On the importance of pre- and post-conditioning procedures for speaker recognition systems based on total variability factors

Pierre-Michel Bousquet, Jean-François Bonastre

University of Avignon - LIA, France

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

Abstract

The total variability factor approach in the speaker recognition field draws on the reduction of a high-dimensional representation of utterances based on the Gaussian Mixture Model of Universal Background model (GMM-UBM) paradigm. Once the low-rank total variability factor (referred to as *i-vector*) has been extracted, Gaussian modelings of speaker and session have proven to be efficient in terms of performance, but only if conditioning procedures are applied to vectors before modeling. We show in this article that pre-conditioning procedures (applied to high-dimensional representation before or during i-vector extraction) are also needed to achieve the best performance. The impact assessment of these pre- and post-procedures is evaluated, with the state-of-the-art representation and with the semi-parametric GMM-UBM based representation referred to as *binary key*. This evaluation shows the importance of both procedures, which lead to a relative reduction of 48% and 30% in detection errors.

1. Introduction

The i-vector representation of speech utterances provides a feature vector of low dimension (less than 600), independent of the length of the utterance. The i-vector solution first relies on a Gaussian mixture model (GMM-UBM)-based high-dimensional representation of utterances. Then, a dimensionality reduction is applied to this representation, extracting low-rank factors referred to as i-vectors. The Bayesian generative model designed to provide a consistent probabilistic framework for i-vectors is the PLDA. Gaussian PLDA (G-PLDA) assumes that speaker and residual components have Gaussian distributions. To deal with severe within-class distortions and increase the robustness to outliers, post-conditioning procedures (after extraction) are applied to i-vectors, which we contributed to introduce in the field. We analyzed in a previous article [1] the importance of these post-conditioning procedures. In this article, we complete this study by revealing the importance of pre-conditioning procedures in the quality of systems. These procedures are implicit in Factor analysis. We uncover, in this probabilistic dimensionality reduction, a standardization of supervectors and a kind of normalization according to 0^{th} order statistics (the probabilistic quantity of information per Gaussian component). To better assess the relevance of pre-conditioning procedures, we use an alternative representation of utterances, the speaker binary keys, which we designed and contributed to enhance. This approach draws on GMM-UBM but differs of the statistical modeling of the speaker information by using a semi-parametric way. Total variability factors can be extracted from binary keys and we show that the post-conditioning procedures and PLDA modeling carried out for i-vectors can be applied to these factors. We present a new pre-conditioning procedure suited to binary keys, referred to as "Equalization per Gaussian Component", and justify its use in the semi-parametric context of the binary key model. An evaluation of systems based on these different representations, with or without pre- and post-conditioning, is carried out. The summary of these experiments assesses the impact of the conditioning procedures in a speaker recognition system based on total variability factor.

2. GMM-UBM representations

2.1. Baum-Welch 0^{th} and 1^{th} order statistics

Let G be the number of components in the GMM-Universal Background model (UBM) and F the dimension of the acoustic feature vectors. We denote by $\mathcal{X} = \{x_t\}_t$ the entire collection of labeled frames for a segment. The 0^{th} order statistics of the segment are the G dimensional vector whose the g^{th} component is:

$$\sum\nolimits_t \gamma_g(t) \tag{1}$$

where $\gamma_g(t)$ is the occupation probability of x_t for the g^{th} component.

Let $N_{\mathcal{X}}$ be the $FG \times FG$ block diagonal matrix whose $F \times F$ diagonal blocks are $n_g\mathbf{I}$ where $\mathbf{I}$ denotes the $F \times F$ identity matrix and:

$$n_g = \sum\nolimits_t \gamma_g(t) \tag{2}$$

n_g is the number of frames which are accounted for by the given mixture component.

The 1^{th} order centered statistics of the segment are the FG-dimensional vector $S_{\mathcal{X}}$ obtained by stacking the F-dimensional component vectors $S_{\mathcal{X},g}$ defined by:

$$S_{\mathcal{X},g} = \sum\nolimits_t \gamma_g(t)(x_t - \mu_g) \tag{3}$$

where μ_g is the g^{th} component of the *world* (UBM) overall mean μ.

