Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Iterative Bayesian and MMSE-based noise compensation techniques for speaker recognition in the i-vector space

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

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

Abstract

Dealing with additive noise in the i-vector space can be challenging due to the complexity of its effect in that space. Several compensation techniques have been proposed in the last years to either remove the noise effect by setting a noise model in the i-vector space or build better scoring techniques that take environment perturbations into account. We recently presented a new efficient Bayesian cleaning technique operating in the i-vector domain named I-MAP that improves the baseline system performance by up to 60%. This technique is based on Gaussian models for the clean and noise i-vectors distributions. After I-MAP transformation, these hypothesis are probably less correct. For this reason, we propose to apply another MMSE-based approach that uses the Kabsch algorithm. For a certain noise, it estimates the best translation vector and rotation matrix between a set of train noisy i-vectors and their clean counterparts based on RMSD criterion. This transformation is then applied on noisy test i-vectors in order to remove the noise effect. We show that applying the Kabsch algorithm allows to reach a 40% relative improvement in EER(%) compared to a baseline system performance and that, when combined with I-MAP and repeated iteratively, it allows to reach 85% of relative improvement.

keywords: i-vector, additive noise, Kabsch algorithm, I-MAP

1. Introduction

State of the art speaker recognition systems achieve high recognition rates in clean environments but can suffer considerably in presence of environment noise. Due to the number of factors affecting the recognition decision and to the variety of noise sources, a universal fast and efficient noise compensation technique is not yet available.

We recently presented a new efficient Bayesian cleaning technique operating in the i-vector domain named I-MAP [1, 2, 3]. It is a "data-driven" i-vector cleaning method based on an additive noise model in the i-vector space. It estimates a clean i-vector given its noisy version and the noise distribution using MAP approach. It uses a full-covariance Gaussian modeling of the clean i-vectors and noise distributions in the i-vector space. Even though the noise is known to be non-additive in this space, using such model with a MAP estimator makes the derivations very simple while giving up to 60% of relative improvement compared to the baseline system performance. Since applying I-MAP does not guarantee that the Gaussianity hypothesis is true for residual noise (thus we can't use I-MAP iteratively on noisy test data), we propose to complement this technique by applying another MMSE-based approach that uses the Kabsch algorithm. By doing so, we achieve two goals: On one hand, we want to improve the recognition performance of I-MAP by combining it with another algorithm that uses a different optimization criterion (even though a Bayesian technique performs in general better than an MMSE-based algorithm, combining the two can outperform each of the two). On the other hand, we want to be able to use these techniques (I-MAP+Kabsch) iteratively to achieve even better recognition performance and remove the noise effect more efficiently while using the same train data.

In this paper, we present a MMSE-based approach for noise compensation in the i-vector space that complements I-MAP using the Kabsch algorithm. Originally developed in cheminformatics to compare molecular structures [4, 5] and used in bioinformatics to compare protein structures [6], the Kabsch algorithm estimates the best translation vector and rotation matrix between two paired sets of points based on root mean squared deviation (RMSD) criterion. It is possible to use this algorithm in a speaker recognition application by supposing that the effect of additive noise can be modeled in the i-vectors space by a translation followed by a rotation. In this context, the algorithm uses two paired sets of clean and noisy i-vectors affected by a certain noise to estimate the best translation vector and rotation matrix that transform the noisy i-vectors to their clean counterparts. Once the translation vector and the rotation matrix corresponding to a certain noise are estimated, the two transformations are applied on noisy test i-vectors in order to remove the noise effect.

We show that using this algorithm allows to reach a 40% relative improvement in EER(%) compared to a baseline system performance. Since deriving a noise model in the i-vector space based on its additive effect in the temporal domain is a complex task due to the number of transformations included in the i-vector extraction process [7], we will combine this algorithm with I-MAP in a second experiment and show that using the two techniques can achieve up to 80% of relative EER(%) improvement. Then, we will show that these two algorithms (I-MAP+Kabsch) complement each other and that applying them iteratively can yield i-vectors of better quality achieving up to 85% or relative EER(%) improvement.

