INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Nonparametrically Trained Probabilistic Linear Discriminant Analysis for *i*-Vector Speaker Verification

Abbas Khosravani, Mohammad Mehdi Homayounpour

Laboratory for Intelligent Multimedia Processing (LIMP)
Amirkabir University of Technology (AUT), Iran
{a.khosravani,homayoun}@aut.ac.ir

Abstract

In this paper we propose to estimate the parameters of the probabilistic linear discriminant analysis (PLDA) in text-independent *i*-vector speaker verification framework using a nonparametric form rather than maximum likelihood estimation (MLE) obtained by an EM algorithm. In this approach the between-speaker covariance matrix that represents global information about the speaker variability is replaced with a local estimation computed on a nearest neighbor basis for each target speaker. The nonparametric between- and within-speaker scatter matrices can better exploit the discriminant information in training data and is more adapted to sample distribution especially when it does not satisfy Gaussian assumption as in *i*-vectors without length-normalization. We evaluated this approach on the recent NIST 2016 speaker recognition evaluation (SRE) as well as NIST 2010 core condition and found significant performance improvement compared with a generatively trained PLDA model.

Index Terms: speaker recognition, PLDA, nonparametric, NIST SRE

1. Introduction

The speaker recognition technology based on *i*-vector framework dominates the research field due to its state-of-the-art performance and its suitability for machine learning techniques [1, 2, 3]. This representation provides a fixed-length low-dimensional vector from an arbitrary duration speech segment which models both speaker and channel variability. Either an unsupervised Gaussian mixture model (GMM) or a supervised ASR acoustic model can be used to compute the frame-level soft alignments required to extract an *i*-vector, where the latter integrates the information from speech content directly into the model. In order to produce speaker verification score, a log likelihood ratio between the same-speaker and different-speaker hypotheses is computed for each pair of *i*-vectors corresponding to the segments in the verification trial.

Probabilistic Linear Discriminant Analysis (PLDA) [4] is a generative model that provides a probabilistic approach to model both between-speaker variability which characterizes speaker information and within-speaker variability which characterizes channel or distortion observed in different utterances of individual speakers in *i*-vector space. The Gaussian PLDA is based on the assumption that the speaker and channel components are statistically independent with Gaussian distribution. However, the non-Gaussian behavior of speaker and channel effects can best be modeled by a heavy-tailed version of PLDA which replaces Gaussian with Student's *t* distribution [5]. But an alternative approach would be to reduce the non-Gaussian behavior of *i*-vectors through length-normalization which provides similar performance and is more preferred in practice [6]. The Gaussian PLDA which allows for direct evaluation of log-likelihood ratio verification score has obtained the state-of-the-art performance on length-normalized *i*-vectors.

An estimation of the verification score using a discriminative rather than a generative model has been proposed in [7]. In this approach the speaker verification score for a pair of *i*-vectors is computed using the same functional form derived from the PLDA generative model but the parameters are estimated using a discriminative training criterion that discriminates between same-speaker and different-speaker trials rather than using maximum likelihood estimation (MLE) of PLDA model parameters [8]. Training can be performed by means of support vector machines (SVMs) and a suitable kernel derived from the PLDA generative model [9]. However, it has been shown that discriminative training of a probabilistic model needs more training data and only provides competitive performance with the one obtained by a generative model [10, 11].

In this paper, we propose to estimate PLDA model parameters to compute verification score for a pair of *i*-vectors representing a trial using nearest neighbor (NN) technique rather than the standard MLE. The proposed method is inspired by the recent success of the nonparametric discriminant analysis (NDA) [12] in speaker recognition [13, 14]. In the proposed approach the between-speaker covariance matrix that represents global information about the speaker variability is replaced with a local estimation for each target speaker and is estimated using speakers with the most similarity to that target speaker. The new formulation in which both the within-speaker and between-speaker scatter matrices are redefined in nonparametric form can better exploit the discriminant information in training data. Moreover, they lead to a model more adapted to sample distribution especially in non-Gaussian case of *i*-vectors without length-normalization. The estimated parameters will be used to compute verification score in the same way as in generative PLDA model. Our approach is similar in spirit to the nearest neighbor technique so it is appropriate to use the term nearest neighbor PLDA (NN-PLDA) to refer to the model we propose. We evaluated our approach on the NIST 2010 speaker recognition evaluation (SRE) core telephony trial condition as well as the most recent 2016 SRE. Our main contributions will be to show that modeling between- and within-speaker variability using a nonparametric form for each target speaker leads to significant gains in speaker verification accuracy.

