Unsupervised Training of PLDA with Variational Bayes

Jesús Villalba and Eduardo Lleida*

ViVoLab, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida}@unizar.es

Abstract. Speaker recognition relays on models that need a large amount of labeled development data. This models are successful in tasks like NIST SRE where sufficient data is available. However, in real applications, we usually do not have so much data and the speaker labels are unknown. We used a variational Bayes procedure to train PLDA on unlabeled data. The method consisted in a generative model where both the unknown labels and the model parameters are latent variables. We experimented on unlabeled NIST SRE data. The trained models were evaluated on NIST SRE10. Compared to cosine distance, unsupervised PLDA improved EER by 28% and minimum DCF by 36%.

Keywords: speaker recognition, PLDA, unsupervised training, variational Bayes, AHC.

1 Introduction

The i-vector approach provides a method to map a speech utterance to a low dimensional fixed length vector while retaining the speaker identity [1]. We can model the i-vector distributions with advanced techniques like probabilistic linear discriminant analysis (PLDA). PLDA is a generative model that decomposes i-vectors into a speaker specific part and a channel noise. PLDA models need to be trained on labeled databases with large number of speakers and sessions per speaker. Unfortunately, in most applications data is scarce and, in many cases, labels are unknown. We intend to train PLDA in this latter case.

There are previous works that intended to reduce dataset shift to be able to use the same PLDA model in different domains. i-Vector length normalization makes the distributions of different datasets closer. For example, between NIST datasets [2] or between different languages [3]. Bayesian evaluation of likelihood ratios also helps with dataset shift, because the predictive distributions that result, if the amount of training data is small, are heavy-tailed [4, 5].

We presented a variational Bayes (VB) method to adapt a full-rank PLDA model from one domain to another with scarce development data [6], where

* This work has been supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

J.L. Navarro Mesa et al. (Eds.): IberSPEECH 2014, LNAI 8854, pp. 69–78, 2014.

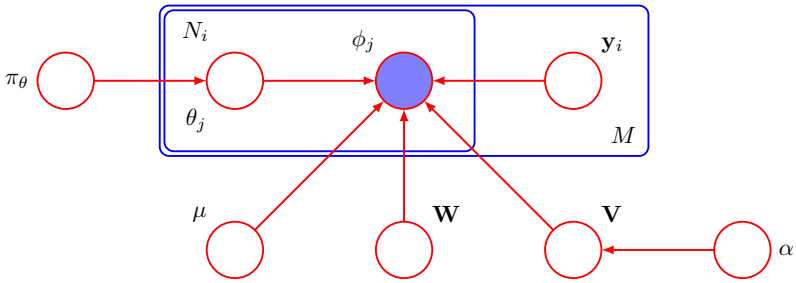


Fig. 1. BN for unsupervised SPLDA

speaker labels were known. Our method was compared with others–parameter or objective function weighting– in the context of the *Domain adaptation challenge* proposed in the 2013 JHU workshop on speaker recognition[1] [7].

The adaptation challenge also promoted adapting models using unlabeled data. We adapted a simplified PLDA model from Switchboard data to NIST SRE [8]. The speaker labels and model parameters were hidden variables whose posterior distributions were iteratively estimated by a VB procedure. In this paper, we intend to evaluate if this procedure is useful to train PLDA from scratch, instead of doing model adaptation. That is, we will not use any labeled data.

Recently, more works about unsupervised adaptation have appeared in relation with the challenge. In [9], agglomerative hierarchical clustering (AHC) is used to obtain the speaker labels of the development set. The clustering is based on the pair-wise scores between i-vectors, computed with an out-of-domain PLDA model. A threshold on the scores, which are unsupervisedly calibrated [10], stops the cluster merging. In [11], several clustering methods were compared (AHC, Markov, infomap). Another approach consists in adding a term accounting for dataset shift to the PLDA model. We can find several flavors of this method [12, 13].

2 Unsupervised SPLDA

2.1 Model Description

Simplified probabilistic linear discriminant analysis (SPLDA) is a linear generative model that assumes that an i-vector ϕ_j of speaker i can be written as:

$$\phi_j = \mu + \mathbf{V}\mathbf{y}_i + \epsilon_j \tag{1}$$

where μ is a speaker independent term, $\mathbf{V}$ is a low rank eigenvoices matrix, $\mathbf{y}_i$ is the speaker factor vector, and ϵ_j is the within class variability term. We put a standard normal prior on $\mathbf{y}_i$ and normal with zero mean and precision $\mathbf{W}$ on ϵ_j.