This paper is structured as follows: Section 2 compares different noise compensation techniques used in speaker recognition (SR) systems. Section 3 presents a new noise compensation technique operating in the i-vector space based on the Kabsch algorithm. Section 4 presents the I-MAP denoising procedure. Finally, Section 5 presents the experimental protocol, the conducted experiments and analyses the findings of this paper.

doi: 10.21437/Odyssey.2016-9

2. Robust speaker recognition in noisy conditions

In order to build robust speaker recognition (SR) systems, different approaches, operating in different domains, have been proposed:

In the temporal domain, spectral and wavelet-based speech enhancement techniques were proven to be noise and SNR level-dependent and can either degrade or improve the recognition performance depending on the environment noise [8, 9]. Different speech enhancement algorithms based on non-negative matrix factorization (NMF) have also been developed for speech-based applications [10, 11] and have been proven to better model non-stationary noise. Despite its consistency (the algorithm does not degrade the recognition performance compared to algorithms described in [8, 9]), the relative improvement in recognition performance reached by NMF in a speaker recognition context [12] is relatively low compared to other methods (10% of relative improvement in EER(%)).

On a feature level, new robust representations have been proposed lately for robust speaker recognition such as Residual Phase Cepstrum Coefficients (RPCC) [13], Generalized perceptual features (GFCC) [14], Cosine Distance Features (CDF) [15], combined formants, wavelets, and neural network features (FWENN) [16], modulation filtering of autoregressive models [17] and Convolutive Sparse Coding of speech spectrograms [18]. In such representations, a relative improvement of EER(%) that varies between 5% and 27% is observed.

Several stochastic compensation techniques in the cepstral domain (such as Trajectory-based stochastic mapping Mapping (TRAJMAP) [19] and Stereo Stochastic Mapping (SSM) [20]) were recently tested in [21] and achieved high recognition rates in noisy environments. But such algorithms assume prior knowledge about the test noise (stereophonic data are used to train two UBM models in parallel: the first is trained using clean frames while the second represents noisy frames. The relationship between the two UBMs (for each Gaussian) is then used to perform the denoising of test frames).

With the rise of deep learning in the last few years, neural networks were successfully used in a large range of tasks including speech recognition [22, 23, 24, 25], facial recognition [26, 27] and object recognition [28, 29] before being applied to speaker recognition. Different architectures were used to either estimate more robust i-vector statistics or compensate noisy features. In [30], a convolutional neural network (CNN) was used to compute posterior probabilities for speech frames replacing the UBM model and has been shown to produce more robust i-vector statistics. In [31], a deep neural network (DNN) was trained to classify speakers based on speech frames. Then during speaker enrollment, the trained DNN is used to extract speaker specific features from the last hidden layer. For each utterance, the average of activations derived from the last hidden layer is used as speaker model. In [32], a DNN is used to enhance cepstral features before extracting i-vectors. In this system, the DNN is trained from parallel data of clean and noise-corrupted speech which are aligned in the frame level. Deep learning techniques achieve on average a relative improvement of EER up to 30% in noisy conditions.

On a model level, a set of algorithms based on vector Taylor series (VTS) were proposed in [33, 34] then developed using ”unscented transforms” [35]. Such algorithms use a non-linear noise model in the cepstral domain and model the relationship between clean and noisy cepstral coefficients. In the recognition phase, the developed noise model is integrated in the i-vector extractor to help estimate a ”cleaned-up” version of noisy i-vectors. Despite their efficiency, such models are rigid and can be hard to adapt (adding a normalization step or changing the used parameters could mean to rewrite the whole technique).

In the scoring phase, a robust backend training called ”multi-style” was proposed in [36] as a possible solution to take into the account for the effect of noise. 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 (up to 30% of relative improvement) but still suboptimal to take into the account for a particular noise because of its generalization (the same system is used for all noises). Alternatively, another class of techniques based on uncertainty propagation have also been proposed lately for robust speaker recognition. Based on this idea, a robust i-vector extractor was proposed in [37, 38] in order to make the i-vector extraction system focus on reliable or reliably enhanced features but showed little improvement compared to other methods. Recently, an SNR-invariant version of PLDA was proposed in [39]. In this framework, i-vectors extracted from utterances falling within a narrow SNR range are assumed to share similar SNR-specific information and used to develop a more robust version of PLDA which decomposes an i-vector in three components: speaker, SNR, and channel. This model showed an average relative improvement of 25% in EER(%) compared to regular PLDA.

In the next Section, we present a new denoising technique operating in the i-vector space based on the Kabsch algorithm.

