ADDITIVE NOISE COMPENSATION IN THE I-VECTOR SPACE FOR SPEAKER RECOGNITION

Waad Ben Kheder, Driss Matrouf, Jean-François Bonastre, Moez Ajili and Pierre-Michel Bousquet

LIA, University of Avignon, France

ABSTRACT

State-of-the-art speaker recognition systems performance degrades considerably in noisy environments even though they achieve very good results in clean conditions. In order to deal with this strong limitation, we aim in this work to remove the noisy part of an i-vector directly in the i-vector space. Our approach offers the advantage to operate only at the i-vector extraction level, letting the other steps of the system unchanged. A maximum a posteriori (MAP) procedure is applied in order to obtain clean version of the noisy i-vectors taking advantage of prior knowledge about clean i-vectors distribution. To perform this MAP estimation, Gaussian assumptions over clean and noise i-vectors distributions are made. Operating on NIST 2008 data, we show a relative improvement up to 60% compared with baseline system. Our approach also outperforms the "multi-style" backend training technique. The efficiency of the proposed method is obtained at the price of relative high computational cost. We present at the end some ideas to improve this aspect.

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

1. INTRODUCTION

Additive noise has always been one of the most important problems in speaker recognition research and dealing with it generally falls into one of four categories: speech enhancement, feature compensation, robust modeling or score compensation. We will not discuss here the latter as it is not dealing directly with additive noise.

At a signal level, [1] proved that spectral and wavelet-based speech enhancement techniques do not perform consistently when used as a pre-processing block to a standard speaker recognition system even if the resultant speech quality increases. It was further shown in [2] that these algorithms might either enhance or degrade the recognition performance depending on the noise type and the SNR level.

At a feature level, [3] carried out an extensive comparison of several spectrum estimation methods and found that the best estimator was related to the noise type and SNR level. Recent work [4, 5], based on vector Taylor series (VTS) then developed using "unscented transforms" [6] 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 technique, adding a normalization step or changing the used parameters could mean to rewrite the whole technique.

On a model level, the parallel model combination (PMC) was first introduced in speech recognition technology [7] before to be adapted to speaker recognition [8] by building a noisy model and using it to decode noisy test segments. To apply 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. In practice, the high computational expense, mainly in the scoring model, of such a procedure makes it non realizable. A robust backend training called "multi-style" [9] was proposed as a possible solution to account for the noise in the scoring phase. This method uses a large set of clean and noisy data (affected with different noises and SNR levels) to build a generic scoring model. The obtained model gives good performance in general but still is 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.

In this paper, we propose an i-vectors "denoising" procedure to deal with additive noise. The advantage of this 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 are made over the clean i-vectors and the noise distributions in the i-vector space. We assume that both clean i-vectors and noise are normally distributed in the i-vector space. The first assumption is justified by the factor analysis model used to extract the i-vectors [10] which supposes a normal distribution for the resulting i-vectors. Regarding the noise, a Gaussian distribution modeling seems to be suitable. Even though, theoretically, the noise is 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". In addition, the approach

is extensible to a mixture of Gaussians to model the noise i-vectors. The originality of this technique is that it uses not only information about the noise but also information about clean i-vectors (the corresponding probability density functions in the i-vector space). Hence, the risk of introducing new distortions in the obtained i-vector is minimized.

2. I-VECTORS DENOISING

This section describes our new i-vectors "cleaning" technique. De-noising the i-vector directly allows to use classical state-of-the-art scoring models based on generative models like two-cov [11], G-PLDA [12] or HT-PLDA [13] estimated using clean data without any adaptation to the test noise.

Formally, given a noisy i-vector Y_0, our goal is to estimate the corresponding clean version $\hat{X_0}$. Let's define two random variables X and Y corresponding respectively to the clean and noisy i-vectors. Let the noise random variable N be:

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

We consider that clean i-vectors X are normally distributed as described in [10], 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 :

$$f(X) = \mathcal{N}(\mu_X, \Sigma_X) \quad (2) \qquad f(N) = \mathcal{N}(\mu_N, \Sigma_N) \quad (3)$$

where $\mathcal{N}(\mu_i, \Sigma_i)$ denotes a normal distribution with mean μ_i and full covariance matrix Σ_i. Referring to (1),(2) and (3) we can express $f(Y_0|X)$ for a given Y_0 as:

$$f(Y_0|X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_N|^{\frac{1}{2}}} exp^{-\frac{1}{2}(Y_0 - X - \mu_N)^t \Sigma_N^{-1}(Y_0 - X - \mu_N)} \tag{4}$$