The supervector of the collection $\mathcal{X}$ is the FG-dimensional vector s of adapted world means, obtained by:

$$S_{\mathcal{X}} = N_{\mathcal{X}}(s - \mu) \tag{4}$$

2.2. Speaker binary key

In [2][3] a new approach for speaker recognition, denoted *Speaker Binary Key*, was presented. Unlike 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. 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 0^{th} and 1^{th} order statistics. Detailed descriptions of this new approach can be found in [2][3].

3. I-vector extraction

3.1. With 0^{th} and 1^{th} order statistics

The i-vector paradigm assumes that a supervector can be modeled as:

$$s = \mu + \mathbf{T}\mathbf{w} \tag{5}$$

where μ is the world mean supervector, $\mathbf{T}$ is a $FG \times p$ matrix (where $p << FG$) of bases spanning the subspace covering the important variability (both speaker- and session-specific) in the supervector space, and $\mathbf{w}$ is the standard-normally distributed i-vector.

The matrix $\mathbf{T}$ is estimated by using Expectation-Maximization algorithm based on maximum of likelihood (EM-ML). With the same notations than in paragraph 2.1, the extracted i-vector $\mathbf{w}$ is equal to:

$$\mathbf{w} = \left(\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}_{\mathcal{X}}\mathbf{T}\right)^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}_{\mathcal{X}}\left(s-\mu\right) \tag{6}$$

where Σ is the world covariance matrix.

3.2. With speaker binary keys

The high-dimensional binary keys provided by this model are projected onto a PCA subspace and handled as i-vectors for modeling. Note that the 0 and 1 values of a binary key indicate that the specificities have or not to be selected for representing the utterance, thus are occurrences of a categorical variable. Therefore, the dimensions of a binary key are categorical variables. The equivalent-to-PCA technique for categorical variables is the *Multiple Correspondence Analysis* (MCA) but it turns out that in the special case of binary variables (all variables have only two levels), one demonstrates that the MCA is equivalent to PCA on the binary coded vectors: the eigenvectors provided by both techniques are identical.

4. Gaussian-PLDA

Introduced in [4] and adapted for speaker recognition in [5], Gaussian Probabilistic Linear Discriminant Analysis (PLDA) is a generative i-vector model. The most common PLDA model in speaker verification assumes that each p-dimensional i-vector $\mathbf{w}$ of a speaker s can be decomposed as

$$\mathbf{w} = \mu + \mathbf{\Phi}\mathbf{y}_s + \varepsilon \tag{7}$$

The mean vector μ is a global offset, $\mathbf{\Phi}$ is a $p \times r$ matrix whose columns provide a basis for the eigenvoice subspace, the r-dimensional vector y_s is the speaker factor and ϵ is the residual term. Therefore, the speaker-specific part $\mu + \mathbf{\Phi}\mathbf{y}_s$ represents the between-speaker variability and is assumed to be tied across all utterances of the same speaker. G-PLDA assumes that all latent variables are statistically independent. Standard normal prior is assumed for the speaker factor y_s and normal prior for the residual term ϵ with mean 0 and full covariance matrix $\mathbf{\Lambda}$. The maximum of likelihood (ML) point estimates of the model parameters are obtained from a large collection of development data using an expectation-maximization (EM) algorithm as in [4].

5. I-vector conditioning procedures

5.1. Post-conditioning

A post-conditioning procedure step (applied immediately after i-vector extraction) has been introduced in [6, 7], following WCC Normalization and cosine-scoring technique of [8]. 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 summarized as follows:

$$\mathbf{w} \leftarrow \frac{\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)}{\left\|\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)\right\|} \tag{8}$$

Data are standardized according to the mean μ and a variability matrix $\mathbf{A}$ of a training corpus, then length-normalized. Parameters are computed for the i-vectors present in the training corpus and applied to test i-vectors. The matrix $\mathbf{A}$ can be the total covariance matrix $\mathbf{\Sigma}$ or, as proposed in [9], the within-class covariance matrix $\mathbf{W}$. We denote $\mathbf{L\Sigma}$, $\mathbf{LW}$ these transformations. Details on their properties and benefits can be found in [7, 6, 9, 1].

