A NOISE ROBUST I-VECTOR EXTRACTOR USING VECTOR TAYLOR SERIES FOR SPEAKER RECOGNITION

Yun Lei ⋆ *Lukáš Burget* † *Nicolas Scheffer* ⋆

⋆ Speech Technology and Research Laboratory, SRI International, California, USA
{yunlei,scheffer}@speech.sri.com
† Brno University of Technology, Czech Republic
burget@fit.vutbr.cz

ABSTRACT

We propose a novel approach for noise-robust speaker recognition, where the model of distortions caused by additive and convolutive noises is integrated into the i-vector extraction framework. The model is based on a vector taylor series (VTS) approximation widely successful in noise robust speech recognition. The model allows for extracting ”cleaned-up” i-vectors which can be used in a standard i-vector back end. We evaluate the proposed framework on the PRISM corpus, a NIST-SRE like corpus, where noisy conditions were created by artificially adding babble noises to clean speech segments. Results show that using VTS i-vectors present significant improvements in all noisy conditions compared to a state-of-the-art baseline speaker recognition. More importantly, the proposed framework is robust to noise, as improvements are maintained when the system is trained on clean data.

Index Terms— speaker recognition, Vector Taylor Series, i-vector, noisy speaker verification, noise compensation

1. INTRODUCTION

Recently, the speaker verification community has seen a significant increase in accuracy from the successful application of the i-vector extraction paradigm [1]. Along with a Bayesian back-end such as probabilistic linear discriminant analysis (PLDA) [2, 3, 4], it has become the state of the art in speaker verification. In this framework, each speech utterance with variable duration is projected into an i-vector – a single low-dimensional feature vector, typically of a few hundred components. More specifically, an i-vector is a point estimate of a latent variable vector representing a Gaussian mixture model (GMM) adapted to the corresponding utterance. A PLDA model is then used to compare i-vectors representing different utterances and to produce verification scores.

This work is focused on the robustness of speaker verification systems in the presence of noisy speech. With recent widespread use of speech-enabled services for consumers and growing importance of speaker recognition in security and defence, the need for noise-robust techniques is on the rise. Although current state-of-the-art speaker recognition systems achieve very high performance on clean data, there are few studies of noisy conditions. In a previous study [5], we have successfully proposed a robust strategy to compensate for degradations from noise by adopting a multi-style training approach for the PLDA backend. While significant improvements were obtained, worse performance by an order of magnitude are still observed when comparing clean to degraded conditions. In this work, we propose to tackle the problem at an earlier stage, where the i-vector extractor explicitly takes into account the potential degradations in the speech data.

Our approach is inspired by a successful acoustic modeling technique for noise robust automatic speech recognition (ASR) [6, 7], where a VTS approximation is used to model non-linear distortions in the mel-cepstral domain caused by both additive and convolutive noise. In ASR, the VTS approximation is used to synthesize acoustic model of noisy speech from a given clean speech model and from estimated noise distributions. Results observed in [7, 8, 9] show that a significant improvement can be obtained from the VTS approach in noisy environments.

In contrast to ASR, where VTS is used to synthesize noisy model, we use the approach in a somewhat opposite manner where our goal is to obtain a clean version of an i-vector. In our work, VTS is used to decompose the GMM adapted to a noisy speech segment into i) a clean GMM represented by ”clean” i-vector and ii) the distributions of the noise. One of the main benefit is that the resulting i-vector can be used in a standard PLDA backend.

It is worth to point out the similarity between our technique and joint factor analysis (JFA) [10], where the low-dimensional GMM representation is also decomposed into speaker and channel factors. However, the channel factors, which are responsible for modeling the unwanted variability (such as additive and convolutive noise), can only model linear additive effects in the GMM mean supervector domain. In contrast, our technique considers highly non-linear effects that an additive noise has on GMM all parameters (both means and covariances). Moreover, our noise compensation technique is integrated into the more modern i-vector framework, which has been shown to be superior to JFA [1].

2. UBM ADAPTATION USING VTS

The first step of the standard i-vector extraction is to compute the zero and first order sufficient statistics for a universal background model (UBM). In our approach, the sufficient statistics are collected from a noisy UBM synthesized for each speech segment using the VTS based distortion model from the UBM trained on clean data and from the additive and convolutive noise distributions. Such VTS noise adaptation is essentially the same as the one in noise robust ASR [7] for HMM models.