Based on the noise model (1) and the two previously defined distributions, we can estimate for a given noisy i-vector Y_0 its clean version $\hat{X_0}$ using a MAP estimator :

$$\hat{X_0} = \underset{X}{\operatorname{argmax}}\{\ln f(X/Y_0)\} \tag{5}$$

Using the Bayesian rule, we can write :

$$\hat{X_0} = \underset{X}{\operatorname{argmax}}\{\ln f(Y_0/X) f(X)\} \tag{6}$$

Finding $\hat{X_0}$ becomes equivalent to solve:

$$\frac{\partial}{\partial X}\{\ln f(Y_0/X) + \ln f(X)\} = 0 \tag{7}$$

By developing (7) using (2) and (4), we end up with:

$$\begin{aligned} &\frac{\partial}{\partial X}\{(Y_0 - X - \mu_N)^t \Sigma_N^{-1}(Y_0 - X - \mu_N)\} \\ &+ \frac{\partial}{\partial X}\{(X - \mu_X)^t \Sigma_X^{-1}(X - \mu_X)\} = 0 \end{aligned} \tag{8}$$

After the derivation, the final expression of the clean i-vector $\hat{X_0}$ given the noisy version Y_0 and both X and N distributions parameters is:

$$\hat{X_0} = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1}(\Sigma_N^{-1}(Y_0 - \mu_N) + \Sigma_X^{-1}\mu_X) \tag{9}$$

In i-vector-based speaker recognition systems [10], length-normalization was proved to improve the overall performance [14]. In our case, it is important to mention that all used noisy and clean i-vectors were initially length-normalized.

3. EXPERIMENTAL PROTOCOL

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 toolkit is used for the estimation of the total variability matrix and the i-vectors extraction. The used algorithms are described in [15]. Finally a two-covariance-based scoring [11] is applied. The equal-error rate (EER) over the NIST SRE 2008 male test data on the "short2/short3" task under the "det7" conditions [16] (all trials involving only English language telephone speech spoken by a native U.S. English speaker in training and test). It will be used as a reference to monitor the performance improvements compared to the baseline system and to the "multi-style" backend in noisy conditions.

We use 6 noise samples from the free sound repository FreeSound.org [17] as background noises (crowd noise, air-cooling noise, rain, cars traffic noise, nature noise and engine noise). The open-source toolkit FaNT [18] was used to add these noises to the full waveforms generating new noisy audio files for each noise / SNR level.

4. 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 de-noising 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 initially 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 and then $f(N)$.

We focus in the following on minimizing the number of train files used to build $f(N)$ along with their selection criteria. We will work with two different noises (crowd and air-cooling) on three SNR levels (10dB, 5dB and 0dB) using 3000 clean train speech segments (SNR $>$ 25dB).

4.1. Number of i-vectors needed to estimate $f(N)$

In a "clean enrollment / noisy test" setup and for each one of the previously described six configurations, the EER is evaluated using a different number of train i-vectors to estimate $f(N)$ going from 400 to 3000. (9) is used every time prior to the scoring phase using the selected i-vectors to estimate μ_N and Σ_N. Figure 1 shows the obtained results:

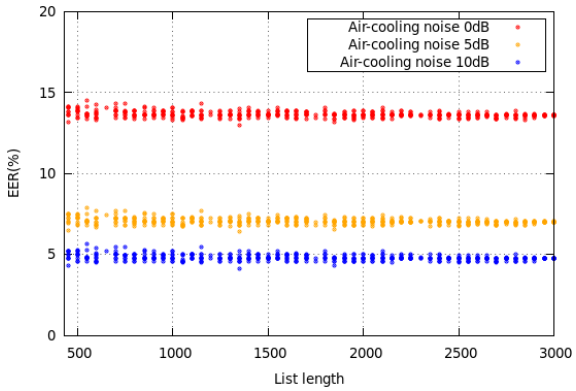

Fig. 1. EER variation with the amount of i-vectors used to estimate the noise distribution $f(N)$ for the "air-cooling noise" at 0dB, 5dB and 10dB (10 measures for each length).