3. The Kabsch algorithm

Given two paired sets of points $\{x_i\}_{i=1..N}$ and $\{y_i\}_{i=1..N}$ defined in an M-dimensional space where to each point x_i in the first set corresponds a unique point y_i in the second (hence the term ”paired” sets), it is possible to arrange the corresponding coordinates in a matrix format as:

$$P_X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{pmatrix} \quad (1)$$

$$P_Y = \begin{pmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,M} \\ y_{2,1} & y_{2,2} & \dots & y_{2,M} \\ \vdots & \vdots & & \vdots \\ y_{N,1} & y_{N,2} & \dots & y_{N,M} \end{pmatrix} \quad (2)$$

Given P_X and P_Y, the orthogonal Procrustes problem is a matrix approximation problem which aims at finding the best orthogonal matrix R that maps P_X to P_Y according to:

$$R = \underset{R}{\operatorname{argmin}} \|RP_Y - P_X\|_F \quad (3)$$

where: $R^T R = I_N$ and $\|.\|_F$ denotes the Frobenius norm.

It is possible to constrain this problem by only allowing rotation matrices (i.e. orthogonal matrices with determinant equal to 1). In that context, the solution can be found using the Kabsch algorithm [4]. This algorithm allows to estimate the best rotation matrix R which transforms the set $\{x_i\}_{i=1..N}$ (P_X) onto $\{y_i\}_{i=1..N}$ (P_Y) based on root mean squared deviation (RMSD) criterion.

In a speaker recognition context, we will use the Kabsch algorithm to find, for a certain noise, the best rotation matrix R

between a set of noisy i-vectors and their clean versions. By doing so, it will be possible to apply the resultant transformation to noisy test i-vector and recover a ”cleaned-up” version. The algorithm starts with two sets of paired i-vectors represented in a matrix format (clean i-vectors matrix P_X corresponding to $\{x_i\}_{i=1..N}$ and noisy i-vectors matrix P_Y corresponding to $\{y_i\}_{i=1..N}$ where i-vectors are arranged by rows). The estimated rotation matrix R will then characterize the noise present in $\{y_i\}_{i=1..N}$.

The Kabsch algorithm allows to find the best rotation matrix which transforms P_Y onto P_X and follows three steps:

1. Both sets of points (P_X and P_Y) are translated so that their centroid coincide with the origin of the coordinate system.
2. The rotation matrix is estimated using the two centered matrices (using SVD decomposition).
3. During the test phase, the rotation matrix is applied to noisy test i-vectors.

Step 1: Translation of the two sets of points:

1. Computing the centroids of the clean and noisy sets of i-vectors:
   - $\overline{P_X} = centroid(P_X)$
   - $\overline{P_Y} = centroid(P_Y)$
2. Centering all points of P_X and P_Y around the origin of the coordinate system:
   - $\tilde{P_X}_i = {P_X}_i - \overline{P_X}$ for each row ${P_X}_i$ of P_X.
   - $\tilde{P_Y}_i = {P_Y}_i - \overline{P_Y}$ for each row ${P_Y}_i$ of P_Y.

Step 2: Estimation of the rotation matrix:

1. Estimation of a covariance matrix: $A = \tilde{P_X}^T \tilde{P_Y}$
2. SVD decomposition of A: $A = VSW^T$
3. Computing $d = sign(det(WV^T))$
4. Estimation of the rotation matrix R as:

$$R = W \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & 1 & \vdots \\ 0 & \dots & 0 & d \end{pmatrix} V^T \qquad (4)$$

Step 3: Application of the rotation on test data:

Given a set of noisy test i-vectors $\{t_i\}_{i=1..N}$:

1. Centering test i-vectors:
   $\tilde{t_i} = t_i - \overline{P_Y}$ for all i in $i = 1..N$.
2. Rotating test i-vectors:
   $\hat{t_i} = R\tilde{t_i} + \overline{P_X}$ for all i in $i = 1..N$.

Then, the resultant i-vectors $\hat{t_i}$ can be used with a clean backend since they are supposed to be noise-free. It is important to note that since the centroid $\overline{P_X}$ is noise-independent, it can be computed once in a off-line step over a large set of clean i-vectors and used in all transformations involving different noises / SNR levels. Also, in order to have a good estimate of the covariance matrix A, we will be working in a setup where $M > N$.