We first present the formulas for adapting the UBM to noisy

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the authors and do not necessarily represent the views of the funding agencies. This work was done while Lukáš Burget was at SRI International.

speech while assuming known distributions of the additive and convolutive noise. We then derive the expectation-maximization (EM) algorithm to estimate the noise distribution directly from the speech segments. More detailed discussion and derivation of the presented formulas can be found in [8].

2.1. UBM adaptation to noisy speech

The VTS approach is based on the knowledge of the speech feature extraction process. Here the mel-frequency cepstrum coefficient (MFCC) features are used to derive the adaptation formulas. In the cepstrum extraction process, the noisy speech y can be modeled as

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + g(\mathbf{n} - \mathbf{x} - \mathbf{h}), \tag{1}$$

where $\mathbf{y}, \mathbf{x}, \mathbf{h}, \mathbf{n}$ are the cepstrum vectors corresponding to the noisy speech, clean speech, channel, and additive noise, respectively. The non-linear function g is:

$$g(\mathbf{n} - \mathbf{x} - \mathbf{h}) = \mathbf{C}\log(1 + \exp(\mathbf{C}^{\dagger}(\mathbf{n} - \mathbf{x} - \mathbf{h}))), \tag{2}$$

where $\mathbf{C}$ is the discrete cosine transform (DCT) matrix and $\mathbf{C}^{\dagger}$ is its pseudo-inverse.

Assuming simple Gaussian distributions for both additive and convolutive noise, the mean vector of the m-th component of the noise adapted UBM can be approximated using a VTS expansion at $(\boldsymbol{\mu}_{x_m 0}, \boldsymbol{\mu}_{n0}, \boldsymbol{\mu}_{h0})$ as

$$\begin{aligned} \boldsymbol{\mu}_{y_m} \approx\ & \boldsymbol{\mu}_{x_m 0} + \boldsymbol{\mu}_{h0} + g(\boldsymbol{\mu}_{n0} - \boldsymbol{\mu}_{x_m 0} - \boldsymbol{\mu}_{h0}) \\ & + \mathbf{G}_m(\boldsymbol{\mu}_{x_m} - \boldsymbol{\mu}_{x_m 0}) + \mathbf{G}_m(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0}) \\ & + \mathbf{F}_m(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n0}), \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}_{x_m}$ is the mean of the corresponding component in the clean UBM, $\boldsymbol{\mu}_n$ and $\boldsymbol{\mu}_h$ are the means of the additive and convolutive noise distributions, respectively. $\mathbf{G}_m$ and $\mathbf{F}_m$ are defined as:

$$\mathbf{G}_m = \mathbf{C} \cdot \mathrm{diag}\left(\frac{1}{1 + \exp(\mathbf{C}^{\dagger}(\boldsymbol{\mu}_{n0} - \boldsymbol{\mu}_{x_m 0} - \boldsymbol{\mu}_{h0}))}\right) \cdot \mathbf{C}^{\dagger} \tag{4}$$

$$\mathbf{F}_m = \mathbf{I} - \mathbf{G}_m. \tag{5}$$

To synthesize the noisy UBM, the VTS expansion is done at the point ($\boldsymbol{\mu}_{x_m 0} = \boldsymbol{\mu}_{x_m}, \boldsymbol{\mu}_{n0} = \boldsymbol{\mu}_n, \boldsymbol{\mu}_{h0} = \boldsymbol{\mu}_h$), which reduces (3) to

$$\boldsymbol{\mu}_{y_m 0} \approx \boldsymbol{\mu}_{x_m 0} + \boldsymbol{\mu}_{h0} + g(\boldsymbol{\mu}_{n0} - \boldsymbol{\mu}_{x_m 0} - \boldsymbol{\mu}_{h0}). \tag{6}$$

The more general formula (3) is nevertheless useful for the the following derivations.

The noise-adapted covariance matrix can be approximated as

$$\boldsymbol{\Sigma}_{y_m} \approx \mathbf{G}_m \boldsymbol{\Sigma}_{x_m} \mathbf{G}_m^T + \mathbf{F}_m \boldsymbol{\Sigma}_n \mathbf{F}_m^T, \tag{7}$$

where $\boldsymbol{\Sigma}_{x_m}$ is covariance matrix of m-th Gaussian component from the clean UBM, $\boldsymbol{\Sigma}_n$ is the additive noise covariance matrix and $\boldsymbol{\Sigma}_h$ is set to zero since the channel is usually considered to be fixed.

