INTRA-CLASS COVARIANCE ADAPTATION IN PLDA BACK-ENDS FOR SPEAKER VERIFICATION

Srikanth Madikeri Marc Ferras Petr Motlicek
Subhadeep Dey

Idiap-RR-05-2017

FEBRUARY 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

INTRA-CLASS COVARIANCE ADAPTATION IN PLDA BACK-ENDS FOR SPEAKER VERIFICATION

Srikanth Madikeri[1], *Marc Ferras*[1], *Petr Motlicek*[1] *and Subhadeep Dey*[1,2]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École polytechnique fédérale de Lausanne, Lausanne, Switzerland
{msrikanth, mferras, pmotlic, sdey}@idiap.ch

ABSTRACT

Multi-session training conditions are becoming increasingly common in recent benchmark datasets for both text-independent and text-dependent speaker verification. In the state-of-the-art i-vector framework for speaker verification, such conditions are addressed by simple techniques such as averaging the individual i-vectors, averaging scores, or modifying the Probabilistic Linear Discriminant Analysis (PLDA) scoring hypothesis for multi-session enrollment. The aforementioned techniques fail to exploit the speaker variabilities observed in the enrollment data for target speakers. In this paper, we propose to exploit the multi-session training data by estimating a speaker-dependent covariance matrix and updating the intra-speaker covariance during PLDA scoring for each target speaker. The proposed method is further extended by combining covariance adaptation and score averaging. In this method, the individual examples of the target speaker are compared against the test data as opposed to an averaged i-vector, and the scores obtained are then averaged. The proposed methods are evaluated on the NIST SRE 2012 dataset. Relative improvements of up to 29% in equal error rate are obtained.

Index Terms— i-vectors, PLDA, multi-session training

1. INTRODUCTION

Speaker verification (SV) is typically addressed as a hypothesis testing problem in which we compare the same-speaker hypothesis against the different-speaker hypothesis [1]. The same-speaker hypothesis states that the target speaker is present in the test recording, while the different-speaker hypothesis rejects the claim. In conventional SV systems, the target speaker and the test recording are represented by i-vectors, which are fixed low-dimensional representation of speech in the audio [2]. I-vector based SV systems often employ a Probabilistic Linear Discriminant Analysis (PLDA) classifier to obtain a log-likelihood ratio based score to evaluate these hypotheses [3, 4].

Recent SV benchmark datasets, such as the NIST SRE 2012 ([5]), RSR2015 ([6]) and Reddots ([7]), have multiple audio files to enroll target speakers. In order to exploit multiple examples, various methods have been explored [8, 9]. As the PLDA scoring framework is a well established back-end classifier, a majority of these methods target to exploit it in order to leverage its effectiveness. Generally, such methods extract an i-vector for each available session. These models are either compared to the test recording either individually or after averaging them to obtain one single model for the target speaker [8, 10]. In the former case, the scores need to be combined (for instance, through simple averaging) [11]. Another method to obtain a single i-vector for the target speaker is to pool the sufficient statistics from all training samples [12]. In [13], the PLDA scoring method was extended for speaker with multiple enrollment utterances, but no improvement was observed over the methods mentioned earlier. From our experience with the RSR2015 dataset in [10], averaging the scores performed better than averaging the i-vectors. Thus, there is a lack of clarity in the strategy to be selected in such cases of multi-session enrollment.

With multiple examples for enrollment, all the above mentioned methods fail to explicitly model the observed speaker-specific intra-class variability. In this paper, we show that by carefully modeling the assumptions on the intra-speaker covariance matrix it is possible to further exploit multiple enrollment samples for speakers in the existing PLDA scoring implementations. This is further extended by adapting intra-speaker covariance matrices for each speaker in the evaluation set. The advantages of the proposed approaches primarily lie in the simplicity of the implementation given the performance improvements that can be achieved for speakers with multi-session data. Very little adjustment to the original PLDA approach is required to obtain significant gains on the SV performance. We show that we can improve relatively by 29% in terms of Equal Error Rate (EER) with the proposed approach compared with simple i-vector averaging.

