Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System

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

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

Abstract. This paper focuses on the analysis of the i-vector paradigm, a compact representation of spoken utterances that is used by most of the state of the art speaker verification systems. This work was mainly motivated by the need to quantify the impact of their steps on the final performance, especially their ability to model data according to a theoretical Gaussian framework. These investigations allow to highlight the key points of the approach, in particular a core conditioning procedure, that lead to the success of the i-vector paradigm.

1 Introduction

Recent advances in speaker verification have revealed the discriminant power of a new representation of spoken utterances, referred as i-vector[1]. Easy to work with and bringing back the speaker recognition problem to a more traditional biometric pattern recognition problem, i-vectors are now largely used in the most recent speaker verification systems. A classical i-vector system can be briefly decomposed in three stages. First, the acoustic space is structured using the GMM-UBM approach [2] and each speech utterance is represented by a high-dimensional representation denoted "'supervector"'. Then, a low-dimensional representation of this supervector is extracted thanks to a factor decomposition approach. Lastly, a scoring module obtains the final score for a given test, taking advantages of the compact speech utterance representation. Quite often, an additional data conditioning procedure is applied before the scoring step.

The goal of this paper is to assess the impact of each of these stages in terms of global performance. This is important as i-vector approach allows in the past years a drastic progress in terms of performance. A better understanding of the origins of these progresses should allow further improvements and/or some simplification in the quite complex chain of processing. More precisely, we wish to quantify the role of the optional conditioning procedure as we suspect that this module plays a more important role than expected in the performance of i-vector systems.

At all three stages, data modelings have been designed to meet the constraints of a parametric approach, based on Gaussian probabilistic assumptions. The conditioning procedure is also known to help achieve these modeling goals.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 278–285, 2013.

To examine independently each of the stages, we proceed by replacing one by one these modules by methods based on deterministic or non-parametric approaches. The gaps of performance are compared with that involved by the conditioning procedure, then summarized in order to assess the impact of the different approaches. Moreover, replacing methods by others measures the robustness of concepts on which they rely. Results of these investigations can thus highlight the key points in the chain of processing that lead to the success of the i-vector paradigm.

The paper is organized as follows: Sections 2, 3, 4 describe the i-vector based speaker verification system on which we focus. Section 5 presents the alternative methods used at each stage of the system. The experimental results are presented and commented in Sections 6, 7 and conclusions are drawn in Section 8.

2 GMM Framework and i-Vector Extraction

Speaker information is modeled by using the Gaussian Mixture Model/Universal Background model (GMM/UBM) paradigm [2] where a weighted sum of Gaussian distributions performs a direct acoustic modeling of the acoustic space. A model of a given speech segment is represented by the Baum-Welch zero and first order statistics of its feature vectors, according to UBM prior distribution. This model is denoted "‘supervector"’. The i–vector model [3] constrains the supervector **s** of a given speech segment to live in a single subspace following the linear model of a Factor Analysis:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where **m** is the supervector corresponding to the UBM, **T** is a low-rank rectangular matrix with $G \times F$ rows and r columns, G and F are the number of GMM components and feature dimension, respectively. The r columns of **T** are vectors spanning the “total variability” space, and **w** is a random vector of size r having a standard normal prior distribution. Determination of **T** by using EM-ML procedure and explicit formula of the extracted i-vector **w** can be found in [1].

3 I-Vector Models and Scorings

The first i-vector based speaker verification systems were based on the LDA–WCCN approach [1], which performs intersession compensation thanks to Linear Discriminant Analysis (LDA) [1], where all the i-vectors belonging to the same speaker are associated with the same class. This technique projects the input data into a much lower dimensional space with minimal loss of discriminative ability, as the ratio of between-speaker and within-speaker variations is maximized. These speaker features are finally normalized by a Within Class Covariance Normalization (WCCN) [4]. The final scores are then computed using a cosine distance scoring [3].