5.2. Pre-conditioning

5.2.1. With 0^{th} and 1^{th} order statistics

As matrices $\mathbf{N}_{\mathcal{X}}$ and $\mathbf{\Sigma}$ of Equation 6 are diagonal, this equation can be rewritten:

$$\mathbf{w} = \left[\left(\mathbf{I} + \widetilde{\mathbf{T}}^t\mathbf{N}_{\mathcal{X}}\widetilde{\mathbf{T}}\right)^{-1}\widetilde{\mathbf{T}}^t\mathbf{N}_{\mathcal{X}}\right]\mathbf{\Sigma}^{-\frac{1}{2}}\left(s-\mu\right) \tag{9}$$

where $\widetilde{\mathbf{T}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{T}$. As is shown by this equation, FA-extraction includes an implicit pre-conditioning procedure. On the one hand, $\mathbf{w}$ is obtained by projecting the standardized version $\mathbf{\Sigma}^{-\frac{1}{2}}\left(s-\mu\right)$ of supervector s [1]. On the other hand, the projection matrix $\widetilde{\mathbf{T}}$ is adapted depending on the amount of informations per Gaussian component, expressed in $\mathbf{N}_{\mathcal{X}}$.

To assess the importance of this pre-conditioning procedure, we carry out a simple *Probabilistic Principal Component Analysis* PPCA [10],[11] on supervectors, non previously standardized. This dimensionality reduction model extracts i-vector by using an EM-ML procedure under Gaussian assumptions. A unique projection matrix is estimated then applied to any supervectors, without taking into account the amount of informations per Gaussian component.

5.2.2. With speaker binary keys

The speaker binary key model first draws on a discrete coverage of the features space (each point of this coverage is referred to as

[1] according to the world mean and covariance matrix.

"specificity"). Let c denote the Gq-dimensional count vector of a segment, where G is the number of components in the GMM-UBM and q the number of specificities per Gaussian component (q is a parameter). Given the collection of frames of an utterance, a *count* vector is generated which cumulates the amount of calls to most-likely specificities per frame. Let $\mathcal{C}$ denote the index subset of the n highest values of c (n is a parameter). The binary key representation b of the overall segment is generated by setting to 1 the subset $\mathcal{C}$ of b, 0 otherwise.
Let $c_{g,k}$ denote the value of c for the k^{th} specificity of the g^{th} GMM component. We propose to introduce a pre-conditioning transformation, inserted between the determinations of c and b. The binary key b is generated by using a conditioned version $\widehat{c}$ of c, defined by:

$$\widehat{c}_{g,k} = \frac{c_{g,k}}{\sum_{k=1}^{q} c_{g,k}} \tag{10}$$

The count vector c becomes a *frequency* vector $\widehat{c}$ (the sum of values of $\widehat{c}$ for a Gaussian component is equal to 1). We refer to this pre-conditioning procedure as *Equalization per Gaussian Component*.

Figure 1 shows an example of the effect of this conditioning procedure. The GMM-UBM is comprised of two components (left and right sides of the figure). Each of these latter is covered by 29 specificities. The vertical bars depict the count values of a segment. The $n = 8$ highest count values set to 1 the binary key of the segment. The 0 and 1 values of the resulting binary key are indicated below the x-axis. Without equalization (top-side), only specificities of the first component are selected: a region of high density and a unique pick of density. The specificities of the second component have too little count values to be selected. This approach is close to the statistical approach of the 0^{th} and 1^{th} order statistics of the GMM-UBM. With equalization (bottom side), specificities of the second component are also selected, revealing picks of energy (high density specificities compared to their neighborhood). Moreover, the high density region of the first component is summarized by only four specificities. This approach is motivated by the double aim to fulfill the statistical assumptions of "confident interval" (the latter region of component 1 is a sufficiently likely region of high density to narrow its amplitude) and the non-parametric logic of "exception" (taking into account local abnormality, like the single selected specificity of the second component, which is locally significant).