[1] `http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/`

Figure 1 depicts the Bayesian network of this model where the labels θ of the training data are hidden. θ partitions N i-vectors into M speakers. θ_j is a latent variable comprising a 1–of–M binary vector with elements θ_{ji} with $i = 1, \ldots, M$. Note that the distribution of each speaker is assumed to be Gaussian with mean $\mu + \mathbf{V}\mathbf{y}_i$ and precision $\mathbf{W}$. The set of all the speakers forms a GMM where θ corresponds to the component occupations. The conditional distribution of θ given the mixture weights π_θ is

$$P(\theta|\pi_\theta) = \prod_{j=1}^{N}\prod_{i=1}^{M} \pi_{\theta_i}^{\theta_{ji}} \ . \tag{2}$$

We put a Dirichlet prior on the weights:

$$P(\pi_\theta|\tau_0) = \mathrm{Dir}(\pi_\theta|\tau_0) = C(\tau_0)\prod_{i=1}^{M} \pi_{\theta_i}^{\tau_0 - 1} \tag{3}$$

where, by symmetry, we choose the same τ_0 for all the components, $C(\tau_0)$ is the normalization constant,

$$C(\tau_0) = \frac{\Gamma(M\tau_0)}{\Gamma(\tau_0)^M} \tag{4}$$

and Γ is the Gamma function.

2.2 Model Priors

We chose the model priors based on Bishop's paper about VB PPCA [14]. We introduced a *hierarchical* prior $P(\mathbf{V}|\alpha)$ over $\mathbf{V}$ through a conditional Gaussian distribution of the form:

$$P(\mathbf{V}|\alpha) = \prod_{q=1}^{n_y} \left(\frac{\alpha_q}{2\pi}\right)^{d/2} \exp\left(-\frac{1}{2}\alpha_q \mathbf{v}_q^T \mathbf{v}_q\right) \tag{5}$$

where $\mathbf{v}_q$ are the columns of $\mathbf{V}$ and n_y is the speaker factors dimension. Each α_q controls the inverse variance of the corresponding $\mathbf{v}_q$. If a particular α_q has a posterior distribution concentrated at large values, the corresponding $\mathbf{v}_q$ will tend to be small, and that direction of the latent space will be effectively "switched off".

We defined a prior for α:

$$P(\alpha) = \prod_{q=1}^{n_y} \mathcal{G}(\alpha_q|a_\alpha, b_\alpha) \tag{6}$$

where $\mathcal{G}$ denotes the Gamma distribution.

We placed a Gaussian prior for the mean μ:

$$P(\mu) = \mathcal{N}\left(\mu|\mu_0, \beta^{-1}\mathbf{I}\right) \ . \tag{7}$$

Low values of a_α, b_α and β make the priors less informative and vice versa.

Finally, we put informative Wishart priors on $\mathbf{W}$,

$$P(\mathbf{W}) = \mathcal{W}(\mathbf{W}|\mathbf{\Psi}_0, \nu_0) \ . \tag{8}$$

2.3 Variational Bayes with Deterministic Annealing

We approximated the joint posterior of the latent variables by a factorized distribution of the form:

$$P(\mathbf{Y}, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha | \mathbf{\Phi}) \approx q(\mathbf{Y})\, q(\theta)\, q(\pi_\theta) \prod_{r=1}^{d} q(\tilde{\mathbf{v}}'_r)\, q(\mathbf{W})\, q(\alpha) \tag{9}$$

where $\tilde{\mathbf{v}}'_r$ is a column vector containing the r^{th} row of $\tilde{\mathbf{V}} = [\mathbf{V} \quad \mu]$. If $\mathbf{W}$ were diagonal the factorization $\prod_{r=1}^{d} q(\tilde{\mathbf{v}}'_r)$ would not be necessary because it would arise naturally. However, for full $\mathbf{W}$, we have to force the factorization to make the problem tractable.