A key evolution of i-vector approach was introduced in [5], using the Probabilistic Linear Discriminant Analysis (PLDA) [6]. Two assumptions on the prior probability distributions of the PLDA variables (speaker, session and residual factors of eq. 7 in [7]) have been proposed:

- Gaussian PLDA (G-PLDA) assumes that all latent variables are statistically independent. Standard normal priors are assumed for speaker and session factors. The residual term is assumed to be Gaussian with zero mean and diagonal covariance matrix.
- Student's t-distribution is proposed in [5] as an alternative to the Gaussian to model the speaker and channel subspaces in the i-vector space. Heavy-tailed PLDA (HT-PLDA) assumes that all the factors follow an heavy-tailed distribution, scaled by gamma distribution scalars.

The ML point estimates of the model parameters are obtained from a large collection of development data using an EM algorithm as in [6].

4 Pre-conditioning

A pre-processing before any i-vector modeling has been introduced in [8][9]. I-vectors are whitened and length-normalized, in order to make them more Gaussian. The most commonly used whitening technique is a standardization, and the transformation applied to an i-vector $\mathbf{w}$ can be resumed as follows:

$$\mathbf{w} \leftarrow \frac{\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)}{\left\|\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)\right\|} \tag{2}$$

where μ and $\mathbf{A}$ are the mean and a variability matrix of a training corpus. Data are standardized according to a variability matrix $\mathbf{A}$ then length-normalized, confining the i-vectors to the hypersphere of unit radius. Parameters are computed for the i-vectors present in the training corpus and applied to the test i-vectors. The matrix $\mathbf{A}$ can be the total covariance matrix or, as we proposed in [7], the within-class covariance matrix $\mathbf{W}$ defined in eq. 4 of [7].

In [9], it is shown that this technique improves the gaussianity of the i-vectors. It reduces the gap between the underlying assumptions on the data distribution and the real distribution and also reduces the dataset shift between development and trial i-vectors. Moreover, it is shown in [9] that performance of a G-PLDA system with this pre-conditioning is competitive versus the HT-PLDA, when the latter is much more complicated. As proposed in [8][7], these two-steps can be iterated. As a result, i-vectors tend to be simultaneously $\mathbf{A}$-standardized and length-normalized (magnitude 1), involving a number of properties related to intersession compensation. Some of them are detailed in [8][7]. Also in [7], we propose, after $\mathbf{W}$-standardization, a deterministic initialization of PLDA matricial metaparameters $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ of eq. 7 in [7]. It allows a faster convergence of the PLDA EM-ML procedure.

5 Alternative Methods

The state of the art i-vector-based system described below is composed of three stages: representation of segments by Baum-Welch zero and first order UBM-statistics, i-vector extraction using Factor Analysis total-variability (*FA-total-var*), Gaussian-PLDA modeling and scoring with an optional pre-conditioning. We present here the alternative methods that we have implemented for each of these three stages.

5.1 Models and Scorings

To analyze the efficiency of Gaussian-PLDA, we compare this probabilistic modeling with two simplified and deterministic versions. First, the LDA-two-covariance model [10] reduces the dimensionality by using LDA, then full rank matrices $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ of eq.7 in [7] are deterministically estimated (no EM-ML procedure is performed) by $\mathbf{\Phi} = \mathbf{B}^{\frac{1}{2}}$ and $\mathbf{\Gamma} = \mathbf{W}^{\frac{1}{2}}$ where $\mathbf{B}$ and $\mathbf{W}$ are the between- and within-class covariance matrices defined in eq. 3, 4 of [7]. Comparing Gaussian-PLDA and LDA-two-covariance model measures the gain of the probabilistic ML-approach in a generative i-vector modeling. Second, the LDA-Mahalanobis model, introduced in [8] is a particular case of the previous two-covariance model which makes no assumption about the speaker factor distribution (speaker precision matrix $\mathbf{B}^{-1}$ is null). The deterministic Mahalanobis model is useful to estimate the relevance of a between-speaker modeling.