4. The I-MAP denoising procedure

In our previous work [1, 2, 3], we proposed an additive noise model in the i-vector space obeying to the equation:

$$N = Y - X \qquad (5)$$

Where X and Y are two random variables representing respectively clean and noisy i-vectors and N represents the noise. Using full-covariance Gaussian distributions for both clean i-vectors $d_X \sim \mathcal{N}(X; \mu_X, \Sigma_X)$ and noise in the i-vector space $d_N \sim \mathcal{N}(N; \mu_N, \Sigma_N)$, it is possible to write the cleaned-up version $\hat{X_0}$ of a noisy i-vector Y_0 using MAP criterion as [1, 2, 3]:

$$\hat{X_0} = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1}(\Sigma_N^{-1}(Y_0 - \mu_N) + \Sigma_X^{-1}\mu_X) \qquad (6)$$

The derivation of Equation 6 is detailed in [1, 2, 3].

4.1. Estimation of $\mathcal{N}(X; \mu_X, \Sigma_X)$ and $\mathcal{N}(N; \mu_N, \Sigma_N)$

As detailed in [1, 2], the clean i-vectors distribution $\mathcal{N}(X; \mu_X, \Sigma_X)$ and the noise distribution $\mathcal{N}(N; \mu_N, \Sigma_N)$ are two key components in the I-MAP procedure.

Since $\mathcal{N}(X; \mu_X, \Sigma_X)$ is noise-independent, it can 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, $\mathcal{N}(N; \mu_N, \Sigma_N)$ 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 set is 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 system and selecting the low-energy frames) then added to the clean train set in the time domain. Finally, the corresponding noisy i-vectors (Y) are estimated and Equation (6) is used to compute N then $\mathcal{N}(N; \mu_N, \Sigma_N)$. The full algorithm is shown in Figure 1 and more details about the technique can be found in [2]. We also showed that the best performances are reached when using clean train sessions ($SNR > 25dB$) with an average speech duration of 90 seconds. We will use a similar configuration in our experiments. In order to speed-up the algorithm in unknown test conditions, it is possible to fix an SNR threshold beyond which a session is considered clean and I-MAP is not applied.

5. Experiments and results

5.1. Experimental protocol:

Throughout this paper, all conducted 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 (selected using an energy-based VAD). The low-energy frames (corresponding mainly to silence) are removed.

Two SR systems are used in our experiments depending of the speakers gender in enrollment/test data. Two gender-dependent 512 diagonal component UBMs and total variability matrices of low rank 400 are estimated using NIST SRE 2004, 2005, 2006 and Switchboard data. The male models

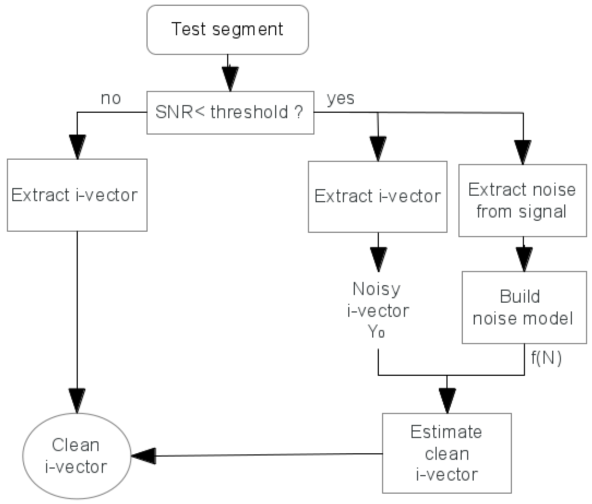

Figure 1: I-vector cleaning procedure using I-MAP.

(male UBM and total variability matrix) were trained using 15660 utterances corresponding to 1147 speakers and the female models (female UBM and total variability matrix) were trained using 24100 utterances corresponding to 2012 speakers. The LIA_SpkDet package of the LIA_RAL/ALIZE toolkit [40] is used for the estimation of the total variability matrix and the i-vector extraction. The algorithms used are described in [41]. Finally a two-covariance-based scoring [42] is applied. The equal-error rate (EER) over the NIST SRE 2008 male test data on the "short2/short3" task under the "det7" conditions (all trials involving only English language telephone speech in training and test) [43] will be used as a reference to monitor the performance improvement compared to the baseline system. In clean "det7" conditions, our system reaches $EER = 1.59\%$ on male data and $EER = 2.66\%$ on female data.