6. Experimental setup

The feature extraction and the 512-components GMM-UBM functionalities used in our experiments are described in [7]. 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 PLDA model. 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. Robust measurements of performance of systems are given by the averages of the three Equal Error Rates (EER) and of the three minimal

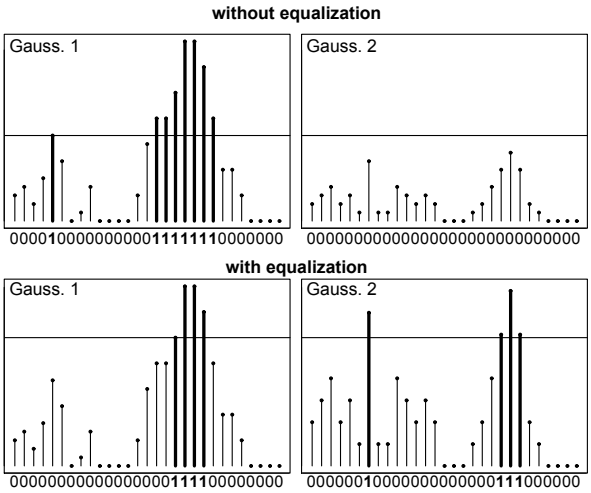

Figure 1: An illustration of the strategy of "Equalization per Gaussian component" pre-conditioning technique

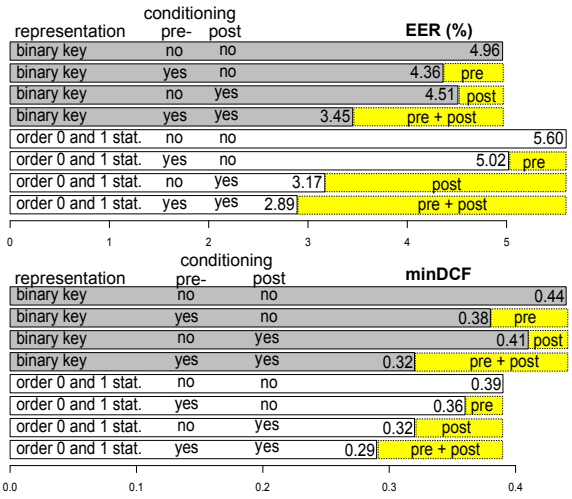

Figure 2: Averages of EER and minimal DCF for the same systems than Table 1. The gains of performance obtained by using conditioning procedures are depicted by yellow colored rectangles.

DCF as defined by NIST [2].

7. Results

Results of the impact assessment of pre- and post-conditioning procedures are summarized in Table 1. For the binary key (first four lines) and for the 0^{th} and 1^{th} order statistics (last four lines) representations, EERs and minimal DCFs (minDCF) are displayed for all the combinations of pre- and post-conditioning procedures (each one applied or not). The averages of EER and minimal DCF are embolded.
Figure 2 illustrates the same average values for each of the eight evaluated systems. The gains of performance obtained by using conditioning procedures are displayed on the figure (yellow colored rectangles). For example, comparing the first two systems, the average of EER for binary keys moves from 4.96%