In addition, the first and second order derivatives (Δ and Δ^2) of the MFCC features are commonly used for speaker recognition. The means and covariances of these dynamic features can be approximated as

$$\boldsymbol{\mu}_{\Delta y_m} \approx \mathbf{G}_m \boldsymbol{\mu}_{\Delta x_m} \tag{8}$$

$$\boldsymbol{\Sigma}_{\Delta y_m} \approx \mathbf{G}_m \boldsymbol{\Sigma}_{\Delta x_m} \mathbf{G}_m^T + \mathbf{F}_m \boldsymbol{\Sigma}_{\Delta n} \mathbf{F}_m^T, \tag{9}$$

where we assume the noise to be stationary so that $\boldsymbol{\mu}_{\Delta n}$ and $\boldsymbol{\mu}_{\Delta h}$ are set to zero for simplicity.

2.2. Noise model estimation

For each utterance, we initialize our noise models using estimates from non-speech portions of the signal. Both additive and convolutive noise models are further updated using several EM iterations to better fit the noise adapted UBM to the noisy speech. The EM auxiliary function can be written as

$$\begin{aligned} Q = & \sum_i \sum_t \sum_m \gamma_{mt}^{(i)} \Big[-\frac{1}{2} \log \left| \boldsymbol{\Sigma}_{y_m}^{(i)} \right| \\ & -\frac{1}{2} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)}) \Big], \end{aligned} \tag{10}$$

where $\gamma_{mt}^{(i)}$ is the posterior probability that the component m from the current noise-adapted UBM generated the frame t from speech segment i. Substituting (3) into (10) and solving for noise means by maximizing the EM auxiliary function gives us the following updates:

$$\begin{aligned} \boldsymbol{\mu}_n^{(i)} = & \boldsymbol{\mu}_{n0}^{(i)} + \left\{ \sum_{t,m} \gamma_{mt}^{(i)} (\mathbf{F}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} \mathbf{F}_m^{(i)} \right\}^{-1} \\ \times & \left\{ \sum_{t,m} \gamma_{mt}^{(i)} (\mathbf{F}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m 0}^{(i)}) \right\} \end{aligned} \tag{11}$$

$$\begin{aligned} \boldsymbol{\mu}_h^{(i)} = & \boldsymbol{\mu}_{h0}^{(i)} + \left\{ \sum_{t,m} \gamma_{mt}^{(i)} (\mathbf{G}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} \mathbf{G}_m^{(i)} \right\}^{-1} \\ \times & \left\{ \sum_{t,m} \gamma_{mt}^{(i)} (\mathbf{G}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m 0}^{(i)}) \right\}, \end{aligned} \tag{12}$$

where $\boldsymbol{\mu}_{y_m 0}^{(i)}$ is given by (6) and symbols with subscript 0 corresponds to the current estimates of the parameters. In ASR, the covariance matrix $\boldsymbol{\Sigma}_n$ is usually diagonalized for efficiency. In our work, however, all covariance matrices, including those in UBM, are full. Since there is no closed-form solution to estimate $\boldsymbol{\Sigma}_n$, we use the L-BFGS-B algorithm [11] to maximize the Q function. For convenience, $\boldsymbol{\Sigma}_n$ is represented using its Cholesky decomposition to assure positive-definiteness of the covariance matrix during the optimization process:

$$\boldsymbol{\Sigma}_n^{(i)} = \mathbf{U}_n^{(i)T} \mathbf{U}_n^{(i)}, \tag{13}$$

where $\mathbf{U}_n^{(i)}$ is the upper triangle matrix. The gradient of the auxiliary function (10) w.r.t. $\mathbf{U}_n^{(i)}$ is:

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{U}_n^{(i)}} = & \frac{\partial}{\partial \mathbf{U}_n^{(i)}} \sum_t \sum_m \gamma_{mt}^{(i)} \Big[-\frac{1}{2} \log \left| \boldsymbol{\Sigma}_{y_m}^{(i)} \right| \\ & -\frac{1}{2} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)}) \Big] \\ = & \sum_t \sum_m \gamma_{mt}^{(i)} \Big[-\mathbf{U}_n^{(i)} (\mathbf{F}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} \mathbf{F}_m^{(i)} \\ & + \mathbf{U}_n^{(i)} (\mathbf{F}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)}) \\ & \times (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} \mathbf{F}_m^{(i)} \Big]. \end{aligned}$$

For the dynamic features, the covariance matrices (e.g., $\boldsymbol{\Sigma}_{\Delta n}$ and $\boldsymbol{\Sigma}_{\Delta^2 n}$) can be estimated in a similar way. From these equations, we observe that the updates for the means and covariance matrices are not independent. Therefore, we alternate the means and covariance updates where the posteriors $\gamma_{mt}^{(i)}$ are recalculated.