It is clear that for the three SNR levels, the EER does not vary much beyond 500. Then, we will fix to 500 i-vectors the noise model training set size for our next experiments.

4.2. Train i-vectors selection

Once fixed to 500 the number of i-vectors needed to estimate $f(N)$, we concentrate on their selection criteria. For the six different configurations, we created a set of 300 lists of 500 elements picked randomly from the original set of 3000 clean audio files which will be used to estimate $f(N)$. For each list, we plot the resultant EER after compensation with respect to the average files speech duration. Figure 2 shows the curve obtained using noisy test data affected with crowd-noise on 10dB.

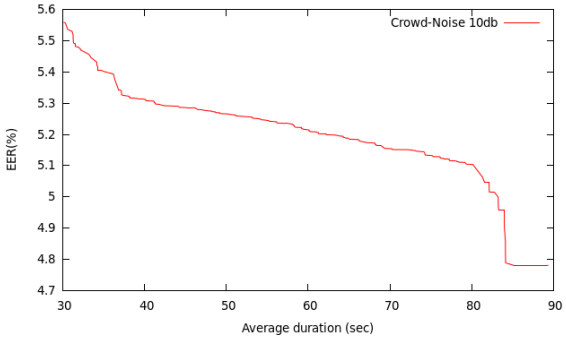

Fig. 2. EER variation with the average speech duration of the segments used to estimate $f(N)$ for the crowd-noise on 10dB.

It is easy to see that the longer speech segments give better results than the short ones. We observed the same shape for the other five configurations. In the following, the longest 500 files (having a speech duration of 90 seconds) will be used as a train set to estimate $f(N)$.

5. INTEGRATION OF THE DENOISING METHOD IN A SPEAKER RECOGNITION SYSTEM

The new i-vector denoising method allows to build a speaker recognition system that takes into account the test signal SNR level as shown in Figure 3. As mentioned in the previous section, the clean i-vectors distribution $f(X)$ and the clean train i-vectors (X) are extracted once and for all prior to the denoising procedure. Before starting, an SNR threshold above which a segment is considered clean has to be specified. Then, the algorithm follows these steps :

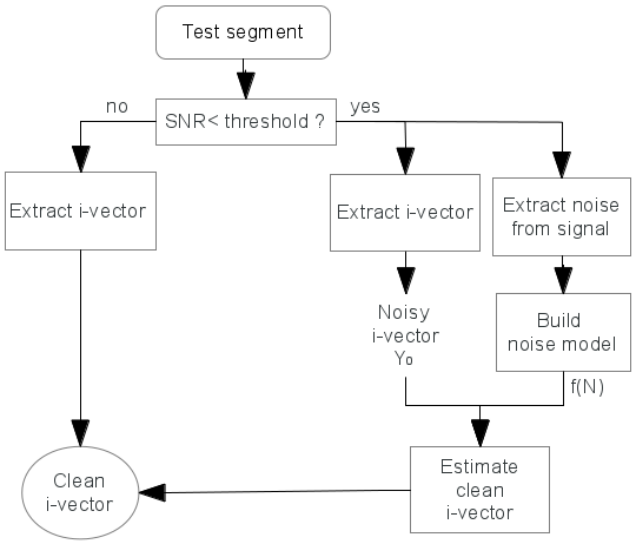

Fig. 3. Clean i-vector extraction algorithm.

- **SNR checking:** The SNR level is estimated for the test segment and compared to the threshold.
- **The clean case:** If the segment is clean, then a standard i-vector extraction is done.
- **The noisy case:** If the segment is noisy :
  1. The corresponding noisy i-vector Y_0 is computed.
  2. A VAD is used to extract the noise part from the signal (by selecting the low-energy frames in the signal corresponding to the non-speech intervals).
  3. The noise is added to a set of clean train files with the SNR of the test file (estimated in the first step).
  4. A standard i-vector extraction is done using the noisy train files (corresponding to the Y data).
  5. The noise distribution $f(N)$ in the i-vector space is estimated using (1).
  6. The new clean i-vector is estimated using (9).