We computed these factors by using Variational Bayes [15] with deterministic annealing (DA) [16]. The formula to update a factor q_j is

$$\ln q_j^*(\mathbf{Z}_j) = \mathrm{E}_{i \neq j} [\kappa \ln P(\mathbf{\Phi}, \mathbf{Z})] + \text{const} \tag{10}$$

where $\mathbf{Z}$ abbreviates the set of all hidden variables, $\mathbf{Z}_j$ are the hidden variables corresponding to the j^{th} factor, and κ is the annealing factor; expectations are taken with respect to all the factors $i \neq j$. We could prove that Equation (10) optimizes the VB lower bound

$$\mathcal{L} = \mathrm{E}[\ln P(\mathbf{\Phi}, \mathbf{Z})] - \mathrm{E}[\ln q(\mathbf{Z})] = \ln P(\mathbf{\Phi}) - \mathrm{KL}(q(\mathbf{Z}) || P(\mathbf{Z}|\mathbf{\Phi})) \tag{11}$$

where expectations are taken with respect to the variational posterior $q(\mathbf{Z})$. $\mathcal{L}$ is an approximation of the marginal likelihood of the data $\ln P(\mathbf{\Phi})$, which becomes equality when approximated posterior is equal to the true posterior. Annealing modifies the VB objective in a way that helps to avoid local maxima. We must set $\kappa < 1$ at the beginning and increase it in each iteration until $\kappa = 1$. The full VB equations can be found in our report [17].

2.4 Initialization with AHC

The speaker labels were initialized with Agglomerative hierarchichal clustering (AHC) [18]. AHC is a greedy bottom-up approach. Initially, each i-vector is its own cluster and, then, clusters are progressively merged using a similarity criterion–We used cosine distance. Thus, we started with the pair-wise score matrix between all the development i-vectors. Then, to merge two clusters, A and B, we tried tree linkage criteria: average, complete and single. The linkage criterion determines the similarity between the clusters A and B, $s(A, B)$, as a function of the pair-wise scores between their elements $s(a, b)$. Thus,

$$s_{\text{avg}}(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} s(a, b) \tag{12}$$

$$s_{\text{complete}}(A, B) = \min \{ s(a, b) | a \in A, b \in B \} \tag{13}$$

$$s_{\text{single}}(A, B) = \max \{ s(a, b) | a \in A, b \in B \} \ . \tag{14}$$

2.5 Model Selection

To select the best model, i.e., the best number of speakers M; we used the same method that in our previous work [8]. We ran the AHC+VB algorithm several times, each time hypothesizing a different M. We assumed that the best model is the one that obtains the largest VB lower bound $\mathcal{L}(M)$. To fairly compare lower bounds for different M, the Dirichlet prior on the speaker weights needs to be such that the product $M\tau_0$ is constant. To select the value of that constant, we tried several values and chose the one that maximized the sum $\sum_M \mathcal{L}(M)$.

3 Experiments

3.1 Experimental Setup

We trained PLDA on an unlabeled version of NIST SRE04-08. The i-vectors for this task were provided by the JHU HLT-COE in the 2013 JHU workshop on speaker recognition [11]. The training data consisted of 33125 segments from 3789 speakers. To perform faster experiments, we also created a subset of 500 speakers. The adapted models were evaluated on the NIST SRE10 det5 (tel-tel) extended condition.

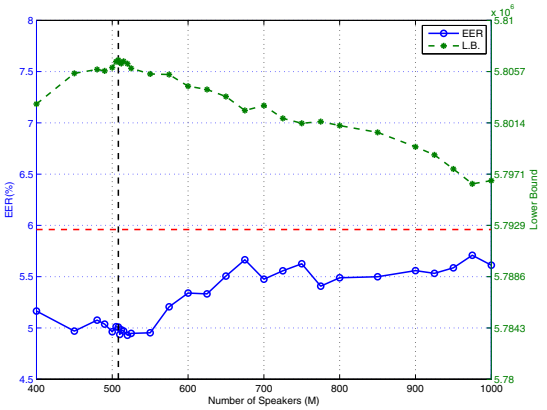
The i-vectors were 600 dimensional. They were extracted using 20 MFCC + Δ with short time mean and variance normalization. The UBM and i-vector extractor were gender independent and used 2048 Gaussians. We applied centering, whitening and length normalization to the i-vectors [2]. The parameters needed for centering and whitening were trained on all NIST SRE data since speaker labels are not required.

