Robust speaker recognition using MAP estimation of additive noise in i-vectors space

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

LIA, University of Avignon, France
{waad.ben-kheder,driss.matrouf,pierre-michel.bousquet,
jean-francois.bonastre,moez.ajili}@univ-avignon.fr

Abstract. In the last few years, the use of i-vectors along with a generative back-end has become the new standard in speaker recognition. An i-vector is a compact representation of a speaker utterance extracted from a low dimensional total variability subspace. Although current speaker recognition systems achieve very good results in clean training and test conditions, the performance degrades considerably in noisy environments. The compensation of the noise effect is actually a research subject of major importance. As far as we know, there was no serious attempt to treat the noise problem directly in the i-vectors space without relying on data distributions computed on a prior domain. This paper proposes a full-covariance Gaussian modeling of the clean i-vectors and noise distributions in the i-vectors space then introduces a technique to estimate a clean i-vector given the noisy version and the noise density function using MAP approach. Based on NIST data, we show that it is possible to improve up to 60% the baseline system performances. A noise adding tool is used to help simulate a real-world noisy environment at different signal-to-noise ratio levels.

Keywords: i-vectors, MAP adaptation, speaker recognition, additive noise

1 Introduction

Recent work on the robustness of i-vector -based speaker recognition systems has been carried out at different levels in order to track and compensate the additive noise effect without altering the speaker-related information. After the success of VTS (Vector Taylor Series) in robust ASR applications [1], a VTS-based i-vectors extractor was proposed in [2, 3] and then developed in [4] using "Unscented transforms" trying to model non-linear distortions in the mel-cepstral domain based on a non-linear noise model in order to compensate both convolutive and additive noises. This compensation scheme tackles the problem on an early stage by computing the "clean i-vector" directly by fitting the corresponding noisy GMM to a given noisy speech segment. That requires information about spectral data distribution to do the link between the two domains. The biggest weakness

of this technique is the complexity of the estimation model and the number of imposed constraints which makes it extremely rigid and hardly extendable. The integration of many interesting techniques (like feature warping [5] for robust channel mismatch) becomes a hard task and requires to rebuild the whole model.

This motivates the development of a new kind of noise models which operates directly in the i-vectors space. We show in this paper that it's possible to reach far better results than VTS-based techniques based on an additive noise model in the i-vectors space using only noise and clean i-vectors distributions. We start by assuming that both clean i-vectors and noise can be modeled by full-covariance Gaussian distributions. Then, we present an i-vectors "cleaning" technique that uses the MAP approach to estimate a clean i-vector given a noisy i-vector version and a normal noise distribution model.

This paper is structured as follows. Section 2 describes the i-vector framework for speaker recognition. Section 3 details the proposed approach. Section 4 presents the experimental protocol, the experiments and the corresponding results.

2 The i-vectors Framework

In this section we present the i-vectors framework along with the scoring procedure that will be used further in our experiments.

2.1 The total-variability Subspace

In this approach, an i-vector extractor converts a sequence of acoustic vectors into a single low-dimensional vector representing the whole speech utterance. The speaker- and session-dependent super-vector s of concatenated Gaussian Mixture Model (GMM) means is assumed to obey a linear model of the form :

$$s = m + Tw \tag{1}$$

where :

- m is the mean super-vector of the Universal Background Model (UBM)
- T is the low-rank variability matrix obtained from a large dataset by MAP estimation [6]. It represents the total variability subspace.
- w is a standard-normally distributed latent variable called "i-vector".

Extracting an i-vector from the total variability subspace is essentially a maximum a-posteriori adaptation of w in the space defined by T . The algorithms for the estimation of T and the extraction of i-vectors are described in [7].

2.2 The i-vectors scoring System

Many dimensionality reduction techniques (such as LDA) and generative models (like PLDA, and the Two-covariance model) have been developed in order to improve the i-vectors comparison in speaker verification trials. The speaker verification score given two i-vectors w_1 and w_2 is the likelihood ratio described by:

$$score = log\frac{P(w_1, w_2|\theta_{tar})}{P(w_1, w_2|\theta_{non})} \tag{2}$$

where the hypothesis θ_{tar} states that inputs w_1 and w_2 are from the same speaker and the hypothesis θ_{non} states they are from different speakers.

We focus in the following on the generative model that we used in our work: the two-covariance scoring model.