5.2 I-Vector Extraction

Factor analysis total variability (FA-total-var) is the state of the art factor decomposition technique used to extract i-vectors. To assess the pertinence of its probabilistic approach, we compare it with the well-known deterministic principal component analysis (PCA). But FA-total-var is based on zero and first order statistics and applying PCA to extract low dimensional vectors (that we will also call *i-vectors*) needs to determine the unique high-dimensional vectorial representation to compress. Some solutions have been suggested [11]. In order to fairly compare probabilistic FA-total-var and deterministic PCA, we introduce an adapted version $\widehat{\mathbf{s}}$ of a supervector $\mathbf{s}$, equal to:

$$\widehat{\mathbf{s}} = N_{\mathcal{X}} \left(\mathbf{\Sigma} + N_{\mathcal{X}}\right)^{-1} \left(\mathbf{s} - \mu\right) \tag{3}$$

$N_{\mathcal{X}}$ is the $GF \times GF$ diagonal matrix composed of F blocks of $N_{\mathcal{X}}^{(g)}\mathbf{I}$ ($g = 1, ..., G$) where $N_{\mathcal{X}}^{(g)}$ are the zero–order statistics estimated on the g-th Gaussian component of the UBM observing the set of feature vectors in the sequence $\mathcal{X}$, and μ and $\mathbf{\Sigma}$ are the UBM mean and diagonal covariance matrix.

In the extreme case of a square and full rank identity matrix $\mathbf{T}$ (no dimensionality reduction applied), eq. 6 of [1] shows that FA-extraction provides an i-vector $\mathbf{w}$ equal to $\widehat{\mathbf{s}}$.

The supervector $\widehat{\mathbf{s}}$ is an adapted version of $\mathbf{s}$, centered and weighted by the amount of informations per Gaussian-component and by the variance per dimension.

5.3 UBM-Based Representation

In [12][13] a new approach for speaker recognition, denoted "Speaker Binary Key", was presented. Contrary to classical speaker recognition based on statistical modeling of the speaker information, this approach proposes to handle directly each piece of speaker specific information in a binary space. Each coefficient of this binary space corresponds to a targeted piece of speaker-specific information which could be present (the coefficient is equal to 1) or non present (the coefficient is equal to 0) in a given acoustic frame or acoustic segment. This new approach allows to exploit temporal or sequential information as a binary vector is extracted for each acoustic frame. It also focuses on speaker specific information in a non-parametric way as each coefficient of the binary space models speaker-specific information. As the binary key representation first ties each input frames with one or several GMM-UBM components (before non-parametric transformation to a binary space), it constitutes a GMM-UBM-based alternative to the zero and first order statistics. High-dimensional binary keys provided by this model are projected onto a PCA subspace (by the lack of a specific Factor Analysis), and handled as i-vectors for modeling and scoring.

6 Experimental Setup

The feature extraction and the 512-components GMM-UBM functionalities used in our experiments are described in [8]. For i-vector extraction, the total variability matrix $\mathbf{T}$ is trained using 15660 speech utterances from 1147 speakers (NIST 2004-05-06, Switchboard II part 1, 2 & 3; Switchboard cellular part 1 & 2, about 14 sessions per speaker). The results are reported with 400-dimensional i-vectors. The same database is used to estimate the parameters of the i-vector models and scorings. In PLDA, channel factor is kept full and speaker factor is varied, as proposed in [5]. Evaluation was performed on the NIST SRE 2008 DET conditions 6 and 7, male only, corresponding to telephone-telephone (all and English-only respectively) enrollment-verification trials, and on the NIST SRE 2010 DET extended condition 5, male only, corresponding to telephone-telephone. A global measurement of performance of a system is given by the average of the three Equal Error Rates (EER). These three conditions are the most currently used in the domain and their average EER is a robust performance measure of a system.