The remainder of this paper is organized as follows. Section 2 describes the standard PLDA modeling technique for speaker verification. Our proposed NN-PLDA technique is presented in Section 3. In Section 4, we describe systems and experimental evaluations on the protocols defined by NIST and present results in Section 5. Finally, Section 6 concludes the paper.

http://dx.doi.org/10.21437/Interspeech.2017-829

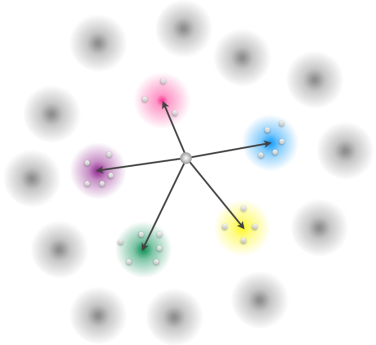

Figure 1: *Illustration of vectors used to compute the between-speaker scatter matrix for a given target speaker. The spheres indicate speakers and small circles represent i-vectors.*

2. PLDA Speaker Verification

Probabilistic Linear Discriminant Analysis (PLDA) provides a powerful mechanism to distinguish between-speaker variability which characterizes speaker information, from all other sources of undesired variability that characterize channel or distortions. Since *i*-vectors are assumed to be generated by some generative model, we can break it down into statistically independent speaker and channel components with Gaussian distributions [6]. A standard Gaussian PLDA assumes that an *i*-vector $\mathbf{w}$, is modelled according to

$$\mathbf{w} = \mathbf{m} + \mathbf{V}\mathbf{y} + \varepsilon. \quad (1)$$

where, $\mathbf{m}$ is the mean of *i*-vectors, the columns of matrix $\mathbf{V}$ contains the basis for the between-speaker subspace, the latent identity variable $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes the speaker factor that represents the identity of the speaker and the residual ε which is normally distributed with zero mean and full covariance matrix $\mathbf{S}_W$, represents within-speaker variability. The maximum likelihood estimation (MLE) of PLDA model parameters are obtained from a large collection of training data using an EM algorithm. For each two trial *i*-vectors $\mathbf{w}_{tar}$ and $\mathbf{w}_{test}$, the Gaussian PLDA score is computed using

$$s = \mathbf{w}_{tar}^T \mathbf{Q} \mathbf{w}_{tar} + \mathbf{w}_{test}^T \mathbf{Q} \mathbf{w}_{test} + 2\mathbf{w}_{tar}^T \mathbf{P} \mathbf{w}_{test} + c, \quad (2)$$

in which

$$\mathbf{Q} = \mathbf{S}_T^{-1} - (\mathbf{S}_T - \mathbf{S}_B \mathbf{S}_T^{-1} \mathbf{S}_B)^{-1}, \quad (3)$$

$$\mathbf{P} = \mathbf{S}_T^{-1} \mathbf{S}_B (\mathbf{S}_T - \mathbf{S}_B \mathbf{S}_T^{-1} \mathbf{S}_B)^{-1}. \quad (4)$$

and $\mathbf{S}_B = \mathbf{V}\mathbf{V}^T$ and $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$.

3. Nearest-Neighbor PLDA (NN-PLDA)

The Gaussian PLDA (G-PLDA) is a probabilistic approach that models both between- and within-speaker variances using the parametric form of the scatter matrix which relies on the underlying assumption that speaker's *i*-vectors satisfy the Gaussian distribution. In this section we describe how we estimate the PLDA model parameters in nonparametric form to better exploit the discriminant information in training data and to deal with the non-Gaussian behavior of the *i*-vectors.