The two-covariance scoring Model:

This model is a particular case of the Probabilistic Linear Discriminant Analysis (PLDA) described in [8]. 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 w of a speaker s can be decomposed in:

$$w = y_s + \varepsilon \tag{3}$$

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

$$P(y_s) = \mathcal{N}(\mu, B) \tag{4}$$

$$P(w|y_s) = \mathcal{N}(y_s, W) \tag{5}$$

$\mathcal{N}$ 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 :

$$B = \sum_{s=1}^{S} \frac{n_s}{n}(y_s - \mu)(y_s - \mu)^t \tag{6}$$

$$W = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(w_i^s - y_s)(w_i^s - y_s)^t \tag{7}$$

where n_s is the number of utterances for speaker s, n is the total number of utterances, 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 (6) and (7), the score from equation (2) can be expressed as:

$$s = \frac{\int \mathcal{N}(w_1|y, W)\mathcal{N}(w_2|y, W)\mathcal{N}(y|\mu, B)dy}{\prod_{i=1,2}\int \mathcal{N}(w_i|y, W)\mathcal{N}(y|\mu, B)dy} \tag{8}$$

the explicit solution of (8) is given in [9].

3 MAP estimation of clean I-vectors

Given a noisy i-vector Y_0, the goal of this section will be to estimate the corresponding clean version $\hat{X}_0$. We will work exclusively in the i-vectors space and build a clean i-vectors estimator based solely on "i-vector space"-related data using a MAP approach.

Let's start by defining two random variables in the i-vectors space :

- X which corresponds the clean i-vectors.
- Y which corresponds the noisy i-vectors.

To model the additive noise in the i-vectors space, we define a third random variable N that links X and Y according to the following expression:

$$N = Y - X \tag{9}$$

We assume that both clean i-vectors (X) and noise data (N) can be represented by two normal distributions in the i-vectors space. We can then define the corresponding probability distribution functions $f(X)$ and $f(N)$ as :

$$f(X) = \mathcal{N}(\mu_X, \Sigma_X) \tag{10}$$

$$f(N) = \mathcal{N}(\mu_N, \Sigma_N) \tag{11}$$

where $\mathcal{N}(\mu_i, \Sigma_i)$ denotes a normal distribution with mean μ_i and full covariance matrix Σ_i.

Referring to hypothesis (9),(10) and (11) we can express $f(Y_0|X)$ for a given Y_0 as:

$$f(Y_0|X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\{(Y_0 - X - \mu_N)^t \Sigma_N^{-1}(Y_0 - X - \mu_N)\} \tag{12}$$

Based on the noise model (9) 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 :

$$\hat{X}_0 = \underset{X}{\text{argmax}}\{\ln f(X/Y_0)\} \tag{13}$$

Using the Bayesian rule, we can write $f(X/Y_0)$ as :

$$f(X/Y_0) = \frac{f(Y_0/X)f(X)}{f(Y_0)} \tag{14}$$

After combining (13) and (14):

$$\hat{X}_0 = \underset{X}{\text{argmax}}\{\ln f(Y_0/X)f(X)\} \tag{15}$$

Finding $\hat{X}_0$ becomes equivalent to solving:

$$\frac{\partial}{\partial X}\{\ln f(Y_0/X) + \ln f(X)\} = 0 \tag{16}$$

By developing (16) using (10) and (12), we end up with:

$$\frac{\partial}{\partial X}\{(Y_0 - X - \mu_N)^t \Sigma_N^{-1}(Y_0 - X - \mu_N) + (X - \mu_X)^t \Sigma_X^{-1}(X - \mu_X)\} = 0 \tag{17}$$

After the derivation, we have :

$$-\Sigma_N^{-1}(Y_0 - \hat{X}_0 - \mu_N) + \Sigma_X^{-1}(\hat{X}_0 - \mu_X) = 0 \tag{18}$$

then, we find the final expression of the clean i-vector $\hat{X}_0$ given the noisy version Y_0 and both X and N distributions parameters:

$$\hat{X}_0 = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1}(\Sigma_N^{-1}(Y_0 - \mu_N) + \Sigma_X^{-1}\mu_X) \tag{19}$$

The estimation of $f(X)$ and $f(N)$ are done as so:

- $f(X)$: μ_X and Σ_X are estimated once and for all over a large set of clean i-vectors. Since this distribution is independent from the noise, there is no constraints on the number of i-vectors to be used.
- $f(N)$: In real-world conditions, the available amount of noisy data is generally limited. Possible improvements of this technique could be proposed in future publications to deal with this constraint. Based on a set of clean and noisy i-vectors pairs corresponding to the same clean utterances, the noise data set in the i-vectors space is firstly computed using $N = Y - X$. Then μ_N and Σ_N are estimated as any regular normal distribution parameters.

In i-vector -based speaker recognition systems, length-normalization was proved to improve the overall system performance [10]. In our case, it's important to mention that all used noisy and clean i-vectors in the estimation process of $\hat{X}_0$, $f(X)$ and $f(N)$ were initially length-normalized.