7 Results

Table 1 shows comparison result of systems applying the different representations, extractors, models and scorings listed above. The first eight systems use

Table 1. Comparison of performance, in terms of EER (%), between systems based on different representations, extractors, models and scorings (without and with pre-conditioning)

	repr.	extract.	conditioning	model and scoring	det 7	det 6	det 5 ext	**average**
1	sv	FA	no	LDA-Maha	5.70	9.5	9.73	**8.31**
2	sv	FA	no	LDA-two-cov	3.23	6.83	5.97	**5.34**
3	sv	FA	no	G-PLDA	3.39	6.37	6.38	**5.38**
4	sv	FA	WCCN-cosine	LDA-WCCN-cosine	3.26	6.29	3.69	**4.41**
5	sv	FA	$\mathbf{L\Sigma}$	LDA-Maha	1.86	5.06	2.62	**3.18**
6	sv	FA	$\mathbf{L\Sigma}$	LDA-two-cov	1.53	4.93	2.36	**2.94**
7	sv	FA	$\mathbf{L\Sigma}$	G-PLDA	1.63	4.80	2.45	**2.96**
8	sv	FA	L**W**	G-PLDA	1.58	4.80	2.28	**2.89**
9	BK	PCA	no	G-PLDA	2.84	5.82	4.42	**4.36**
10	sv	PCA	no	G-PLDA	3.17	6.59	5.80	**5.19**
11	BK	PCA	L**W**	G-PLDA	2.16	5.26	2.87	**3.43**
12	sv	PCA	L**W**	G-PLDA	1.99	5.24	2.47	**3.23**

high-dimensional representation by zero and first order UBM statistics (**sv** for supervector) and Factor Analysis on total variability (**FA**) as i-vector extractor. Performance are given without (**no**) and with pre-conditioning: $\mathbf{L\Sigma}$, L**W** for standardization according to total $\mathbf{\Sigma}$ or within-class **W** covariance matrix, or **WCCN-cosine** as implicit normalization of LDA-WCCN-cosine scoring. HT-PLDA scoring has not been carried out, as pre-conditioning and Gaussian-PLDA are able to match its performance.

The state of the art system (line 8) yields the best result: average EER of 2.89 and best EERs for all the individual conditions. But, first, the gap between ML (lines 7 and 8) and deterministic approach (line 6) for i-vector modeling is slight or null (average EER of 2.89 and 2.96 vs 2.94). This observation is strengthened by the fact that the best system (line 8) deterministically initializes PLDA metaparameters then requires only 10 EM-ML iterations to converge, against 100 using the randomly initialized system (line 7). Second, comparison of systems without and with pre-conditioning shows that the quality of the modeling is, in a major proportion, the consequence of the conditioning: 5.34 to 2.94 for the best deterministic approach, 5.38 to 2.89 for the probabilistic approach. It is worth noting that the gap between the less efficient system (LDA-Mahalanobis) and the others is particularly significant in the absence of pre-conditioning (8.31 vs 5.34 without, 3.18 vs 2.94 with). This shows that the initial lack of gaussianity in the extracted i-vectors is mainly due to the within-speaker distribution.

The four last lines give comparison result between systems using representation by speaker binary key (**BK**) or by zero and first order UBM statistics, all using i-vector extraction techniques by PCA (**PCA**), each time without and with pre-conditioning (L**W** only, since it gives the better performance in the previous experiments). Comparing the extraction techniques (lines 8 and 12), FA brings a relative improvement of 10.5% of average EER: 2.89 vs 3.23 with PCA.

This slight gain recalls that i-vector extraction falls into the family of compression techniques rather than factor decompositions. Comparing representations for PCA-based systems (lines 11 and 12), the binary key representation yields a performance close to that of zero and first order UBM statistics (3.43 vs 3.23) with, which must be taken into account, a 32 times lower amount of information[1]. But once again, the improvement of performance is mainly due to the conditioning step. Systems based on different representations and dimensionality reductions are able to provide interesting performance but only if they include a pre-conditioning procedure.