The rest of the paper is organized as follows. Section 2 describes the baseline i-vector PLDA system. Section 3 presents our proposed modifications. In Section 4, the experimental setup and results are given. The conclusions from our results are given in the final section.

2. I-VECTOR PLDA FRAMEWORK

The i-vector extractor projects Gaussian mean supervectors on a low-dimensional subspace called *total variability space* (TVS) [2]. The variability model underlying i-vector extraction is given by

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

where $\mathbf{s}$ is the supervector adapted with respect to a Universal Background Model-Gaussian Mixture Model (UBM-GMM) from a speech recording. The vector $\mathbf{m}$ is the mean of the supervectors,$\mathbf{T}$ is the matrix with its columns spanning the total variability subspace and $\mathbf{w}$ is the low-dimensional i-vector representation. In the above model, the i-vector is assumed to have a standard Normal distribution as prior.

The i-vectors obtained from a speech utterance are further projected onto a discriminative space using techniques such as LDA, WCCN [14, 2] and PLDA [3, 15], which together form the back-end of the i-vector system. Length normalization is also often applied prior to PLDA modeling ([4]) in order to aggressively deal with the non-Gaussian behavior of the i-vectors [16]. Using PLDA parameters two i-vectors can be compared as belonging to the same class or as belonging to two different classes, thus generating a likelihood ratio (LR) to score a pair of speech utterances.

2.1. PLDA modelling

While there are several variants of the PLDA algorithm [17], a generic implementation models the interclass and intraclass variances as follows

$$\mathbf{w}_{\mathrm{ij}} = \boldsymbol{\mu} + \mathbf{F}\mathbf{v}_{\mathrm{i}} + \mathbf{E}\mathbf{u}_{\mathrm{ij}} + \boldsymbol{\epsilon}_{\mathrm{ij}}, \tag{2}$$

where $\mathbf{w}_{\mathrm{ij}}$ refers to the i-vector of the j^{th} example of the i^{th} speaker. The matrices $\mathbf{F}$ and $\mathbf{E}$ model the interclass and intraclass variabilities, respectively. The hidden variable $\mathbf{v}_{\mathrm{i}}$ is the class identity and $\mathbf{u}_{\mathrm{ij}}$ explains the deviation from the class means. The residue $\boldsymbol{\epsilon}_{\mathrm{ij}}$ follows a standard normal distribution.

The Gaussian PLDA (GPLDA) approach simplifies the modelling of interclass and intraclass covariances by subsuming the latter into the residual covariance matrix [4]. The above model is simplified as follows

$$\mathbf{w}_{\mathrm{ij}} = \boldsymbol{\mu} + \mathbf{F}\mathbf{v}_{\mathrm{i}} + \boldsymbol{\epsilon}_{\mathrm{ij}}, \tag{3}$$

where the residue is now modelled by a normal distribution with mean $\mathbf{0}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$.

An equivalent implementation of PLDA is available in the Kaldi toolkit [18]. The implementation is based on [15], in which PLDA is introduced as a natural probabilistic interpretation of LDA. The probabilistic interpretation arises from the assumptions imposed on the distribution of the hidden variables. These assumptions lead to estimates of the covariance matrices, which are then used to compute the projection as in the case of the conventional LDA algorithm. The interclass ($\boldsymbol{\Phi}_b$) and intraclass covariances ($\boldsymbol{\Phi}_w$) are modelled according to a shared orthonormal projection matrix $\mathbf{A}$ so that

$$\mathbf{w}_{\mathrm{ij}} = \boldsymbol{\mu} + \mathbf{A}\mathbf{u}_{\mathrm{ij}}, \tag{4}$$

where $\mathbf{u}_{\mathrm{ij}}$ follows a normal distribution $\mathcal{N}(\mathbf{v}_{\mathrm{i}}, \mathbf{I})$ with $\mathbf{v}_{\mathrm{i}}$ being the class identity. The hidden variable $\mathbf{v}$ follows a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the interclass covariance diagonalized by $\mathbf{A}$ such that $\boldsymbol{\Phi}_b = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}^T$.

