Intersession compensation and scoring methods in the i-vectors space for speaker recognition

Pierre-Michel Bousquet, Driss Matrouf, Jean-François Bonastre

University of Avignon (LIA), France

{pierre-michel.bousquet, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

Abstract

The total variability factor space in speaker verification system architecture based on Factor Analysis (FA) has greatly improved speaker recognition performances. Carrying out channel compensation in a low dimensional total factor space, rather than in the GMM supervector space, allows for the application of new techniques. We propose here new intersession compensation and scoring methods. Furthermore, this new approach contributes to a better understanding of the session variability characteristics in the total factor space.

Index Terms: Joint Factor Analysis, i-vectors, Total variability space, speaker recognition.

1. Introduction

The use of Gaussian Mixture Models (GMM) in a GMM-UBM framework has been a standard in speaker verification [1]. This generative model has been extended to model jointly the speaker component and the session (or channel) component. This extension is named Joint Factor Analysis (JFA) [2].

Inspired by the JFA approach, N. Dehak [6] proposed to extract from the GMM super-vector a compact version, named i-vector, containing both the speaker and the session information. In this case, the session variability is taken into account during the scoring process, unlike JFA which takes it into account during the speaker modelling given an utterance. In the i-vector space, the session variability is modeled by using a global covariance matrix modeling the correlation between components in vectors containing only the session information. These session-only-dependent vectors are obtained by subtracting, from a given i-vector, the mean of all sessions belonging to the speaker corresponding to that i-vector.

This idea is very promising as it opens a wide panel of perspectives in connection with the data analysis domain. N. Dehak has proposed to use LDA and WCCN to process the i-vectors [5]. In this paper we propose a new i-vector treatment method, adapted to the spatial distribution of these vectors in the total factor space. We will validate this new method on NIST Speaker Recognition Evaluation 2006 and 2008.

2. I-vector extraction

As proposed by P. Kenny [3], the Joint Factor Analysis model for session h belonging to speaker s can be written as:

$$m_{(s,h)} = m + Dy_s + Vz_s + Ux_h \tag{1}$$

where m is the speaker- and channel-independent supervector, which can be taken to be the UBM supervector, $m_{(s,h)}$ is the session-speaker dependent mean super-vector. U is the session variability matrix of low rank R (a $MD \times R$ matrix where M is the features dimension and G is the number of gaussians in the UBM) and x_h are the channel factors (an R vector). V is the speaker variability matrix of low rank S (a $MD \times S$ matrix) and z_s are the speaker factors (an S vector). D is a $(MD \times MD)$ diagonal matrix and y_s the residual speaker vector (a MD vector). Both, y_s, z_s and x_h are assumed to be normally distributed among $\mathcal{N}(0, I)$. $DD^t + VV^t$ represents the variability of the speaker mean super-vectors. UU^t represents the session variability. Hence, as we assume that the speaker and session variabilities are independent, the total variability is: $DD^t + VV^t + UU^t$.

By merging the information belonging to the subspaces generated by V and U and by ignoring the residual speaker component, we obtain:

$$m_{(s,h)} = m + Tw_{(s,h)} \tag{2}$$

where T is the low rank variability matrix of speaker and session and $w_{(s,h)}$ are the total variability factors. $w_{(s,h)}$ is assumed to be normally distributed among $\mathcal{N}(0, I)$. The T matrix is estimated iteratively using the algorithm detailed in [4] (in which the term Dy_s is ignored). We define $w_{(s,h)}$ obtained using the data frames of session h of the speaker s as i-vector.

3. Intersession compensation

We first present in paragraph 3.1 the i-vectors method proposed by N. Dehak [5][6]. Then we propose in paragraph 3.2 a new set of methods able to carry out channel compensation. These methods, based on new linear and non-linear transformations, prepare the data for new scoring method described in paragraph 4. To simplify notations, we will ignore subscripts h and s in this description. The following acronyms will be used: LDA for Linear Discriminant Analysis, NAP [7] for Nuisance Attribute Projection and WCCN [8] for Within-Class Covariance Normalization.