The SPLDA models were gender independent with speaker factors of dimension $n_y = 400$ when training with 500 speakers; and $n_y = 600$ when training with all the speakers. Given the results in our previous work [8], we put informative priors on the model parameters. Our priors were based on the average total variance of the data s_0^2–average across dimensions. From our previous work, we assumed that the average variance of the speaker space was approximately 15% of s_0^2 and the channel variance was the remaining 85%. Thus, we computed s_0^2 from the training data. Then, for α (prior of the inverse eigenvalues), we placed a wide prior with mode $1/(0.15s_0^2)$ by setting $a_\alpha = 2$ and $b_\alpha = 0.15s_0^2$. For $\mathbf{W}$, we used a Wishart prior with expectation $1/(0.85s_0^2)\mathbf{I}$ by setting $\nu_0 = 602$ and $\Psi_0 = 1/(0.85s_0^2\nu_0)\mathbf{I}$. Note that, for the Wishart prior to be proper, we need $\nu_0 > d$, this means that the prior will have an important influence on the posterior unless we have a number of training segments $N >> d$. We set $\tau_0 = 400/M$.

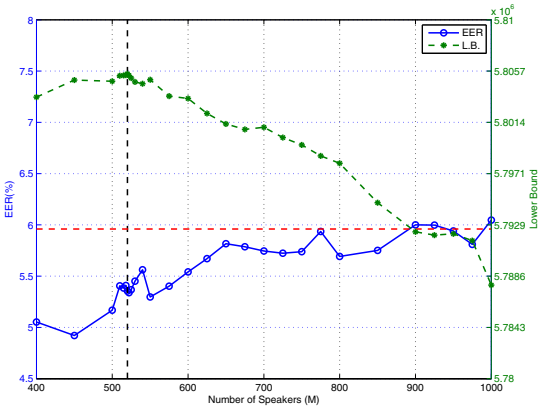
The expectations of the model parameters given the VB posteriors were used to compute the likelihood ratios of the evaluation set in the standard way.

3.2 Experiments Results

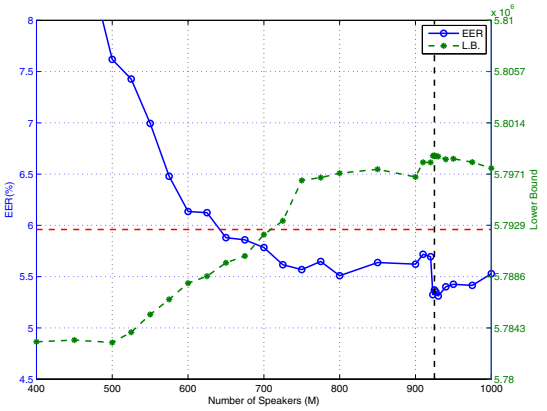
First, we focus on the results obtained by training the PLDA with 500 speakers. Figure 2 plots the EER and VB lower bound against the number of hypothesized speakers M. Each subfigure corresponds to one of the linkage criteria used in



(a) Average linkage clustering.

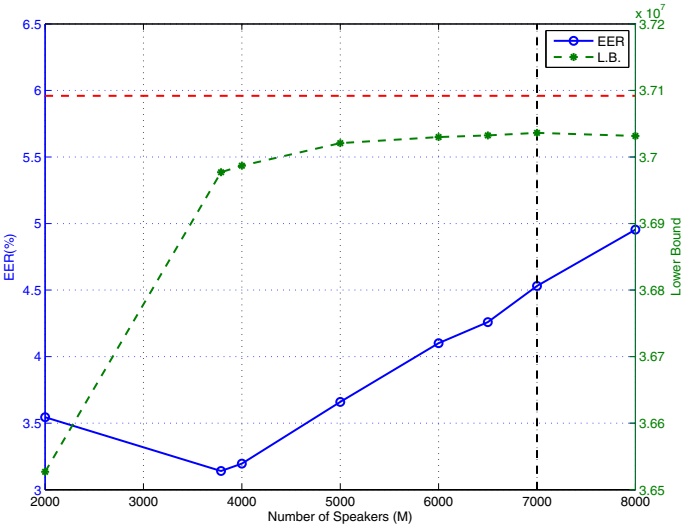


(b) Complete linkage clustering.