6. RECOGNITION PERFORMANCE

In this section, the new estimated clean i-vectors (corresponding to either test or enrollment segments) will be referred to as "I-MAP" vectors. The LIA speaker verification baseline system reaches an EER=1.59% in clean conditions. We will be comparing three performances in this section :

- Noisy i-vectors used with the baseline system (clean backend).
- Noisy i-vectors used with a multi-style backend (the scoring model is built using clean and noisy data affected with different noises and SNR levels.
- I-MAP vectors used with a clean backend (the algorithm described in Section 5 is used for each i-vector).

The enrollment and test data have been altered using two different sets of noises (crowd noise, rain and engine noise) for enrollment and (air-cooling, cars traffic and nature noise) for test at five different SNR levels: 0dB, 5dB, 10dB, 15dB and 20dB. Each utterance has been affected by one noise at a fixed SNR level. Figure 4 shows the performances of the three systems in matched/mismatched SNR scenarios. An average relative improvement of 43% is observed in all conditions compared to the baseline performance and of 28% compared to a "multi-style" backend performance. The results validate the method efficiency in matched and mismatched SNR conditions while using different noises.

To prove the validity of our technique in a situation where the noise level is varying randomly, we perform a last experiment. In this experiment, all the speech files (for enrollment and test) are corrupted by a noise with a varying SNR level between 0dB to 20dB, the SNR level is selected randomly. Table 1 shows the obtained results with the three systems.

The same range of improvement is also observed in this condition. This result validates the use of the proposed algorithm in unknown test/enrollment conditions.

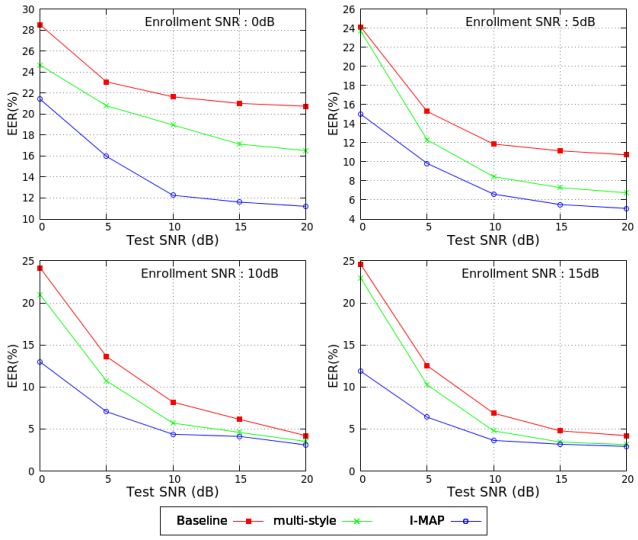

Fig. 4. Each figure corresponds to a different enrollment SNR. The x-axis corresponds to the SNR level in the test segments and the y-axis gives the resultant EER.

Table 1. Performance comparison in an heterogeneous setup.

	EER (%)
Baseline	27.65
"multi-style" backend	23.12
I-MAP + clean backend	16.27

7. CONCLUSION

In this work, we introduced an i-vector cleaning technique working only inside the i-vector domain. Our approach assumes that both the clean i-vectors and the noise distributions (in the i-vectors space) are normally distributed. It allows to estimate the noise for a given test and to eliminate its influence inside the corresponding i-vector.

Significant improvement was observed using our approach compared to a baseline system or a "multi-style" backend system (60% to 43% of relative improvement). An experiment using a randomly mixed setup, in terms of noise level inside both train and test files, showed that our approach still allows a large improvement compared to the two other systems (16.27% of EER to be compared with 27.65% of EER for the baseline system). These results demonstrate clearly the potential of our approach.

Further improvements could be achieved to deal with the computational cost of the proposed algorithm. One solution is to build a noise distribution database in the i-vector space estimated using a large number of noise categories and different SNR levels, then select the appropriate distribution depending on the test i-vector. In terms of performance, it seems also interesting to us to use a mixture of Gaussians in place of a simple Gaussian to model the noise inside the i-vector space.

8. REFERENCES

[1] 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.

[2] 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.

[3] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4769–4772.

[4] Yun Lei, Lukas Burget, and Nicolas Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6788–6791.

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

[6] 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.

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

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

[9] Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4253–4256.

[10] 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.

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

[12] 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.

[13] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, p. 14.

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

[15] 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.

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

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

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