In order to test the response to the developed techniques in adverse environments, we used 3 noise samples (air-cooling noise, crowd noise and car-driving noise) from the free sound repository FreeSound.org [44] as background noises. The open-source toolkit FaNT [45] was used to add these noises to in the temporal domain generating new noisy audio files for each noise / SNR level. All clean train data used in the our experiments have an average speech duration of 90 seconds and an SNR level greater than 25dB. Also, the SNR threashold used

Table 1: Recognition performance on male data in different test conditions using clean enrollment and noisy test data. The number of iterations indicates how many times I-MAP and Kabsch are applied successively.

		EER(%)				
Test condition		**Baseline**	**Kabsch**	**I-MAP**	**I-MAP + Kabsch (1 iteration)**	**I-MAP + Kabsch (2 iterations)**
Air-cooling noise	**0dB**	26.85	17.18	13.21	8.86	**7.24**
	5dB	15.21	10.34	7.25	4.71	**3.89**
	10dB	9.51	5.70	4.85	2.94	**2.55**
	15dB	5.41	3.40	2.85	1.82	**1.63**
Car-driving noise	**0dB**	25.54	15.83	12.05	7.91	**6.37**
	5dB	14.54	9.30	6.65	3.63	**3.04**
	10dB	8.32	5.15	3.78	1.99	**1.82**
	15dB	4.82	3.22	2.36	1.79	**1.65**
Crowd noise	**0dB**	24.24	16.48	11.55	8.24	**6.78**
	5dB	13.94	9.20	5.09	4.18	**3.62**
	10dB	7.77	5.20	3.05	2.02	**1.81**
	15dB	4.01	2.52	2.02	1.84	**1.63**

Table 2: Recognition performance on female data in different test conditions using clean enrollment and noisy test data. The number of iterations indicates how many times I-MAP and Kabsch are applied successively.

		EER(%)				
Test condition		**Baseline**	**Kabsch**	**I-MAP**	**I-MAP + Kabsch (1 iteration)**	**I-MAP + Kabsch (2 iterations)**
Air-cooling noise	**0dB**	27.19	16.95	13.53	10.80	**9.49**
	5dB	16.77	10.45	8.34	6.66	**5.85**
	10dB	9.01	5.61	4.48	3.58	**3.14**
	15dB	6.42	4.00	3.19	2.75	**2.70**
Car-driving noise	**0dB**	24.82	15.47	12.35	9.86	**8.66**
	5dB	14.90	9.28	7.41	5.92	**5.20**
	10dB	8.65	5.39	4.30	3.43	**3.02**
	15dB	5.89	3.67	3.12	2.95	**2.74**
Crowd noise	**0dB**	25.44	15.85	12.66	10.11	**8.88**
	5dB	14.37	8.95	7.15	5.71	**5.01**
	10dB	8.77	5.46	4.36	3.48	**3.06**
	15dB	5.78	3.60	3.04	2.92	**2.81**

Table 3: Recognition performance on male data in different test conditions using noisy enrollment and test data. The number of iterations indicates how many times I-MAP and Kabsch are applied successively.

		EER(%)				
Test & enrolment condition		**Baseline**	**Kabsch**	**I-MAP**	**I-MAP + Kabsch (1 iteration)**	**I-MAP + Kabsch (2 iterations)**
Air-cooling noise	**0dB**	35.06	22.08	15.42	8.36	**6.76**
	5dB	28.47	17.65	13.09	7.25	**6.13**
	10dB	23.68	15.86	13.02	5.92	**4.73**
	15dB	18.14	11.60	9.43	3.99	**3.44**
Car-driving noise	**0dB**	30.97	20.44	16.41	8.98	**6.57**
	5dB	26.09	16.17	11.21	6.52	**5.73**
	10dB	22.46	14.37	9.20	5.81	**4.64**
	15dB	17.85	11.60	7.49	4.99	**4.08**
Crowd noise	**0dB**	34.55	21.42	14.51	7.94	**6.56**
	5dB	27.31	16.93	11.74	7.21	**6.06**
	10dB	22.99	14.02	9.19	6.50	**5.74**
	15dB	17.26	10.35	7.59	4.48	**3.45**

Table 4: Recognition performance on female data in different test conditions using noisy enrollment and test data. The number of iterations indicates how many times I-MAP and Kabsch are applied successively.