In our experience, the GPLDA algorithm and Kaldi's implementation perform similarly. Thus, we chose to explore the latter following our i-vector system implemented for the same toolkit [19]. The methods presented in this paper can be easily extended to the GPLDA framework as well.

2.2. PLDA scoring

In this section, we describe the PLDA scoring strategy for speakers with only one training session. The scoring involves computing the ratio between the likelihood of the test audio and the target speaker sharing the same model, and the probability that the two i-vectors being compared are independent. If $\mathbf{w}_{\mathrm{tst}}$ is the test i-vector and $\mathbf{w}_{\mathrm{trn}}$ is the target speaker's i-vector, the former term corresponds to the probability $P(\mathbf{w}_{\mathrm{tst}}|\mathbf{w}_{\mathrm{trn}})$, which is obtained by marginalizing the joint distribution of the enrollment and test i-vectors over a common identity hidden variable. According to Eq. 4, this is equivalent to $P(\mathbf{u}_{\mathrm{tst}}|\mathbf{u}_{\mathrm{trn}})$, where $\mathbf{u}_{\mathrm{tst}}$ and $\mathbf{u}_{\mathrm{trn}}$ correspond respectively to $\mathbf{w}_{\mathrm{tst}}$ and $\mathbf{w}_{\mathrm{trn}}$. The simplified LR after ignoring common factors is given by

$$\mathrm{LR} = \mathcal{N}(\mathbf{u}_{\mathrm{tst}}; \mathbf{u}_{\mathrm{trn}}, \mathbf{I} + \mathbf{S}) / \mathcal{N}(\mathbf{u}_{\mathrm{tst}}; \mathbf{0}, \mathbf{B}), \tag{5}$$

where $\mathbf{S} = \frac{\boldsymbol{\Psi}}{\boldsymbol{\Psi}+\mathbf{I}}$ and $\mathbf{B} = \boldsymbol{\Psi}+\mathbf{I}$. We will continue referring to only the projected i-vector $\mathbf{u}$ and ignore the original i-vector $\mathbf{w}$ in the rest of the text.

2.3. Multiple i-vector enrollment

In case of multiple samples for enrollment, the numerator in Eq. 5 is naturally extended from the definition of joint probabilities of i-vectors. The means and covariance matrix are modified as follows

$$\mathrm{LR} = \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{S}_n\mathbf{u}_{\mathrm{trn}}, \frac{1}{n}\mathbf{S}_n + \mathbf{I}\right) / \mathcal{N}(\mathbf{u}_{\mathrm{tst}}; \mathbf{0}, \mathbf{B}), \tag{6}$$

where $\mathbf{S}_n = \frac{n\boldsymbol{\Psi}}{n\boldsymbol{\Psi}+\mathbf{I}}$ and n is the number of training samples for the target class, and $\mathbf{u}_{\mathrm{trn}}$ is the average over training samples. This type of scoring is termed multisession scoring in this paper. While the above equation is a natural extension of the scoring strategy by definition (Eq 4), we observe the bias in the covariance of the joint distribution. We also observe that in case of multiple i-vectors the estimate of this covariance can be adapted to the observed intra-speaker covariance, thereby obtaining a speaker-dependent intra-speaker covariance matrix.

2.4. Baseline methods

In this subsection, we discuss two baseline strategies to utilize multiple enrollment samples for a target speaker: i-vector averaging and score averaging. In the former technique, the i-vectors of a speaker are averaged. In Eq. 6, this corresponds to ignoring the number of examples.

In score averaging, the multiple examples are combined during scoring. The scores are obtained for each training session from Eq. 5 as though they belong to different speakers. The individual scores are then averaged to obtain one single score for the target speaker.