8 Conclusion

The aim of this work was to assess the benefits of the different steps in a classical i-vector based speaker verification system. In particular, we quantify the role of the optional conditioning procedure in the good probabilistic modeling of data. As all stages of the system try to take into account the constraints of a Gaussian framework, we replace one by one these modules by a deterministic or non-parametric method and compare the gap of performance with that involved by the conditioning procedure. These comparisons also allow to measure the robustness of concepts involved in the i-vector approach. The results of this analysis can be summarized by the following key points:

- All the systems presented here rely on the GMM-UBM. Their good performance, following however various ways, show the robustness of the GMM-UBM to structure the acoustic feature space.
- High-dimensional UBM-based representations are stacking a fixed-length set of vectors from the feature space. The low gaps between systems with various representations and extractors show that any dimensionality reduction of stacked vectors built by using UBM, according to the total variability, is able to capture and summarize correlated behaviors between UBM-components. As remarked in the introduction of [14], the i-vector random variables can be viewed as principal components of utterances. The coordinates represent physical quantities, which are constant for a given utterance but which differ from one utterance to another.
- Resulting low-dimensional vectors do not match the assumptions of an usual probabilistic framework. More than FA-total-var or PLDA decompositions, the conditioning procedure mainly contributes to make vectors compatible with a linear-Gaussian modeling and scoring. WCCN-cosine-scoring can be decomposed into an inner-product applied to standardized and length-normalized vectors, as done in eq. 2. A core procedure, composed of standardization according to a target variability, followed by length-normalization

[1] In our configuration of 512-components GMM-UBM, 50-dimensional feature space and, for binary modeling, 100 specificities per component, the size of a binary key is 6.4 KB and the size of double precision zero and first order UBM-statistics is 208.9 KB.

(which ignores the magnitude to focus on the directional information), turns out to be decisive in the final performance.

Works about the properties of the conditioning and dimensionality reduction procedures are presented in [1][9][8][7]. But we are now continuing a thorough study of their properties, in order to better explain their impact in the performance and improve further i-vector based speaker verification systems.

References

1. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19, 788–798 (2011)
2. Reynolds, D.A.: A Gaussian mixture modeling approach to text-independent speaker identification. PhD thesis, Georgia Institute of Technology (1992)
3. Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P.: Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In: International Conference on Speech Communication and Technology, pp. 1559–1562 (2009)
4. Hatch, A.O., Kajarekar, S., Stolcke, A.: Within-Class Covariance Normalization for SVM-based Speaker Recognition. In: International Conference on Speech Communication and Technology, pp. 1471–1474 (2006)
5. Kenny, P.: Bayesian speaker verification with heavy-tailed priors. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2010)
6. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
7. Bousquet, P.M., Larcher, A., Matrouf, D., Bonastre, J.F., Plchot, O.: Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)
8. Bousquet, P.M., Matrouf, D., Bonastre, J.F.: Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In: International Conference on Speech Communication and Technology, pp. 485–488 (2011)
9. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of i-vector length normalization in speaker recognition systems. In: International Conference on Speech Communication and Technology, pp. 249–252 (2011)
10. Brummer, N., de Villiers, E.: The speaker partitioning problem. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2010)
11. Campbell, W.M., Sturim, D., Borgstrom, B.J., Dunn, R., McCree, A., Quatieri, T.F., Reynolds, D.A.: Exploring the impact of advanced front-end processing on nist speaker recognition microphone tasks. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)
12. Bonastre, J.F., Bousquet, P.M., Matrouf, D., Anguera, X.: Discriminant binary data representation for speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 5284–5287 (2011)
13. Bonastre, J.F., Anguera, X., Sierra, G.H., Bousquet, P.M.: Speaker modeling using local binary decisions. In: International Conference on Speech Communication and Technology, pp. 485–488 (2011)
14. Kenny, P.: A small footprint i-vector extractor. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)