3.1. LDA+WCCN and cosine-fast scoring

N. Dehak proposed [5][6] to carry out channel compensation in the total factor space using several channel compensation techniques working in this space. The best performances were obtained by the process LDA+WCCN+Fast scoring. The Linear Discriminant Analysis (LDA) is a supervised method of dimensionality reduction. It defines new spatial axes that minimize the intra-class variance caused by channel effects and maximize the variance between speakers. So, i-vectors are subjected to the projection matrix obtained by LDA. The WCCN and cosine-fast scoring approach computes a cosine score, according to the metric of a within-class covariance matrix using a set of background impostors having several sessions each. This within class covariance matrix, introduced by Andrew Hatch [8] in the context of SVM classifiers, is calculated in a similar manner to the LDA within class covariance. For i-vectors, the calculation is done in the projected space of the LDA. Finally, a fast scoring calculates a cosine score between two test vectors according to the WCCN matrix. This i-vectors treatment method nowadays yields the best results in the most common speaker recognition evaluations.

3.2. Proposed method

The proposal approach aims to solve the three following points:

(i) The i-vectors w of eq.2 have to be theoretically normally distributed among $\mathcal{N}(0, I)$. This constraint produces i-vectors with independant dimensions and identical standard deviations, which is important for matricial-products-based scoring. It seems important for us to guarantee that this constraint is verified.

(ii) In [5] §3.2. and in [6] it is clearly shown that the channel effects carry out not only a linear deviation but also a non-linear dilatation of a given speaker vector (”radial” effect). In these works, the linear effect is removed thanks to the LDA (but simultaneously with a dimension reduction) before to cancel the dilatation effect using cosine-scoring. We consider that the radial effect should be first cancelled, independently of dimension reduction.

(iii) In [5] [6] the i-vectors issued from the FA dimensionality reduction technique are immediately projected by the LDA onto a low rank subspace (from 400 dimensions for total factor space to 200 for LDA). We consider that discriminant transformations could take advantage of the full rank total factor space compared to similar transformations in the LDA lower rank space.

3.2.1. I-vectors transformations

To deal with (i), we consider the difference between i-vectors (mean, covariance) and $(0, I)$ as a residue, which can be usefully removed, therefore that the i-vectors have to be standardized. To deal with (ii), i-vectors can be length-normalized. We chose to divide them by their norm but according to the accuracy matrix metric (inverse of covariance matrix). As desired in (iii), these two transformations do not involve a dimensionality loss. But when applying length-normalization on standardized data, the resulting i-vectors are no longer standardized, and reciprocally. We will see later in this paragraph how to reconcile the goals i) and ii).

Let $\overline{w}$ and V denote the empirical mean and covariance matrix of a broad set of training i-vectors. To apply standardization and length-normalization, the covariance matrix V is decomposed by diagonalization into PDP^t where P is the eigenvectors matrix of V -in columns- and D is the diagonal version of V. A train i-vector w is transformed to $w^{'}$ such that:

$$w^{'} = \frac{D^{-\frac{1}{2}} P^t \left(w - \overline{w}\right)}{\sqrt{\left(w - \overline{w}\right)^t V^{-1} \left(w - \overline{w}\right)}} \qquad (3)$$

Note that the numerator is equivalent by rotation to $V^{-\frac{1}{2}} \left(w - \overline{w}\right)$. The euclidean norm of $w^{'}$ is equal to 1. The same transformation is applied to the test i-vectors, using the training set parameters $\overline{w}$ and V as estimations of test set parameters.

As shown in Figure 1, from an initial training set (Fig.1 a.) previously centered, a rotation is applied (Fig.1 b.) around principal axes of total variability by application of P^t. Then, applying $D^{-\frac{1}{2}}$ achieves a standardization of vectors (Fig.1, c.). After length-normalization (Fig.1, d.) the i-vectors $w^{'}$ lie on the surface area of the unit hypersphere.