4 Experimental protocol and Results

In this section, we present the configuration used in the LIA speaker recognition system along with the training and test data sets. Then, the noise adding procedure and the realized experiments are detailed.

4.1 The LIA speaker recognition baseline System

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 256 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 implemented algorithms are described in [7]. Finally a two-covariance-based scoring scheme is applied.

4.2 Noise Adding

We will use two different noises in our analysis:

- A crowd-noise
- An air-cooling noise

The open-source toolkit FaNT [11] (Filtering and Noise Adding Tool) was used to add these noises at different SNR levels generating new noisy audio files.

In order to have a good estimation of the clean normal i-vectors distribution, we have selected the 6000 utterances from the training data having an SNR greater than 30dB.

For each test condition, we used 3000 pairs of clean and noisy i-vectors to estimate the normal noise distribution model. N is firstly computed with $N = Y - X$ then $f(N)$ is estimated by computing μ_N and Σ_N.

At the end, six trial conditions will be evaluated for each noise:

- Noisy test/target data with "Crowd-noise" at SNR levels 10db, 5db and 0db.
- Noisy test/target data with "Air-cooling noise" at SNR levels 10db, 5db and 0db.

4.3 Test data and performance Evaluation

The equal-error rate (EER) over the NIST SRE 2008 test data will be used as a reference to monitor the performance improvement compared to the baseline system in noisy conditions. We will be only focused on the "short2/short3" task under the "det7" conditions [12]. In order to help visualize the improvement in the error-rate in each test configuration, the relative improvement measure (RI%) will be added.

The two studied noises have been used to create noisy versions of the test and target data over 10db, 5db and 0db SNR levels.

4.4 Experiments and Results

The LIA speaker verification baseline system reaches EER=1.59% in clean conditions. This error-rate will be the lower bound that helps evaluate the gain of the proposed technique compared to the noisy baseline performance.

In the following tables, the estimated clean i-vectors corresponding to noisy test or target i-vectors will be referred to as "I-MAP" vectors.

The system performances will be presented in two different configurations:

- Clean target i-vectors and noisy test i-vectors.
- Noisy target i-vectors and noisy test i-vectors.

First, we evaluate the baseline system performances before and after the application of our method when all noisy data (test and target noisy i-vectors) are produced by the same noise.

Clean target i-vectors and noisy test I-vectors (Crowd-Noise):

The table 1 summarizes the baseline system performance while used with noisy test i-vectors (Crowd-noise) and clean target i-vectors compared to the proposed method performance:

Table 1. System performance using noisy test data (Crowd-noise)

	EER (%)		
	Baseline system	**with I-MAP test**	**RI (%)**
SNR=10db	5.86	3.18	**45.73**
SNR=5db	9.53	4.34	54.46
SNR=0db	17.08	8.43	50.64

We observe more than 50% relative improvement in average at the three SNR levels. This encourages the use of clean target models when available with noisy test data.

Clean target i-vectors and noisy test I-vectors (Air-cooling Noise):

The table 2 summarizes the baseline system performance while used with noisy test and target i-vectors (Air-cooling noise) and clean target i-vectors compared to the proposed method performance:

Table 2. System performance using noisy test data (Air-cooling noise)

	EER (%)		
	Baseline system	**with I-MAP test**	**RI (%)**
SNR=10db	7.47	4.78	**36.01**
SNR=5db	15.68	7.3	53.44
SNR=0db	27.33	13.89	49.18

We observe more than 46% relative improvement in average at the three SNR levels. The overall performance is comparable to the previous one and validates the proposed method for different noisy test conditions.

Noisy target i-vectors and noisy test I-vectors (Crowd-Noise):

In real speaker recognition applications, clean target data could not be available, so it's important to check the validity of the proposed method in noisy target i-vectors conditions.

The table 3 summarizes the baseline system performance while used with noisy test and target i-vectors (Crowd-noise) compared to the proposed method performance:

Table 3. System performance using noisy test and target data (Crowd-noise)

	EER (%)		
	Baseline system	**with I-MAP target and I-MAP test**	**RI (%)**
SNR=10db	10.72	4.34	59.51
SNR=5db	17.79	8.15	54.19
SNR=0db	24.77	13.44	**45.74**

We observe more than 53% relative improvement in average at the three SNR levels. It's important to note that our method keeps its efficiency even with noisy target i-vectors.