		EER(%)				
Test & enrolment condition		**Baseline**	**Kabsch**	**I-MAP**	**I-MAP + Kabsch (1 iteration)**	**I-MAP + Kabsch (2 iterations)**
Air-cooling noise	**0dB**	36.10	22.50	14.00	9.65	**8.27**
	5dB	28.89	18.00	11.20	7.72	**6.62**
	10dB	24.12	15.02	9.35	6.45	**5.53**
	15dB	19.24	11.98	7.46	5.14	**4.41**
Car-driving noise	**0dB**	31.54	19.65	12.23	8.43	**7.23**
	5dB	27.65	17.21	10.72	7.39	**6.34**
	10dB	23.68	14.77	9.18	6.33	**5.42**
	15dB	18.65	11.61	7.23	4.98	**4.27**
Crowd noise	**0dB**	35.02	21.82	13.58	9.36	**8.03**
	5dB	28.01	17.45	10.86	7.49	**6.42**
	10dB	23.54	14.68	9.13	6.29	**5.39**
	15dB	18.65	11.61	7.23	4.98	**4.27**

for I-MAP is equal to 25dB.

For I-MAP, the number of train i-vectors N needed to estimate the noise distribution for each noise $\mathcal{N}(N;\mu_N,\Sigma_N)$ was investigated in [2]. We will use $N = 500$ in all our experiments and the same set of train data will be used in the Kabsch algorithm to compute the rotation matrix R and translation vector $\overline{P_Y}$ corresponding to each noise.

5.2. Recognition performance using the Kabsch algorithm

The LIA speaker verification baseline system reaches an EER=1.59% in clean conditions. We will compare five systems performances in these experiments (a clean backend is used for all systems):

- **Baseline system:** Noisy i-vectors used with the baseline system.
- **Kabsch algorithm:** for each noise n / SNR level s:
  1. A set of clean train audio segments are affected with the noise n at sdB in the temporal domain.
  2. The i-vectors corresponding to the resultant noisy segments $\{y_i\}_{i=1..N}$ and their clean counterparts $\{x_i\}_{i=1..N}$ are extracted.
  3. Steps 1 and 2 of the Kabsch algorithm are applied to $\{x_i\}_{i=1..N}$ and $\{y_i\}_{i=1..N}$ and both the translation vector $\overline{P_Y}$ and the rotation matrix R are estimated.
  4. Step 3 of the Kabsch algorithm is applied on noisy test i-vectors using R and $\overline{P_Y}$.
- **I-MAP + Kabsch algorithm (1 iteration):** for each noise n / SNR level s, the I-MAP transformation is applied to both noisy test and noisy train i-vectors, then the Kabsch algorithm is applied:
  1. A set of clean train audio segments are affected with the noise n at sdB in the temporal domain.
  2. The i-vectors corresponding to the resultant noisy segments $\{y_i\}_{i=1..N}$ and their clean counterparts $\{x_i\}_{i=1..N}$ are extracted.
  3. Equation 5 is applied then the noise distribution $f(N)$ is estimated.
  4. I-MAP is applied (Equation 6) to noisy test i-vectors $\{t_i\}_{i=1..N}$ (generating $\{t_i^{'}\}_{i=1..N}$).
  5. I-MAP is applied to the set of noisy train i-vectors $\{y_i\}_{i=1..N}$ (generating $\{y_i^{'}\}_{i=1..N}$).

6. Steps 1 and 2 of the Kabsch algorithm are applied to $\{x_i\}_{i=1..N}$ and $\{y_i^{'}\}_{i=1..N}$ and both the translation vector $\overline{P_{Y'}}$ and the rotation matrix R are estimated.

7. Step 3 of the Kabsch algorithm is applied to the noisy test i-vectors transformed with I-MAP ($\{t_i^{'}\}_{i=1..N}$) using R and $\overline{P_{Y'}}$.

- **I-MAP + Kabsch algorithm (2 iterations):** The procedure described in the pervious system is applied twice on noisy test and train data.