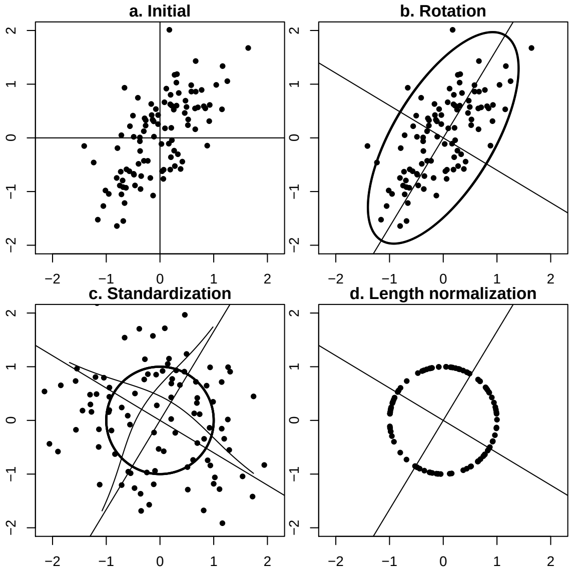

Figure 1: *Effect of the operations of standardization and length-normalization*

As noted before, transformed i-vectors are no longer standardized. But the previous process (compute $\overline{w}$ and V then apply eq.3) is iterated: then$\overline{w}$ tends to 0 and the V matrix tends to a diagonal matrix with identical diagonal values. To evaluate the convergence of V we measure the distance between the current occurence of V and the closest diagonal matrix with identical diagonal values. We use Frobenius matrix norm of the trace $A \mapsto \|A\|_F = \sqrt{tr(A^t A)}$ and compute the Least-squares error LSE between V and its projection onto $vect\,\{I\}$[1] :

$$LSE = \left\| V - \frac{tr\,(V)}{tr\,(I)} I \right\|_F \qquad (4)$$

Table 1: Distance (Least Square Error LSE) between covariance matrix V and family $vect\,\{I\}$ of diagonal matrices with identical diagonal-values, for 6 iterations of the transformation.

iter. 0 (LSE before iterations): 1×10^{-2}					
iter.	LSE	iter.	LSE	iter.	LSE
1	4×10^{-3}	3	1×10^{-4}	5	6×10^{-6}
2	6×10^{-4}	4	2×10^{-5}	6	1×10^{-6}

The LSE for the six first iterations are reported in Table 1. The distance between V and $vect\,\{I\}$ quickly converges to 0, hence the covariance matrix of i-vectors (which are length-normalized) becomes approximately diagonal with identical diagonal-values.

The transformations described above take into account the three initial goals i), ii) and iii). The resulting i-vectors are now well conditionned in order to apply a simple scoring method.

[1] $vect\,\{I\} = \{\lambda I,\ \lambda \text{ real}\}$

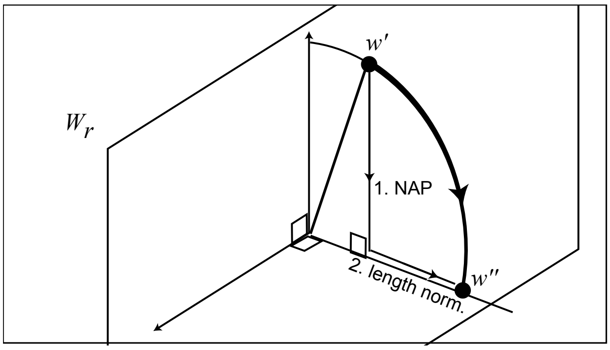

Figure 2: *Effect of radial-NAP technique*

Additional dimensionality reduction techniques can also be applied on these vectors.

3.2.2. Additional radial-NAP technique

The intersession compensation in the i-vectors space can still be improved using an additional dimensionality reduction technique. We present here a NAP technique adapted to our i-vectors lying on the surface area of an hypersphere, that we call radial-NAP.

Presented in [7], the usual NAP technique estimates session variability as a subspace of intermediate rank obtained using principal axes (eigenvectors having the largest eigenvalues) of the within-class covariance matrix, and projects the i-vectors into the orthogonal complementary subspace, assumed to be the speaker space. The within class covariance matrix W of the training set is calculated as follows:

$$W = \sum_{s=1}^{S} \frac{n_s}{n} W_s = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \left(w_i^s - \overline{w_s}\right) \left(w_i^s - \overline{w_s}\right)^t \quad (5)$$

where W_s is the covariance matrix of speaker s, n_s is the number of utterances for speaker s, n is the total number of utterances, w_i^s are the train i-vectors of sessions of speaker s and $\overline{w_s}$ their mean[2] .

In order to adapt NAP to our i-vectors conditioned to lie on an hypersphere surface area, we assume a radial distribution of a given speaker i-vectors. Radial-NAP follows this idea by first suppressing the nuisance dimensions and then normalizing the vector norm in order to stay on the hypersphere. Geometrically speaking, using radial-NAP, each i-vector is rotated to get orthogonal to the first r principal axes of the session-subspace. Figure 2 presents, in three dimensions, the geometrical result of this compensation technique.