Noisy target i-vectors and noisy test I-vectors (Air-cooling Noise):

The table 4 summarizes the baseline system performance while used with noisy test and target i-vectors (Air-cooling noise) compared to the proposed method performance:

Table 4. System performance using noisy test and target data (Air-cooling noise)

	EER (%)		
	Baseline system	**with I-MAP target and I-MAP test**	**RI (%)**
SNR=10db	16.14	6.83	57.68
SNR=5db	20.73	10.5	49.35
SNR=0db	32.89	20.5	**37.67**

We observe more than 48% relative improvement in average at the three SNR levels. The relative improvement with this noise is also comparable with the "clean target - noisy test" performance. This validates the robustness of the proposed method in different noisy target and test conditions.

It's easy to see the considerable leap between the baseline system performance and the one obtained after the MAP estimation of the clean i-vectors in all previous conditions. For each of the two noises, the average relative improvement exceeds 48% in 10dB and 5dB SNR levels conditions. One of the most interesting results is the efficiency of this method even on very low SNR levels (0dB).

Noisy data in real-world applications could be affected by different noise sources. Based on this idea, it's interesting to evaluate the performance of this technique in test conditions where more than one noise is present. To test this possibility, we mixed evenly for each SNR level the noisy i-vectors coming from both noises. This way, for every SNR level, 50% of the noisy test i-vectors are related to the "crowd-noise" and the other 50% is related to the "air-cooling" noise. The same mixing scheme is done on noisy target i-vectors in the "noisy test - noisy target" configuration.

The following tables summarizes the baseline system performance before and after the application of our technique.

Clean target i-vectors and noisy test I-vectors (two noises):

The table 5 summarizes the system performance before and after the application of our technique for clean target data and mixed noisy test i-vectors (coming from two different noises):

Table 5. System performances for clean target and mixed noisy test i-vectors

	EER (%)		
	Baseline system	**with I-MAP target and I-MAP test**	**RI (%)**
SNR=10db	7.06	3.92	**44.47**
SNR=5db	13.24	5.92	55.28
SNR=0db	22.55	11.86	47.40

We observe more than 50% relative improvement in average at the three SNR levels. The overall performance is maintained compared to the first configurations when we used only one noise. These results validate the efficiency of the proposed method.

Noisy target i-vectors and noisy test I-vectors (two noises in both):

The table 6 summarizes the system performance before and after the application of our technique for mixed noisy target and test i-vectors (coming from two different noises):

Table 6. System performances for mixed noisy test and target i-vectors

	EER (%)		**RI (%)**
	Baseline system	**with I-MAP target and I-MAP test**	
SNR=10db	15.49	6.16	60.23
SNR=5db	24.15	9.79	59.46
SNR=0db	34.16	22.78	**33.31**

Similar performance is also observed in this condition (51% average relative improvement) showing the validity of the used method in the "noisy target - noisy test" configuration.

5 Conclusion

In this work, we introduced a new clean i-vector estimation technique referring to a noisy version based on a normal distribution model of both clean i-vectors and noise in the i-vectors space using a MAP approach. The observed improvement compared to the baseline system performance reaches 60% in low SNR test conditions and outperforms recently developed robust speaker recognition techniques (like VTS-based i-vector extractors).

Further improvements could be achieved by extending the noise distribution model in the i-vectors space (using Gaussian mixtures instead of unimodal Gaussian distributions for example). The use of a factor analysis -based technique like PLDA could also be explored to improve the quality of the i-vectors used to build the noise distribution model.

References

1. Alex Acero, Li Deng, Trausti T Kristjansson, and Jerry Zhang. Hmm adaptation using vector taylor series for noisy speech recognition. In *INTERSPEECH*, pages 869–872, 2000.

2. 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*, pages 6788–6791. IEEE, 2013.
3. Yun Lei, Mitchell McLaren, Luciana Ferrer, and Nicolas Scheffer. Simplified vts-based i-vector extraction in noise-robust speaker recognition. *submitted to) ICASSP, Florence, Italy*, 2014.
4. David Martınez, Lukáš Burget, Themos Stafylakis, Yun Lei, Patrick Kenny, and Eduardo Lleida. Unscented transform for ivector-based noisy speaker recognition.
5. Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. 2001.
6. Patrick Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.
7. 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*, pages 1242–1245, 2007.
8. 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*, pages 1–8. IEEE, 2007.
9. Niko Brümmer and Edward De Villiers. The speaker partitioning problem. In *Odyssey*, page 34, 2010.
10. Daniel Garcia-Romero and Carol Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011.
11. FaNT - Filtering and Noise Adding Tool. `http://dnt.kr.hsnr.de/download.html`. [Online; accessed 15-May-2014].
12. The NIST year 2008 speaker recognition evaluation plan. `http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08\_evalplan\_release4.pdf`, 2008. [Online; accessed 15-May-2014].