3. PROPOSED PLDA SCORING

Our proposed methods focus on adapting the intra-speaker covariance based on the observed i-vectors for the target speaker. In order to do this, we first observe that the intra-speaker covariance for a speaker with n examples is biased. Thus, at first we simply adjust the intra-speaker covariance as $\mathbf{S}_n + \mathbf{I}$. The LR is re-defined as

$$\mathrm{LR} = \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{S}_n \mathbf{u}_{\mathrm{trn}}, \mathbf{S}_n + \mathbf{I}\right) / \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{0}, \mathbf{B}\right). \tag{7}$$

This method is referred to as covariance scaling.

With the above scoring strategy, we propose to adapt the intra-class covariance based on the i-vectors observed. Let for a target speaker, the observered i-vector projections be $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$. The observed intraclass covariance is

$$\hat{\mathbf{\Psi}} = \frac{1}{n} \sum_{k=1}^{n} \left(\mathbf{u}_k - \mathbf{S}_n \mathbf{u}_{\mathrm{trn}}\right)^2. \tag{8}$$

Therefore, Eq 7 is rewritten for covariance adaptation as

$$\mathrm{LR} = \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{S}_n \mathbf{u}_{\mathrm{trn}}, \mathbf{S}_n + \mathbf{I} + \hat{\mathbf{\Psi}}\right) / \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{0}, \mathbf{B}\right). \tag{9}$$

In all the strategies discussed above, the mean of the target speaker's distribution is unchanged. However, in many cases such averaging can be aggressive. The mean could lead to many examples modelled poorly by the intra-speaker variance, especially in the baseline system. Such cases occur when there are training samples available from varied conditions. Thus, the effect of using multiple examples can be sub-optimal. Therefore, we propose to combine the covariance adaptation strategy and score averaging. In this technique, we compare individually the test audio with each of the enrollment i-vectors. The LR is re-defined as

$$\mathrm{LR} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{u}_i, \mathbf{S}_n + \mathbf{I} + \hat{\mathbf{\Psi}}\right) / \mathcal{N}\left(\mathbf{u}_{\mathrm{tst}}; \mathbf{0}, \mathbf{B}\right). \tag{10}$$

The equal weight for each example in Eqs. 9, 10 could lead to poor samples to affect the result. For instance, when the input samples vary considerably in length, equal weight to both short and long utterances could impact the system performance when comparing with long test recordings. To handle such scenarios, we propose an automatic weighting scheme based on the target speaker distribution.

The weight γ_i for sample i of the target speaker is given by

$$\gamma_i = \frac{\mathcal{N}\left(\mathbf{u}_{\mathrm{i}}; \mathbf{S}_n \mathbf{u}_{\mathrm{trn}}, \mathbf{S}_n + \mathbf{I}\right)}{\sum_{j=1}^{n} \mathcal{N}\left(\mathbf{u}_{\mathrm{j}}; \mathbf{S}_n \mathbf{u}_{\mathrm{trn}}, \mathbf{S}_n + \mathbf{I}\right)}. \tag{11}$$

The weights replace the uniform weights in Eq. 10 and filter outliers by controlling their impact on the final score. These weights also reflect the influence of individual samples on all the scoring strategies, including baseline PLDA scoring, mentioned earlier. By applying weights during covariance estimation, we aimed to control the influence of each training example for the speaker based on the intra-speaker variance observed in the development dataset.

4. EXPERIMENTS

Speaker verification experiments are conducted on the female subset of the NIST 2012 SRE data, following the official protocol [20]. System performance is evaluated on conditions 2 and 5 (similar to [21]), labelled as cond2 and cond5 in this article, and results are given as EERs and minDCF (minimum Decision Cost Function). While cond5 involves test recordings with added noise, cond2 does not.

4.1. I-vector system configuration

The front-end used 20 MFCC features along with delta and acceleration parameters, extracted every 10 ms using a window of 30 ms. They were further processed through a short term Gaussianization module with a context of 300 frames [22].