The nonparametric between-speaker scatter matrix measures between-speaker scatter on a local basis in the neighborhood of the target speaker. It estimates the contribution of the nearest neighbor speakers for the calculation of the between-speaker scatter matrix which leads to a more flexible and accurate estimation of the between-class variability. Given a collection of speakers and $\mathcal{R}_s = \{\mathbf{w}_1(s), \mathbf{w}_2(s), ..., \mathbf{w}_{R_s}(s)\}$ *i*-vectors for each speaker s (corresponding to different recordings of the speaker) for training, we define for each target speaker $\mathbf{w}_{tar}$ the nonparametric between-speaker scatter matrix as

$$\mathbf{S}_B^{tar} = \frac{1}{K} \sum_{s \in NN_{tar}^K} (\mathbf{w}_{tar} - \mathbf{w}(s))(\mathbf{w}_{tar} - \mathbf{w}(s)))^T \quad (5)$$

where NN_{tar}^K is a set of K speakers in the training set with the most similarity to the target *i*-vector $\mathbf{w}_{tar}$ and $\mathbf{w}(s) = \frac{1}{R_s} \sum_{r=1}^{R_s} \mathbf{w}_r(s)$. An illustration of the vectors used to compute between-speaker scatter matrix is given in Figure 1. The within-speaker scatter matrix is also defined in a nonparametric form as

$$\mathbf{S}_W = \frac{1}{RK} \sum_s \sum_{r=1}^{R_s} \sum_{k=1}^{K} (\mathbf{w}_r(s) - NN_k(\mathbf{w}_r(s), s)) (\mathbf{w}_r(s) - NN_k(\mathbf{w}_r(s), s))^T. \quad (6)$$

where $R = \sum_s R_s$ and $NN_k(\mathbf{w}_r(s), s)$ is the kth nearest neighbor *i*-vector to $\mathbf{w}_r(s)$ from the same speaker. Since the number of recordings available for each speaker is usually limited, setting K to 10 almost encompass all the recordings of each speaker.

In order to find the K nearest neighbor speakers, we first set K to the number of speakers available in the training set and find the most similar ones based on their verification scores. An empirical value between 1000-1500 for K achieved the best performance on the protocol defined by the NIST.

In order to compute the verification score, each target speaker $\mathbf{w}_{tar}$ is associated with a specific between-speaker scatter matrix $\mathbf{S}_B^{tar}$ computed during enrolment. For each pair of trial *i*-vectors $\mathbf{w}_{tar}$ and $\mathbf{w}_{test}$ the verification score is computed using the same functional form (Eq. 2) but $\mathbf{S}_B$ and $\mathbf{S}_W$ are replace with $\mathbf{S}_B^{tar}$ and $\mathbf{S}_T^{tar} = \mathbf{S}_B^{tar} + \mathbf{S}_W$, respectively.

4. Experiments

4.1. Experimental Protocol

4.1.1. Evaluation Data

We evaluated our system on telephone speech portion of the NIST 2010 SRE (SRE'10) core condition (det 5) as well as on the more recent NIST 2016 SRE (SRE'16). Our system submission based on *i*-vector/PLDA provided remarkable results on the fixed condition of SRE'16 evaluation protocol. This motivates us to evaluate NN-PLDA on this protocol as well to demonstrate that our proposed methods can generalize well. In comparison to 2010 evaluation, NIST marked a major shift from English towards Austronesian and Chinese languages making language mismatch a major challenge. The focus was on telephone speech data recorded over a variety of handset types with varying duration collected outside North America, spoken in Tagalog and Cantonese (referred to as the major languages used in test set) and Cebuano and Mandarin (referred to as the minor languages used in development set).