The enrollment and test data have been altered using three noises (air-cooling noise, car driving and crowd-noise) at 4 different SNR levels: 0dB, 5dB, 10dB and 15dB.

First, we present the system performance using clean enrollment data, then we compare them with the results obtained in different noisy enrollment configurations.

5.2.1. System performance using clean enrollment and noisy test data

For three different test noises (air-cooling noise, car-driving noise and crowd noise), clean test data are corrupted in the time domain and the corresponding i-vectors are evaluated before and after the application of Kabsch, I-MAP and I-MAP+Kabsch. Tables 1 and 2 show respectively the five systems performance on male and female data for different test noises.

When the Kabsch algorithm is used, a relative improvement range between 33% and 40% is observed, whereas the use of I-MAP followed by the Kabsch algorithm gives a range of 65% up to 85% of relative improvement compared to the baseline system.

It is important to highlight the power behind the combination of these two techniques. Indeed, when the two algorithms are compared separately, I-MAP performs better than Kabsch due to its Bayesian nature. But using both algorithms (either for one or many iterations) can be highly efficient since the two algorithms use different optimization criteria (MAP for I-MAP and RMSD for Kabsch), hence iteratively improving the quality of the cleaned-up i-vectors. Also, the application of I-MAP produces residual noise that does not necessarily obey the Gaussianity hypothesis (thus we can't use I-MAP iteratively on noisy test data). This problem can be corrected by the Kabsch algorithm and can explain the good fit of the two techniques when used more than once. Indeed, Table 1 shows that using the Kabsch algorithm after I-MAP (either for 1 or 2 iterations) is a good combination and that it improves the recognition performance achieving 85% of relative improvement compared to the baseline system.

5.2.2. System performance using noisy enrollment and test data

For three different test noises (air-cooling noise, car-driving noise and crowd noise), clean test data are corrupted in the time domain and the corresponding i-vectors are evaluated before and after the application of I-MAP. Tables 3 and 4 show respectively the five systems performance for male and female data when each one of the three noises are used to affect the test data.

When the Kabsch algorithm is used, a relative improvement range between 33% and 40% is observed, whereas the use of I-MAP followed by the Kabsch algorithm gives a range of 65% up to 83% of relative improvement compared to the baseline system. This proves that combining the two techniques is till efficient even when noisy enrollment i-vectors are used. Also, using two iterations of I-MAP+Kabsch can fearther improve the recognition performance.

5.2.3. System performance in a heterogeneous setup

We performed another experiment to prove the validity of our technique in a situation where the noise level is varying randomly between the enrollment/test segments. In this experiment, all the speech files (for enrollment and test) are corrupted by a random noise with a randomly-selected SNR level between 0dB to 20dB. As a result, each noisy session is affected by a unique noise at a fixed SNR level. Table 5 shows the obtained results with the five systems.

Table 5: Performance comparison in a heterogeneous setup for male and female data.

	EER (%)	
	Male	**Female**
Baseline	29.65	31.02
Kabsch	18.78	19.95
I-MAP	16.27	17.46
I-MAP + Kabsch (1 iter.)	8.67	10.62
I-MAP + Kabsch (2 iter.)	**7.39**	**9.28**

It is easy to see that while the Kabsch algorithm and I-MAP improve the recognition performance respectively by 36% and 45%, the combination of the two techniques allows to achieve 75% in an heterogeneous setup.

Even though these techniques (I-MAP and the Kabsch algorithm) achieve high recognition rates in noisy environments, it is worth noting that both algorithms are based on a set of paired i-vectors (clean i-vectors and their noisy versions) and that generating these data for each noisy test session can be computationally expensive in a real SR system. A possible solution is to build a noisy i-vector distribution database offline (using many noises and SNR levels). Then, for each noisy test i-vector, select the closest distribution from database (using a likelihood measure) and use it as train data for both I-MAP and Kabsch algorithms. This system could be the subject of a future work.

6. Conclusion

In this paper, we introduced a new cleaning technique operating in the i-vector space based on the Kabsch algorithm. We showed that using this technique leads to a relative gain in EER(%) that reaches 40% and that combining it with I-MAP allows to reach 85% of relative improvement.

For a given noise, we estimate the best translation vector and rotation matrix between a set of train noisy i-vectors and their clean counterparts based on RMSD criterion and show that applying this transformation to noisy test i-vectors achieves 40% of relative improvement. Then, we combined this algorithm with I-MAP, a recently proposed i-vector denoising technique and showed that using the two algorithms iteratively allows to reach 85% of relative improvement in EER.

