives better performance than the *clean* sPLDA, as already observed in [8, 12]. Second, both the sVTS and the sUT techniques outperform CMVN, and of course, the case without noise robustness. For sVTS, iteration 3 seems to be optimal for both the *clean* and *multistyle* sPLDA. The reader should note that in every iteration the utterance-dependent log-likelihood (LLK) function of the noise-adapted UBM is increased, but this increase in LLK does not guarantee an increase in the recognition performance. We believe that more than 3 iterations overfit the data and the updates stop being useful. On the other hand, for sUT more iterations seem to be more useful. With the *clean* sPLDA iteration 7 seems to be optimal for all SNRs. For the case with *multistyle* sPLDA, the addition of noisy data in the sPLDA training makes the training to converge faster, and the best results are obtained with 3 iterations, except for the case of 8 dBs, for which the best results are obtained in iteration 5. The sUT gives better performance than the sVTS in the noisiest case, with an SNR of 8dBs. Recall that UT is an alternative to better model nonlinear distortions in the MFCC domain caused by noise, and thus, the higher the noise level, the larger the nonlinear effect, the worse the first order VTS approximation, and the larger the benefit obtained with sUT. In terms of EER and for SNR=8 dBs, with the *clean* sPLDA a 6.89% relative improvement is obtained with sUT

	EER(100%)				minDCF10			
SNR	clean	20 dB	15 dB	8 dB	clean	20 dB	15 dB	8 dB
No Robust	1.059	4.179	14.008	22.135	0.249	0.489	0.859	0.946
CMVN	0.772	2.143	3.167	7.750	0.182	0.317	0.488	0.717
sVTS it 1	0.851	1.864	3.029	7.262	0.197	0.286	0.451	0.728
sVTS it 3	0.912	1.591	2.607	6.689	0.172	0.252	0.409	0.659
sVTS it 5	0.842	1.765	2.696	6.478	0.180	0.284	0.415	0.697
sVTS it 7	0.788	1.809	2.594	6.357	0.190	0.298	0.412	0.693
sUT it 1	0.811	2.093	3.343	8.120	0.191	0.310	0.455	0.714
sUT it 3	0.712	1.956	3.189	6.805	0.154	0.323	0.466	0.699
sUT it 5	0.971	1.978	2.899	6.279	0.182	0.322	0.444	0.728
sUT it 7	0.970	1.877	2.819	5.919	0.190	0.304	0.423	0.682

Table 1. *Results for the clean sPLDA*

	EER(100%)				minDCF10			
SNR	clean	20 dB	15 dB	8 dB	clean	20 dB	15 dB	8 dB
No Robust	0.802	1.994	10.296	11.942	0.216	0.327	0.791	0.970
CMVN	0.694	1.786	2.304	4.261	0.177	0.278	0.381	0.635
sVTS it 1	0.859	1.521	2.261	4.459	0.182	0.245	0.319	0.583
sVTS it 3	0.846	1.447	1.918	4.292	0.169	0.233	0.338	0.584
sVTS it 5	0.794	1.673	2.104	4.450	0.179	0.275	0.388	0.626
sVTS it 7	0.848	1.790	2.281	4.514	0.184	0.276	0.388	0.627
sUT it 1	0.844	1.564	2.311	4.284	0.191	0.263	0.334	0.573
sUT it 3	0.717	1.412	1.940	4.087	0.155	0.241	0.327	0.568
sUT it 5	0.879	1.675	1.975	3.670	0.148	0.250	0.306	0.556
sUT it 7	0.932	1.639	2.074	3.708	0.180	0.260	0.320	0.582

Table 2. *Results for the multistyle sPLDA*

over sVTS, and in the *multistyle* case the relative improvement is of 14.50%, taking in both cases the optimal iterations of each technique. As final remark, note that the sVTS results are slightly different to the ones published in [12] because the feature extraction is different, and because in this work the VAD of noisy files is computed with the noisy speech, whereas there it was computed from the clean signal.

4. CONCLUSIONS

In this paper, the UT is presented for a speaker recognition task as an alternative to the first order VTS to approximate the nonlinearities caused by noise in the model space. The UT samples in the clean space, transforms the sampled features with the nonlinear function that relates clean and noisy MFCCs, and obtains the mean and covariances of the noise-adapted UBM in the transformed space. Unlike first order VTS, which is a linear approximation, the UT is expected to be more accurate when the distortions are far from being locally linear. The results show improvements for very low SNRs. In terms of EER, a 6.89% relative improvement is obtained for a sPLDA trained with only clean speech, and a 14.50% for a sPLDA trained with clean and noisy speech. To avoid the high computational load of the iVector modelling in the proposed noisy environment, a simplified version is followed, where the sufficient statistics are normalized with their corresponding utterance-dependent noise-adapted UBM. Finally, it is also concluded that the noise-adapted UBM calculation converges faster in sVTS than in sUT.