The Fisher English Part I and II data were used to train the UBM-GMM system and the T-matrix. LDA and PLDA parameters were trained with the following datasets: The NIST datasets - SRE 2004, 2005, 2006, 2008 and 2008 extended, Switchboard Part II and Part III, and Switchboard Cellular Part I and II.

A UBM-GMM with 2048 components and i-vector extractor of 400 dimensions were trained. The i-vector dimension was reduced to 150 after LDA, followed by length normalization prior to PLDA scoring.

The Kaldi toolkit [18] was used for LDA and PLDA training. A standard i-vector extractor was implemented for Kaldi as well (see footnote 1 in Page 2), based on the baseline system described in [23].

4.2. Results

The results are given in Table 1. The EERs and minDCFs of all the systems presented earlier are compared. The minDCF is computed with a prior probablity of 0.01 for the target speaker. As a part of our baseline results, we consider three techniques: i-vector averaging, score averaging and multisession scoring (Eq. 6). In both conditions, i-vector averging performs the best among all three baslines with EERs of 2.4% and 3.0% on cond2 and cond5, respectively. These results are better than score averaging by about 14% and 6% relative in EER. The minDCFs of the i-vector averaging system also confirm its superiority over the other two baseline techniques. The multisession scoring strategy perfoms the worst among the three baselines.

Next, the performances of the proposed systems are presented. First, we test the validity of covariance scaling (Eq. 7). There is a considerable improvement from the baseline multisession scoring technique to justify the removal of a bias from the intraspeaker variance estimate. The EERs improves by about 33% and 31% relative on cond2 and cond5, respectively. With respect to i-vector averaging, the improvements in EER are approximately 20% and 22%, respectively, and the improvements in minDCF are about 12% and 28%. The second system applies variance adaptation (Eq. 9). There were marginal gains in EERs and DCF from the system using variance scaling.

The covariance adaptation with weights (γ_i) in Eq. 11 is termed Weighted Variance adaptation in Table 1. While there is a decrease observed in the performance, we note that the number of examples used per speaker is significantly lesser than the actual number of training samples available. For instance, when the number examples per speaker exceeds 5, an average of 20% of the samples have γ values over 0.01. One major reason is attributed to the peaky nature of distribution assumed in Eq. 5. The results confirm that these models still do not fully exploit the multisession enrollment condition.

Our final results combine score averaging with variance scaling and variance adaptation as defined in Eq. 10. The former is presented to justify the proposed adaptation procedure. When averaging scores obtained from individual samples with covariance scaling (labelled as Variance scaling + Score averaging), there is once again degradation between 4% to 5% relative in EER on the two conditions when compared with the variance scaling procedure. However, significant improvements are achieved when the individual scores after applying variance adaptation (labelled as Variance adaptation + Score averaging) are averaged. When compared to the baseline system using i-vector averaging, relative improvements in EER of 28% and 29% for cond2 and cond5, respectively, are observed. The corresponding relative improvements in DCFs are the same as that obtained with simply scaling the variance. Overall, systems using simple variance adaptation, that is, when the samples have uniform weights, outperform all other variants.

Table 1. Results on the NIST 2012 dataset for the all scoring methods presented in this work. The Equal Error Rate (in %)/minimum Decision Cost Function are reported.

Method	cond2	cond5
Baseline		
I-vector averaging	2.4/0.24	3.0/0.28
Score averaging	2.8/0.38	3.2/0.35
Multisession scoring	3.0/0.39	3.5/0.46
Proposed		
Covariance scaling	2.0/0.21	2.4/0.24
Covariance adaptation	1.9/**0.20**	2.4/**0.23**
Weighted Covariance adaptation	2.1/0.22	2.5/0.26
Covariance scaling + Score averaging	2.1/0.25	2.5/0.29
Covariance adaptation + Score averaging	**1.7**/0.21	**2.1**/0.24