3. NOISE COMPENSATED I-VECTOR EXTRACTION

We lay out the new i-vector framework that fits with the VTS compensation scheme proposed earlier. In the standard i-vector framework, (clean) speech frames $\mathbf{x}^{(i)}$ from i-th speech segment are assumed to be generated from a GMM:

$$\begin{aligned} \mathbf{x}^{(i)} &\sim \sum_m \pi_m N(\boldsymbol{\mu}_{x_m 0} + \mathbf{T}_m \boldsymbol{\omega}^{(i)}, \boldsymbol{\Sigma}_{x_m}), \\ \boldsymbol{\omega}^{(i)} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{14}$$

where $\mathcal{N}(\boldsymbol{\mu}_{x_m 0}, \boldsymbol{\Sigma}_{x_m})$ and π_m are UBM Gaussian components and their weights, $\mathbf{T}_m$ matrices describe a low-rank subspace (called total variability subspace) in which GMM means can be adapted to a particular speech segment and $\boldsymbol{\omega}^{(i)}$ is a segment-specific standard normal distributed latent vector. For a speech segment, the i-vector is extracted as the maximum a posteriori (MAP) point estimate of the latent vector $\boldsymbol{\omega}^{(i)}$.

The model for i-vector extraction can be now adapted to noise by substituting the clean model (14) into equations (3) and (7). We perform the VTS expansion at $(\boldsymbol{\mu}_{x_m 0}, \boldsymbol{\mu}_{n0}, \boldsymbol{\mu}_{h0})$ that corresponds to the clean UBM and noise means estimated using the EM algorithm from the previous section (i.e. $\boldsymbol{\mu}_{n0}$ and $\boldsymbol{\mu}_{h0}$ are set to values obtained form updates (11) and (12), respectively). This results in the following noise-adapted model:

$$\mathbf{y}^{(i)} \sim \sum_m \pi_m N(\boldsymbol{\mu}_{y_m 0}^{(i)} + \mathbf{G}_m^{(i)} \mathbf{T}_m \boldsymbol{\omega}^{(i)}, \boldsymbol{\Sigma}_{y_m}^{(i)}), \tag{15}$$

where $\mathbf{G}_m^{(i)}$, $\boldsymbol{\mu}_{y_m 0}^{(i)}$ and $\boldsymbol{\Sigma}_{y_m}^{(i)}$ are given by equations (4), (6) and (7). This noise-adapted model can be used for i-vector extraction where the resulting i-vectors should be (to a large extent) independent of additive and convolutive noise. They can therefore better represent the remaining variability present in speech segments, which is likely to be informative for speaker recognition.

For the convenience, let us define the following statistics collected from a noisy speech segment using the noise-adapted UBM:

$$\begin{aligned} \mathbf{f}_{y_m}^{(i)} &= \sum_t \gamma_{mt}^{(i)} (\mathbf{G}_m^{(i)})^{-1} (\mathbf{y}_t^{(i)} - \boldsymbol{\mu}_{y_m}^{(i)}) \\ (\hat{\boldsymbol{\Sigma}}_{y_m}^{(i)})^{-1} &= (\mathbf{G}_m^{(i)})^T (\boldsymbol{\Sigma}_{y_m}^{(i)})^{-1} \mathbf{G}_m^{(i)}. \end{aligned} \tag{16}$$

For a fixed soft frame alignment $\mathbf{y}_m^{(i)}$, it can be shown that the posterior distribution of $\boldsymbol{\omega}^{(i)}$ from equation (15) is Gaussian with mean and covariance matrix:

$$\begin{aligned} \langle \boldsymbol{\omega}^{(i)} \rangle &= \mathbf{L}^{(i)} \sum_m \mathbf{T}_m^T (\hat{\boldsymbol{\Sigma}}_{y_m}^{(i)})^{-1} \mathbf{f}_{y_m}^{(i)} \\ \mathbf{L}^{(i)} &= (I + \sum_m \gamma_m^{(i)} \mathbf{T}_m^T (\hat{\boldsymbol{\Sigma}}_{y_m}^{(i)})^{-1} \mathbf{T}_m)^{-1}. \end{aligned} \tag{17}$$

The i-vector extracted for segment s is given by taking a MAP estimate for this distribution $\langle \boldsymbol{\omega}^{(i)} \rangle$.