Let $p_{W_r}(w')$ denote the projection of an i-vector $w^{'}$ onto the range of rank r of the matrix W. A train or test i-vector w' is transformed to $w^{''}$ such that:

$$w^{''} = \frac{w' - p_{W_r}(w')}{\|w' - p_{W_r}(w')\|} \quad (6)$$

4. Mahalanobis metric scoring

The training set of i-vectors can be classified according to the known class of the speaker. Given a new observation w, the goal of a statistical classifier is to identify to which class it belongs. If we assume homoscedasticity (equality of class covariances) and Gaussian conditional density models, the most likely class can be obtained by the Bayes optimal solution. An i-vector w is assigned to the speaker s that minimizes:

$$(w - \overline{w_s})^t \, W^{-1} \, (w - \overline{w_s}) = \|w - \overline{w_s}\|^2_{W^{-1}} \quad (7)$$

where $\overline{w_s}$ is the centroid (mean) of class s and W is the within class covariance matrix of eq.5. It is worth noting that, with these assumptions, the Bayesian approach is similar to the Fisher's geometric approach: w is assigned to the nearest centroid's class, according to the Mahalanobis metric of W^{-1}. To assess the suitability of this method on i-vectors we test it on our training set, described in section 5. The transformation process described in 3.2.1. is applied to the 12399 initial i-vectors from 890 speakers. Each transformed i-vector is then assigned to the nearest centroid's class, according to the W^{-1} metric. 12221 of the 12399 i-vectors (98.56 %) are properly classified.

Motivated by this result, we propose to use a Mahalanobis metric for the speaker detection scoring. The score between two i-vectors w_1 and w_2 is proportional to the log-probability that w_1 and w_2 belong to an unique class following the covariance matrix W. The centroid of this hypothetical class could be w_1 or w_2 or their mean, knowing that each proposition gives equivalent result. The final Mahalanobis-based scoring function is:

$$\mathit{score}\,(w_1, w_2) = -\, \|w_1 - w_2\|^2_{W^{-1}} \quad (8)$$

5. Experiments and results

5.1. Experimental setup

The background model in the experiments is the same as the background model in the LIA submission in the NIST-SRE-2006 campaign (male set only). Training is performed based upon Fisher database and consists of about 10 million of speech frames. Frames are composed of 19 LFCC parameters, its derivatives, and 11 second order derivatives (the frequency window is restricted to 300-3400 Hz). A normalization process is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance. The background model has 512 components whose variance parameters are floored to 50% of the global variance (0.5).

Speaker verification experiments are performed based upon the NIST SRE 2006 and 2008 databases, male speakers only (referred to as 2006 and 2008 protocol). The 2006 protocol consists of 354 speakers, 9720 tests (741 target tests, the rest are impostor trials). The 2008 protocol consists of 3798 speakers, 39433 tests (8290 target tests, the rest are impostor trials). Results are given in terms of equal-error-rate (EER) and the minimum DCF (an a posteriori decision). Train and test utterances contain 2.5 minutes of speech in average (around 30% of speech frames per session have been retained).

For the SVM system, the intersession variability matrix is enrolled on the NIST-SRE-2004 database with 2938 examples with 124 speakers (around 20 iterations to converge). From the same database, 200 impostor speakers are used for score normalization and as negative examples for the SVM classifier.

In the i-vector model, the total variability matrix T is trained using 12933 sessions from 890 speakers (NIST 2004, 2005 and Switchboard, about 15 sessions per speaker). Speaker models are derived by Bayesian adaptation of the Gaussian component means, with a relevance factor of 14. The same

[2] WCCN matrix used by N. Dehak [6] does not use $\frac{n_s}{n}$ in eq.5, considering that all the speakers contributions are equivalent.

database is used to estimate the inter-session matrix W in the i-vector space. The dimension of the i-vectors in the total factor space is 400.

5.2. Results

Our experiments were carried out on the telephone only data (det 3) for the core condition of the NIST 2006 SRE dataset, and on the telephone/non telephone data (det 4,5) and telephone only data (det 6,7,8) for the core condition of the NIST 2008 SRE dataset (only male part is used).