5. CONCLUSION

In this paper, adaptation of the intra-speaker covariance matrix for PLDA scoring was proposed in order to take advantage of the multi-session enrollment conditions. It is observed that simple scaling of the covariance matrix can lead to performance benefits. Relative improvements in EER of about 20% and 22% are obtained on two conditions of the NIST SRE 2012 dataset. The intra-speaker covariance is also adapted with the enrollment data to obtain speaker-dependent covariance. Although such variance adaptation does not directly lead to performance gains, combining it with score averaging resulted in relative improvements of up to 29% in EER when compared to the baseline system using simple i-vector averaging. We also explored probabilistic weighting of individual examples of a target speaker for PLDA scoring. While it did not contribute to any performance improvements, we observed that very few examples contribute to the speaker model when i-vectors are averaged. This justifies the performance of score averaging after variance adaptation.

Acknowledgements

This work was supported by the EU FP7 project ”SPEAKER IDENTIFICATION INTEGRATED PROJECT (SIIP)”. This work was also partially supported by the Swiss CTI grant funding application / CTI no.25356.1 PFES-ES - ”Seamless Customer Identification”.

6. REFERENCES

[1] Tomi Kinnunen and B Haizhou Li, “An overview of text-independent speaker recognition: from features to supervectors,” January 2010, vol. 52(1), pp. 12–40, Speech Communication.

[2] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Tran. on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[3] Simon JD Prince and James H Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE 11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[4] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” August 2011, pp. 249–252, In Proc. of Interspeech.

[5] “The NIST Year 2012 Speaker Recognition Evaluation Plan, https://www.nist.gov/document-6865,” .

[6] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, “Rsr2015: Database for text-dependent speaker verification using multiple pass-phrases.,” in *INTERSPEECH*, 2012, pp. 1580–1583.

[7] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., “The reddots data collection for speaker recognition,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Gang Liu and John HL Hansen, “An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1978–1992, 2014.

[9] Gang Liu, Taufiq Hasan, Hynek Boril, and John HL Hansen, “An investigation on back-end for speaker recognition in multi-session enrollment,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7755–7759.

[10] Subhadeep Dey, Srikanth Madikeri, Marc Ferras, and Petr Motlicek, “Deep neural network based posteriors for text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5050–5054.

[11] Cemal Hanilçi, Tomi Kinnunen, Md Sahidullah, and Aleksandr Sizov, “Classifiers for synthetic speech detection: A comparison,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson, “Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition,” in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4257–4260.

[13] Padmanabhan Rajan, Anton Afanasyev, Ville Hautamäki, and Tomi Kinnunen, “From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification,” *Digital Signal Processing*, vol. 31, pp. 93–101, 2014.

[14] Richard O Duda, Peter E Hart, and David G Stork, *Pattern classification*, John Wiley & Sons, 2012.

[15] Sergey Ioffe, “Probabilistic linear discriminant analysis,” in *Computer Vision–ECCV 2006*, pp. 531–542. Springer, 2006.

[16] Patrick Kenny, “Bayesian speaker verification with heavy-tailed priors.,” in *Odyssey*, 2010, p. 14.

[17] Aleksandr Sizov, Kong Aik Lee, and Tomi Kinnunen, “Unifying probabilistic linear discriminant analysis variants in biometric authentication,” in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2014, pp. 464–475.

[18] D. Povey, A. Ghoshal, et al., “The kaldi speech recognition toolkit,” in *In Proc. of ASRU 2011*, December 2011.

[19] Srikanth Madikeri, Subhadeep Dey, Petr Motlicek, and Marc Ferras, “Implementation of the standard i-vector system for the kaldi speech recognition toolkit,” Tech. Rep., No. EPFL-REPORT-223041 Idiap, 2016.

[20] “The NIST Year 2010 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html,” .

[21] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[22] Jason Pelecanos and Sridha Sridharan, “Feature warping for robust speaker verification,” 2001.

[23] O Glembek et al., “Simplification and optimization of i-vector extraction,” 2011, pp. 4516–4519, In Proc. of ICASSP.