Finally, we derive the corresponding EM algorithm to train the subspace parameters $\mathbf{T}_m$ in the i-vector extraction model (15). In the E-step, the posterior distribution of the latent vector $\boldsymbol{\omega}^{(i)}$ is estimated for each training segment using eq (17). The matrices $\mathbf{T}_m$ can be updated in the M-step using:

$$\begin{aligned} \mathrm{vec}(\mathbf{T}_m) &= \left(\sum_i \left(\gamma_{y_m}^{(i)} \langle \boldsymbol{\omega}^{(i)} (\boldsymbol{\omega}^{(i)})^T \rangle \right) \otimes (\hat{\boldsymbol{\Sigma}}_{y_m}^{(i)})^{-1} \right)^{-1} \\ &\quad \times \mathrm{vec} \sum_i (\hat{\boldsymbol{\Sigma}}_{y_m}^{(i)})^{-1} \mathbf{f}_{y_m}^{(i)} \langle \boldsymbol{\omega}^{(i)} \rangle^T \\ \langle \boldsymbol{\omega}^{(i)} \boldsymbol{\omega}^{(i)T} \rangle &= \mathbf{L}^{(i)} + \langle \boldsymbol{\omega}^{(i)} \rangle \langle \boldsymbol{\omega}^{(i)} \rangle^T, \end{aligned} \tag{18}$$

where $\otimes$ is the Kronecker product and vec is an operator which creates a column vector from a matrix by stacking its columns. The i-vector model for the dynamic features is very similar to the one for the static feature replacing the calculation of $\boldsymbol{\mu}_{y_m}^{(i)}$ in (15) with $\boldsymbol{\mu}_{\Delta y_m}^{(i)} = \mathbf{G}_m^{(i)} \boldsymbol{\mu}_{\Delta 0 m}$.

4. EXPERIMENTAL SETUP

Our speaker recognition system frontend extracts 20 MFCC coefficients (including C0), augmented with first and second order derivatives. A 512 diagonal component UBM is trained in a gender-dependent fashion on NIST telephone data from the speaker recognition evaluation (SRE) 2004 and 2005. A i-vector extractor of dimension 400 is then trained on a larger set (NIST SRE '04, '05, '06, Switchboard, and Fisher). The dimensionality of i-vectors is further reduced to 200 by LDA, followed by length normalization and PLDA.

Results are shown on a part of the PRISM set described in [5, 12], where different noisy speech samples are added to the training, enrollment, and test sets without any overlap at three different signal-to-noise ratios (SNR) (20dB, 15dB, and 8dB). System performance is reported in terms of detection cost function (DCF) on three SNRs. The detection cost function (DCF) effective prior is the one from NIST SRE 2010 [13].

The baseline system employs the above configuration and uses mean and variance normalization (MVN) on the MFCC features estimated using the speech portion of the audio file. We compare this baseline system and a system where MVN was replaced by our VTS compensation. In the case of a VTS compensated system, we first train the i-vector extractor as follows:

1. A UBM model is trained on clean data, with no artificially added noise.
2. The UBM is adapted to each speech segment using 4 iterations of EM described in section 2.2, where the covariance matrices are updated in the second iteration and the means are updated in the others.
3. This noise-adapted UBM is used to extract sufficient statistics (16) from each speech segment.
4. Using 5 EM iterations from section 3:
   (a) Estimate the posterior distribution of the latent variable using (17) for each segment.
   (b) Update matrices $\mathbf{T}_m$ using (18).

After this training process, i-vectors are extracted for each enrollment and test segments using steps 2, 3 and 4a).

5. RESULTS

Table 1 presents the DCF performance of the baseline (MVN) and VTS system at different SNR. Two PLDA backends were evaluated:

a *clean* backend, where the model was trained exclusively on clean data; a *multistyle* backend, where the model was trained on clean and noisy data as proposed in [5]. Results clearly show a very large gains obtained using our VTS based approach over the state-of-the-art system, especially on low SNR conditions.