[2] NIST SRE10 evaluation plans are available at `http://www.itl.nist.gov/`

NIST SRE			2008				2010			
high dim. repr.	pre-condit.	post-condit.	det 7		det 6		det 5 Ext		mean	mean
			EER	DCF	EER	DCF	EER	DCF	EER	DCF
binary key	-	-	3.41	0.20	6.88	0.38	4.59	0.75	**4.96**	**0.44**
binary key	**yes**	-	2.94	0.16	5.83	0.33	4.42	0.65	**4.40**	**0.38**
binary key	-	**yes**	3.02	0.18	6.75	0.36	3.76	0.69	**4.51**	**0.41**
binary key	**yes**	**yes**	2.20	0.13	5.26	0.29	2.89	0.54	**3.45**	**0.32**
0^{th} and 1^{th} order stat.	-	-	3.87	0.18	6.64	0.35	6.29	0.63	**5.60**	**0.39**
0^{th} and 1^{th} order stat.	**yes**	-	3.18	0.17	6.41	0.32	5.48	0.59	**5.02**	**0.36**
0^{th} and 1^{th} order stat.	-	**yes**	1.76	0.13	5.30	0.30	2.45	0.52	**3.17**	**0.32**
0^{th} and 1^{th} order stat.	**yes**	**yes**	1.59	0.12	4.80	0.28	2.27	0.47	**2.89**	**0.29**

Table 1: Results of multiple combinations of representations and conditioning procedures, evaluated across three NIST SRE conditions. Last two columns display the averages of EER and minimal DCF.

to 4.36% when the pre-conditioning procedure is applied. For both representations, the best performance are achieved by combining the two conditioning procedures. With binary keys, the major proportion of improvement is yielded by the pre-conditioning procedure (12% of relative reduction vs 9% in EER, 14% of relative reduction vs 7% in minDCF). The significant reduction provided by combining the two procedures (30% in EER, 27% in minDCF) is worth noting. With 0^{th} and 1^{th} order statistics, the major proportion of improvement is yielded by the post-conditioning procedure (43% of relative reduction vs 10% in EER, 18% of relative reduction vs 8% in minDCF). The overall relative reduction by using the combined procedures is 48% in EER and 26% in minDCF.

8. Conclusion

The Gaussian-PLDA model for i-vectors is able to achieve the best performance in speaker recognition based on total variability factors. But these results are only obtained if transformations are applied to i-vectors, after their generation by the dimensionality reduction technique and before modeling. Today, the benefits of these *post*-conditioning procedures (comprised of standardization and length-normalization) are well known. The evaluation presented in this paper highlights the importance of *pre*-conditioning procedures, applied to high dimensional representations of utterance before their reduction. With the binary key approach, the pre-conditioning procedure that we have designed for this representation turns out to be the most decisive of the two procedures, in terms of system accuracy. With the Baum-Welch 0^{th} and 1^{th} statistics, this evaluation shows that, if the post-conditioning procedure is the most determinant, combining the two procedures is essential to achieve the state-of-the-art performance.

More broadly, this evaluation shows that the strategy of extending the GMM-UBM based speaker recognition systems through an additional stage of total variability factor is only relevant if normalization procedures are applied before and after the dimensionality reduction stage. These procedures remain essential to adapt the low-rank representation vectors to the probabilistic requirements of the existing detection modules.

9. References

[1] Pierre-Michel Bousquet, Jean-François Bonastre, and Driss Matrouf, "Identify the benefits of the different steps in an i-vector based speaker verification system," in *CIARP*, Springer, Ed., 2013, vol. Part II, pp. 278–285.

[2] Jean-Francois Bonastre, Pierre-Michel Bousquet, Driss Matrouf, and Xavier Anguera, "Discriminant binary data representation for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011, pp. 5284–5287.

[3] Jean-Francois Bonastre, Xavier Anguera, Gabriel H. Sierra, and Pierre-Michel Bousquet, "Speaker modeling using local binary decisions," in *International Conference on Speech Communication and Technology*, 2011, pp. 485–488.

[4] Simon J.D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[5] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[6] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[7] Pierre-Michel Bousquet, Driss Matrouf, and Jean-François Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *International Conference on Speech Communication and Technology*, 2011, pp. 485–488.

[8] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[9] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldřich Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[10] Michael E. Tipping and Christopher M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[11] Christopher M. Bishop, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.