4.1.2. Training Data

A portion of the fixed training condition of the SRE'16 which limits the system training to specific data sets is used to build

our system. Since the evaluation data includes only conversational telephone speech, we have utilized the telephone speech data from NIST SRE 2004-2010 and the Switchboard corpora (Switchboard Cellular Parts I and II, Switchboard2 Phase I,II and III) for different steps of system training. These data include 19,556 and 25,835 English utterances from 1,925 male and 2,603 female speakers as well as 1,428 and 2,657 non-English utterances (34 different language dialects) from 274 male and 489 female speakers, respectively. In SRE'10 experiment, however, the SRE'10 evaluation data was held out from training. We have also included the unlabelled CallMyNet development set of 2472 telephone calls from both minor and major languages in SRE'16 system training.

4.1.3. *Performance Measurement*

As performance metrics we used the equal error rate (EER) and minimum detection cost function ($\mathrm{DCF}^{\mathrm{min}}$) defined by NIST in each of the evaluation protocols. In 2016 evaluation, the actual detection cost function ($\mathrm{DCF}_{16}^{\mathrm{act}}$), is used as the primary metric which is computed from trial scores by applying two detection thresholds, $log(\beta_1)$ and $log(\beta_2)$ for $P_{Target_1} = 0.01$ and $P_{Target_2} = 0.005$, respectively, where β is defined as:

$$\beta = \frac{C_{FalseAlarm}}{C_{Miss}} \times \frac{1 - P_{Target}}{P_{Target}}, \tag{7}$$

$C_{FalseAlarm} = 1$ and $C_{Miss} = 1$ are cost of a spurious detection and cost of a missed detection, respectively. Thus, the primary metric is defined as

$$C_{Primary} = \frac{C_{Norm_{\beta_1}} - C_{Norm_{\beta_2}}}{2}, \tag{8}$$

where

$$C_{Norm} = P_{Miss|Target} + \beta \times P_{FalseAlarm|NonTarget}. \tag{9}$$

To compute the final score, $C_{Primary}$ will be calculated on different trial partitions including enrollment (1-segment or 3-segment), language (Tagalog or Cantonese), sex (Male or Female), and phone number match (same or different). The average of $C_{Primary}$'s will form the final score. The counts for each partition will be equalized before pooling (the term Equalized refer to this). Also, $\mathrm{DCF}_{16}^{\mathrm{min}}$will be computed by using the decision thresholds that minimize the detection cost function. The scripts to calculate the primary metric is provided by NIST.

4.2. System Configuration

4.2.1. *Front-end processing*

For speech parametrization, we extracted 60-dimensional MFCC and 39-dimensional PLP (including their first and second order derivatives) as acoustic features. The SRE'16 experiments indicated that the combination of these two feature sets performs particularly well in score fusion. For each utterance, the features are centered using a short-term (3s window) cepstral mean and variance normalization (ST-CMVN) after employing a DNN-HMM speech activity detector (SAD) to drop non-speech frames [15].

We trained on acoustic features, a full covariance, gender-independent UBM model with 2048 Gaussians followed by a 600-dimensional *i*-vector extractor to establish MFCC- and PLP-based GMM *i*-vector systems for the SRE'16 experiment. A scaling factor of 0.33 was used on Baum-Welch statistics which gives a slight edge on detection error tradeoff (DET)

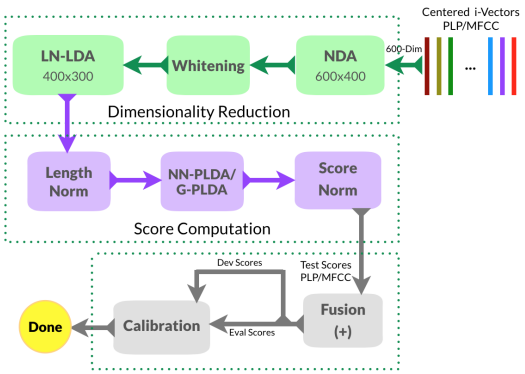

Figure 2: *The flowchart representing the back-end computation of our speaker verification system on NIST 2016 SRE.*