5. ACKNOWLEDGEMENTS

This work was developed thanks to the ideas shared during the workshops Bosaris 2012 and JHU 2013. Thanks to the organizers and all the participants. Thanks to Niko Brümmer for providing optimization functions, and to Olda Plchot for organizing all the data.

David Martínez was funded by the Spanish Government and the European Union (FEDER) under project TIN2011-28169-C05-02.

Lukáš Burget was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015, by Technology Agency of the Czech Republic grant No. TA01011328 and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

6. REFERENCES

[1] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[2] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] Simon Prince and James Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[4] Tomoko Matsui, Tomohito Kanno, and Sadaoki Furui, "Speaker Recognition Using HMM Composition in Noisy Environments," *Computer Speech & Language*, vol. 10, no. 2, pp. 107–116, 1996.

[5] Ji Ming, Timothy Hazen, James Glass, and Douglas Reynolds, "Robust Speaker Recognition in Noisy Conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[6] Luciana Ferrer, Harry Bratt, Lukáš Burget, Jan Černocký, Ondrej Glembek, Martin Graciarena, Aaron Lawson, Yun Lei, Pavel Matějka, Oldich Plchot, and Nicolas Scheffer, "Promoting Robustness for Speaker Modeling in the Community: the PRISM Evaluation Set," in *NIST Workshop*, Atlanta, GE, USA, 2011.

[7] Yun Lei, Lukáš Burget, and Luciana Ferrer, "Towards Noise-Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis," in *ICASSP*, Kyoto, Japan, 2012, vol. 2, pp. 4253 – 4256.

[8] Yun Lei, Lukáš Burget, and Nicolas Scheffer, "A Noise Robust iVector Extractor Using Vector Taylor Series for Speaker Recognition," in *ICASSP*, Vancouver, BC, Canada, 2013, pp. 6788 – 6791.

[9] Pedro Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[10] Alex Acero, Li Deng, Trausti Kristjansson, and Jerry Zhang, "HMM Adaptation Using Vector Taylor Series for Noisy Speech Recognition," in *ICSLP*, Beijing, China, 2000, vol. 2, pp. 869–872.

[11] Ozlem Kalinli, Michael Seltzer, Jasha Droppo, and Alex Acero, "Noise Adaptive Training for Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1889–1901, 2010.

[12] Yun Lei, Mitchell Mclaren, Luciana Ferrer, and Nicolas Scheffer, "Simplified VTS-Based i-Vector Extraction in Noise-Robust Speaker Recognition," in *(submitted to) ICASSP*, Florence, Italy, 2014.

[13] Simon Julier and Jeffrey Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[14] Yu Hu and Qiang Huo, "An HMM Compensation Approach Using Unscented Transformation for Noisy Speech Recognition," in *ISCSLP*, Singapore, 2006, pp. 346–357, Springer Berlin Heidelberg.

[15] Jinyu Li, Dong Yu, Yifan Gong, and Li Deng, "Unscented Transform with Online Distortion Estimation for HMM Adaptation.," in *Interspeech*, Makuhari, Japan, 2010.

[16] Ondrej Glembek, Lukáš Burget, Pavel Matějka, Martin Karafiat, and Patrick Kenny, "Simplification and Optimization of iVector Extraction," in *ICASSP*, Prague, Czech Republic, 2011, number c, pp. 4516–4519.

[17] Jesús Villalba and Eduardo Lleida, "Handling iVectors from Different Recording Conditions Using Multi-Channel Simplified PLDA in Speaker Recognition," in *ICASSP*, Vancouver, BC, Canada, 2013, pp. 6763–6767.

[18] Daniel Garcia-Romero and Carol Espy-Wilson, "Analysis of i-Vector Length Normalization in Speaker Recognition Systems.," in *Interspeech*, Florence, Italy, 2011, pp. 249–252.

[19] Alvin Martin and Craig Greenberg, "The NIST 2010 Speaker Recognition Evaluation," in *Interspeech*, Makuhari, Japan, 2010, number September, pp. 2726–2729.