Tables 2 and 3 give comparison results for male gender between our baseline SVM-FA-ztnorm and three methods using i-vectors: LDA+WCCN with fast scoring using cosine kernel (best LDA dimensionality reduction is obtained with $r = 200$), three iterations of standardization + length-normalization, followed by Mahalanobis metric scoring, and three iterations then additional radial-NAP (best corank r is 50 for 2006 and 100 for 2008) always concluded by Mahalanobis metric scoring. Note that three iterations have been sufficient to obtain best performances. The LDA+WCCN+Fast scoring technique decreases performance of our baseline, which is in contradiction with [5]. The methods we propose in this paper give significant gains in all experimental conditions, in terms of EER and DCF. Radial-NAP yields the best EER performance for 2006 telephone data (Table 2) and for telephone/non telephone and ”all” telephone data 2008 (Table 3, det 5-6). The use of three iteration without radial-NAP yields the best performance for telephone only data 2008 (det 7-8) but only in terms of EER, the radial-NAP remaining the best in terms of DCF. We have also applied zt-normalization (the results are not presented). Score normalization does not bring additional gain, which shows the quality of scores produced by the proposed i-vectors transformations followed by Mahalanobis scoring.

Table 2: Comparison of results from SVM-FA-ztnorm method and methods using i-vectors: LDA + WCCN (rank 200) with fast scoring, three iterations of standardization + length-normalization, and three iterations followed by radial-NAP (corank $r = 50$). The results are given as EER and min.DCF × 100 on the male part of the core condition det 3 (telephone) of the NIST 2006 SRE.

NIST SRE 2006	det 3	
	EER	DCF ×100
baseline (SVM-FA-ztnorm)	2.16 %	1.06
LDA+WCCN+Fast scoring	2.93 %	1.24
stand.+length norm (3 iter.)	1.98 %	**1.00**
stand.+length norm (3 iter.)+radial-NAP	**1.69 %**	1.01

6. Conclusion

The total factor space contains both the speaker and the session information. We presented in this paper a set of simple linear and non-linear transformations to remove the session effects and a simple scoring technique based on a statistical classifier. Compared to our baseline and to LDA+WCCN+cosine scoring, the proposed method gives the best performances. All techniques of intersession compensation and scoring in the i-vectors space show both linear and non-linear natures of this variability, and the necessity of treating completely these two parts to achieve satisfying performances.

Table 3: Comparison of results from SVM-FA-ztnorm method and methods using i-vectors: LDA + WCCN (rank 200) with fast scoring, three iterations of standardization + length-normalization, and three iterations followed by radial-NAP (corank $r = 100$). The results are given as EER and min.DCF × 100 on the male part of the core condition (telephone/non telephone det 4,5 and telephone only det 6,7,8) of the NIST 2008 SRE.

NIST SRE 2008 - **EER (%)**					
	det4	det5	det6	det7	det8
baseline (SVM-FA-ztnorm)	10.44	6.40	6.29	2.72	1.31
LDA+WCCN+ Fast scoring	7.74	6.56	6.63	3.64	2.63
stand.+length norm (3 iterations)	**5.24**	5.61	5.72	**2.03**	**0.95**
stand.+length norm (3 iter.)+radial-NAP	5.68	**4.05**	**5.51**	2.50	1.21

NIST SRE 2008 - **DCF ×100**					
	det4	det5	det6	det7	det8
baseline (SVM-FA-ztnorm)	3.65	2.41	3.57	1.54	1.01
LDA+WCCN+ Fast scoring	3.57	2.66	**3.34**	1.83	1.37
stand.+length norm (3 iterations.)	**2.78**	2.76	3.37	1.63	1.09
stand.+length norm (3 iter.)+radial-NAP	2.82	**2.26**	3.42	**1.50**	**0.85**

7. References

[1] D. A. Reynolds, “A gaussian mixture modeling approach to text-independent speaker identification“, Ph.D. thesis, Georgia Institute of Technology, August 1992.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification“, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 5, pp.980-988, July 2008.

[3] P. Kenny, “Joint Factor Analysis of Speaker and Session Variability : Theory and Algorithms,“ CRIM, Montreal, (Report) CRIM-06/08-13, 2005.

[4] D. Matrouf, N. Scheffer, B. Fauve, and J.F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification“, in International Conference on Speech Communication and Technology, 2007.

[5] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification“, in INTERSPEECH, Brighton, UK, Sept 2009.

[6] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, “Front-end Factor Analysis for Speaker Verification“, IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788-798, August 2010.

[7] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation“, in IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, 2006, vol. 1. pp. 97-100.

[8] A. Hatch, S. Kajarekar, and A. Stolcke, “Withinclass covariance normalization for svm-based speaker recognition“, INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing - ICSLP, vol. 3, pp. 1471-1474, 2006.