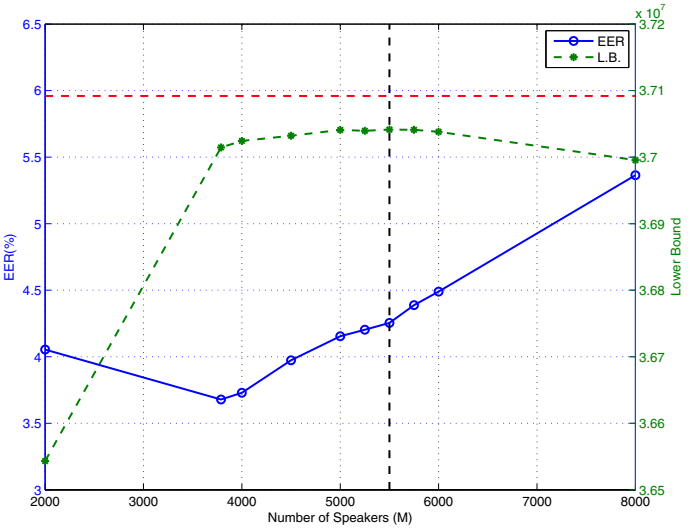


(c) Single linkage clustering.

Fig. 2. EER(%)/$\mathcal{L}$ against the number of hypothesized speakers for different initialization methods and using a subset of 500 development speakers



(a) Average linkage clustering.



(b) Complete linkage clustering.

Fig. 3. EER(%)/$\mathcal{L}$ against the number of hypothesized speakers for different initialization methods and using all the development speakers

the AHC initialization. The left y-axes show the scale of the EER, and the right y-axes show the scale of the lower bound. The horizontal dashed line indicates the baseline–cosine similarity– and the vertical dashed line indicates the point where $\mathcal{L}$ is maximum. Regarding the detection of the number of speakers in the development set, average and complete linkage criteria had their $\mathcal{L}$ maxima

Table 1. EER(%)/MinDCF for different initialization of the VB. The table blocks correspond to training 500 or all the speakers.

	M Actual				M Max $\mathcal{L}$			
	EER(%)	MinDCF	M	$\mathcal{L} \times 10^{-6}$	EER(%)	MinDCF	M	$\mathcal{L} \times 10^{-6}$
Baseline (Cosine)	5.96	0.66	-	-	5.96	0.66	-	-
Oracle labels	**3.02**	**0.50**	**500**	-	**3.02**	**0.50**	**500**	-
VB average link	**4.96**	**0.58**	500	**5.8060**	**5.00**	**0.58**	**508**	**5.8066**
VB complete link	5.16	0.60	500	5.8048	5.35	0.61	520	5.8054
VB single link	7.61	0.77	500	5.7830	5.37	0.61	925	5.7987
Oracle labels	**2.19**	**0.42**	**3789**	-	**2.19**	**0.42**	**3789**	-
VB average link	**3.14**	**0.44**	3789	36.977	4.53	**0.47**	7000	37.036
VB complete link	3.67	0.46	3789	**37.014**	**4.25**	**0.47**	**5500**	**37.041**

close to the actual value. In contrast, single linkage almost doubled the speakers. Nevertheless, the maximum $\mathcal{L}$ was a good criterion to select a model with low EER for the three cases. For average and complete linkage, it selected the point of minimum EER. For complete linkage it did not choose the minimum EER but a point quite near of it. In the three cases, it significantly improved the baseline.

Table 1 compares EER and minimum DCF for multiple cases. The table also shows the number of speakers and the $\mathcal{L}$ obtained in each case. The table has two column blocks. The left block shows results for the model corresponding to the actual M; and the right block to the model that maximizes $\mathcal{L}$. The upper block of rows correspond to the development set of 500 speakers. Average linkage obtained the lowest EER and DCF for both model selection methods (M actual and max. $\mathcal{L}$). Also for both methods average linkage obtained the highest $\mathcal{L}$, so we can use $\mathcal{L}$ to choose the best initialization. The Max $\mathcal{L}$ model improved the baseline by 16% and 12% in terms of EER and DCF. With respect to training with oracle labels, we still have a margin of improvement of 39% and 14% respectively. Single linkage was the worst initialization method so we discarded it for the following experiments.

Figure 3 plots EER and $\mathcal{L}$ against M when training with all the development speakers (3789). In this case, $\mathcal{L}$ was maximum for M much higher than its actual value–almost twice for average linkage. The reason is that, when we increase the number of speakers, we increase the probability of finding speakers with overlapping i-vector distributions and clustering becomes harder. Despite of that, the selected models outperformed the baseline. Average linkage obtained the best EER and DCF for the models with oracle M but complete linkage was better for the model maximizing $\mathcal{L}$. $\mathcal{L}$ was also higher for complete linkage. With respect to the baseline, EER improved by 28.7% and DCF by 36.4%. With respect to oracle model selection, we have a margin for improvement of 26% and 6% respectively; and with respect to oracle labels, margins of 48.5% and 10%. We can see that those margins are still very high. As we increase the amount of data, the margin between the unsupervised and supervised models also increases. As clustering

becomes harder, there is a point where increasing the amount of unsupervised data is not beneficial.