Although multistyle training brought a large improvement for the MVN system, the VTS system using a clean backend still outperforms the latter in the noisy conditions. A multistyle VTS system brings an additional gain which show the complementarity of both approaches. Similar behavior was observed at the equal error rate (EER). Figure 1 shows the DET curves of all four systems at a SNR of 8dB for a more detailed performance comparison.

	clean		multistyle	
Eval. condition	MVN	VTS	MVN	VTS
SNR=8dB	0.975	0.639	0.810	0.480
SNR=15dB	0.661	0.269	0.437	0.234
SNR=20dB	0.350	0.179	0.260	0.170
Clean	0.082	0.146	0.086	0.145

Table 1. *DCF performance of a state-of-the-art baseline system compared to our VTS approach where both clean and multistyle backends were used. The VTS system significantly outperforms the baseline system in low SNR conditions.*

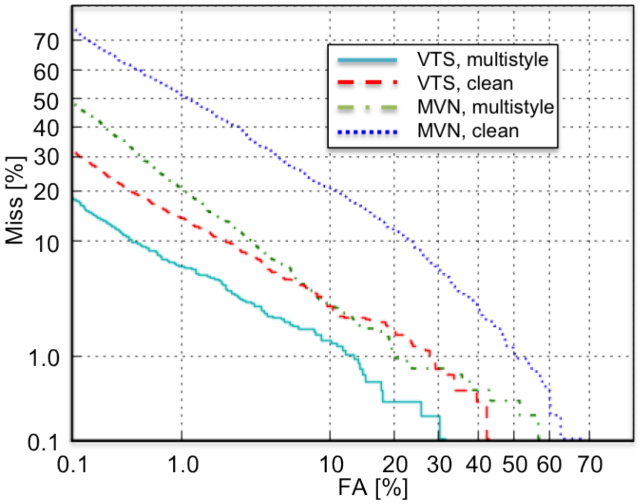

Fig. 1. Comparison of four systems at SNR=8dB. *MVN* means using MVN on MFCC features; *VTS* means using VTS for compensation; *clean* means using a backend model trained on clean data only; *multistyle* means using a backend model trained on clean and noisy data.

6. CONCLUSIONS

In this study, we successfully adapted the VTS approach to speaker recognition by proposing a new i-vector extraction framework. We show how improvements observed for VTS in speech recognition can be also obtained for speaker recognition. The proposed approach, while computationally more expensive than the standard i-vector framework, presents a relative improvement in low SNR conditions (e.g. 15 and 8db). For example, as can be also seen in figure 1, for a miss probability around 10%, the relative improvements in flase alarm rate are on the order of 70% to 80% compared to a state-of-the-art system.

We also show that our approach is robust to new and unseen data as a VTS-based system trained on clean data only outperforms a baseline system trained in a multistyle fashion in noisy conditions. This makes this approach very attractive for realistic operational scenarios where the type of degradation may not be known in advance.

We have identified two directions for future work. First, the computational requirements of the method are very high and it is impractical to scale our UBM beyond 512 Gaussians or the ivector dimension beyond 400. A substantial effort need to be put into optimizations and simplifications of the framework. Second, in speech recognition, VTS is used during the UBM model training as to 'clean up' the model for degradations caused by noise. We will explore a similar strategy for speaker recognition.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, pp. 788–798, May 2010.

[2] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV-11th*. IEEE, 2007, pp. 1–8.

[3] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010-The Speaker and Language Recognition Workshop*. IEEE, 2010.

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

[5] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *ICASSP-2012*. IEEE, March 2012, pp. 4253–4256.

[6] P. J. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[7] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *ICSLP*, 2000, vol. 3, pp. 229–232.

[8] O. Kalinli, M. Seltzer, J. Droppo, and A Acero, "Noise adaptive training for robust automatic speech recognition," *IEEE Trans. ASLP*, vol. 18, pp. 1889–1901, Nov. 2010.

[9] H Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, PhD dissertation, University of Cambridge, Sept. 2007.

[10] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, July 2008.

[11] R. H. Byrd, P. Lu, and J. Noceda, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific and Statistical Computing*, vol. 16, pp. 1190–1208, Nov. 1995.

[12] L. Ferrer, H. Bratt, L. Burget, J. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of NIST 2011 Workshop*, 2011.

[13] "NIST SRE10 evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.