7. References

[1] Waad Ben Kheder, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre, and Moez Ajili, “Ro-

bust speaker recognition using map estimation of additive noise in i-vectors space," in *Statistical Language and Speech Processing*, pp. 97–107. Springer, 2014.

[2] Waad Ben Kheder, Driss Matrouf, Jean-François Bonastre, Moez Ajili, and Pierre-Michel Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4190–4194.

[3] D Matrouf, W Ben Kheder, PM Bousquet, M Ajili, and JF Bonastre, "Dealing with additive noise in speaker recognition systems based on i-vector approach," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2092–2096.

[4] Wolfgang Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[5] Wolfgang Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 34, no. 5, pp. 827–828, 1978.

[6] Ying-Hung Lin, Hsun-Chang Chang, and Yaw-Ling Lin, "A study on tools and algorithms for 3-d protein structures alignment and comparison," in *International Computer Symposium*, 2004, pp. 3–70.

[7] 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.

[8] 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.

[9] 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.

[10] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran, "Speech denoising using nonnegative matrix factorization with priors.," in *ICASSP*, 2008, pp. 4029–4032.

[11] Kevin W Wilson, Bhiksha Raj, and Paris Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising.," in *Interspeech*, 2008, pp. 411–414.

[12] SH Liu, YX Zou, and HK Ning, "Nonnegative matrix factorization based noise robust speaker verification," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015, pp. 35–39.

[13] Jianglin Wang and Michael T Johnson, "Residual phase cepstrum coefficients with application to cross-lingual speaker verification," in *Thirteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2012.

[14] Patrick J Clemins, Marek B Trawicki, Kuntoro Adi, Jidong Tao, and Michael T Johnson, "Generalized perceptual features for vocalization analysis across multiple species," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[15] Kuruvachan K George, C Santhosh Kumar, KI Ramachandran, and Ashish Panda, "Cosine distance features for robust speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] Khaled Daqrouq and Tarek A Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing*, vol. 27, pp. 231–239, 2015.

[17] Shrikanth Ganapathy, Sri Harish Mallidi, and Hynek Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 8, pp. 1285–1295, 2014.

[18] Antti Hurmalainen, Rahim Saeidi, and Tuomas Virtanen, "Noise robust speaker recognition with convolutive sparse coding," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

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

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

[22] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[23] Li Deng, Geoffrey Hinton, and Brian Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[24] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Mike Seltzer, Geoffrey Zweig, Xiaodong He, Julia Williams, et al., "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.

[25] Abdel-rahman Mohamed, Tara N Sainath, George Dahl, Bhuvana Ramabhadran, Geoffrey E Hinton, and Michael A Picheny, "Deep belief networks using discriminative features for phone recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5060–5063.

[26] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[27] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Hybrid deep learning for face verification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1489–1496.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[29] Vinod Nair and Geoffrey E Hinton, "3d object recognition with deep belief nets," in *Advances in Neural Information Processing Systems*, 2009, pp. 1339–1347.

[30] Mitchell McLaren, Yun Lei, Nicolas Scheffer, and Luciana Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions.," in *INTERSPEECH*, 2014, pp. 686–690.

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

[32] Steven Du, Xiong Xiao, and Eng Siong Chng, "Dnn feature compensation for noise robust speaker verification," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015, pp. 871–875.

[33] 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.

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

[35] 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.

[36] 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.

[37] Chengzhu Yu, Gang Liu, Seongjun Hahm, and John HL Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4017–4021.

[38] Dayana Ribas, Emmanuel Vincent, and José Ramon Calvo, "Uncertainty propagation for noise robust speaker recognition: the case of nist-sre," in *Interspeech 2015*, 2015.

[39] Na Li and Man-Wai Mak, "Snr-invariant plda modeling for robust speaker verification," *Proc. Interspeech'15*, 2015.

[40] Anthony Larcher, Jean-Francois Bonastre, Benoit GB Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John SD Mason, and Jean-Yves Parfait, "Alize 3.0-open source toolkit for state-of-the-art speaker recognition.," in *Interspeech*, 2013, pp. 2768–2772.

[41] 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.

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

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

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

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