4 Conclusions

We presented a method to train SPLDA models with unsupervised labels. We designed a generative model where labels and model parameters are hidden variables that are updated iteratively with a variational Bayes procedure. The speaker labels were initialized using AHC with different linkage criteria. The best criteria were average and complete linkage. The VB procedure was run several times, each time hypothesizing a different number of development speakers. We selected the model that maximized the VB lower bound.

We experimented training on unlabeled NIST SRE04-08 data. We evaluated the resulting model on the NIST SRE10 det5 condition. For training with 500 speakers, the algorithm was able to select almost the best model. Compared to cosine distance, EER improved by 16% and minimum DCF by 16%. For training with 3789 speakers, clustering becomes harder and we did not selected the best model. However, EER improved by 28% and DCF by 36%. Despite that these gains were significant, there is still a large margin of improvement to match the results of supervised training.

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(4), 788–798 (2011)
2. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of I-vector Length Normalization in Speaker Recognition Systems. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011, Florence, Italy, pp. 249–252. ISCA (August 2011)
3. Vaquero, C.: Dataset Shift in PLDA based Speaker Verification. In: Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop, Singapore, pp. 39–46. COLIPS (June 2012)
4. Villalba, J., Brummer, N.: Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011, Florence, Italy, pp. 505–508. ISCA (August 2011)
5. Villalba, J., Brummer, N., Lleida, E.: Fully Bayesian Likelihood Ratios vs i-vector Length Normalization in Speaker Recognition Systems. In: NIST SRE 2011 Speaker Recognition Workshop, Atlanta, Georgia, USA (December 2011)
6. Villalba, J., Lleida, E.: Bayesian Adaptation of PLDA Based Speaker Recognition to Domains with Scarce Development Data. In: Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop, Singapore. COLIPS (June 2012)
7. Garcia-Romero, D., McCree, A.: Supervised Domain Adaptation for I-Vector Based Speaker Recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, pp. 4075–4079. IEEE (May 2014)

8. Villalba, J., Lleida, E.: Unsupervised Adaptation of PLDA by Using Variational Bayes Methods. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy. IEEE (May 2014)
9. Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., Vaquero, C.: Unsupervised Domain Adaptation for I-Vector Speaker Recognition. In: Proceedings of Odyssey 2014 - The Speaker and Language Recognition Workshop, Joensuu, Finland, pp. 260–264. ISCA (June 2014)
10. Brummer, N., Garcia-Romero, D.: Generative Modelling for Unsupervised Score Calibration. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, pp. 1699–1703. IEEE (May 2014)
11. Shum, S., Reynolds, D.A., Garcia-Romero, D., McCree, A.: Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems. In: Proceedings of Odyssey 2014 - The Speaker and Language Recognition Workshop, Joensuu, Finland, pp. 265–272. ISCA (June 2014)
12. Aronowitz, H.: Inter Dataset Variability Compensation for Speaker Recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, pp. 4030–4034. IEEE (May 2014)
13. Glembek, O., Ma, J., Matejka, P., Zhang, B., Plchot, O., Burget, L., Matsoukas, S.: Domain Adaptation Via Within-Class Covariance Correction in I-Vector Based Speaker Recognition Systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, pp. 4060–4064. IEEE (May 2014)
14. Bishop, C.: Variational principal components. In: Proceedings of the 9th International Conference on Artificial Neural Networks, ICANN 1999, Edinburgh, Scotland, pp. 509–514. IET (September 1999)
15. Bishop, C.: Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC (2006)
16. Katahira, K., Watanabe, K., Okada, M.: Deterministic annealing variant of variational Bayes method. Journal of Physics: Conference Series International Workshop on Statistical-Mechanical Informatics (IW-SMI 2007), 95 (January 2008)
17. Villalba, J.: Unsupervised Adaptation of SPLDA. Technical report, University of Zaragoza, Zaragoza, Spain (2013)
18. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press (1973)