curve resulting in better detection cost function [16]. In SRE'10 experiment, we developed on MFCC acoustic features, the same dimensional GMM as well as DNN *i*-vector systems. As input to the DNN, we used 20 MFCC features, including energy, along with 10 frames from each side to produce 420 dimensional feature vector. A four-layer fully-connected hidden layer DNN model with 420 input nodes and 1500 nodes per hidden layer was trained with cross-entropy using GMM-HMM alignments which was trained on 300 hours of clean English telephone speech from Switchboard data set. The DNN output layer, which applies a softmax function, consists of 3891 units corresponding to GMM-HMM senones which provides posterior probability to compute Balm-Welch statistics [15]. The open-source Kaldi software has been used for all these processing steps [17].

4.2.2. *Back-end processing*

To compensate for the effect of channel or distortion, a non-parametric linear discriminant analysis (NDA) as a substitute for LDA is used which showed superior performance in both speaker and language recognition tasks [13, 18]. NDA outperformed LDA due to the ability in capturing the local structure and boundary information within and across different speakers. We applied an NDA projection matrix (600×400 in SRE'16 and 600×200 in SRE'10 experiment) computed using the 10 nearest sample information on centered *i*-vectors. The resulting dimensionality reduced *i*-vectors are then passed through a whitening transformation.

Given that the emphasis of the SRE'16 is on language mismatch between training and enrollment/test data, beside from using the unlabeled in-domain development data for whitening and centralizing data, we employed two approaches to alleviate the domain mismatch. First, we used a language normalization technique [19] by extending Source-Normalized LDA (SN-LDA) [20] in order to mitigate variations that separate languages on the multilingual training data. Second, in G-PLDA modeling, we added to the training set replicate copies of the labeled in-domain development set to extend their contribution to almost 10% of the training set which seems reasonable to exploit the language and channel information in the development set that mirror that of evaluation set. For NN-PLDA, however, the labeled in-domain development set was used in the estimation of both the between-speaker and within-speaker scatter matrices. Two within-speaker scatter matrices were computed according to Eq. 6 for both in-domain labeled development set and out-domain training set. An interpolation of the two forms

Table 1: *Performance comparison of the G-PLDA and NN-PLDA on the development and evaluation protocols of NIST SRE 2016.*

		Unequalized			Equalized		
Protocol		EER	$\mathrm{DCF}_{16}^{\mathrm{min}}$	$\mathrm{DCF}_{16}^{\mathrm{act}}$	EER	$\mathrm{DCF}_{16}^{\mathrm{min}}$	$\mathrm{DCF}_{16}^{\mathrm{act}}$
Development	G-PLDA	17.92	0.6216	0.6452	16.11	0.6377	0.6611
	NN-PLDA	**16.55**	**0.5994**	**0.6314**	**15.39**	**0.6113**	**0.6466**
Evaluation	G-PLDA	10.84	0.6747	0.7044	11.23	0.6758	0.6983
	NN-PLDA	**10.55**	**0.6357**	**0.6638**	**10.76**	**0.6329**	**0.6365**

the adapted within-speaker scatter matrix. In our experiment an adaptation of 20% was used.

In SRE'16, our experiments indicate that score normalization have a great impact on the performance of the recognition system. We used the symmetric s-norm proposed in [5] which normalizes the score of each pair of trial *i*-vectors by matching them against the unlabelled in-domain development set. Finally, the calibrated log-likelihood-ratio scores are obtained using the BOSARIS Toolkit [21] and the development protocol as training. To avoid the bias due to training/calibration data overlap, the calibration weights have been computed on the development protocol using a version of models trained without the labeled in-domain set. A flow chart representing the back-end computation of our speaker verification system on NIST SRE'16 is given in Figure 2.

Table 2: *Performance comparison of G-PLDA and NN-PLDA on the NIST 2010 core condition (det5) obtained with GMM i-vectors without length-normalization.*

	EER [%]	$\mathrm{DCF}_{08}^{\mathrm{min}}$	$\mathrm{DCF}_{10}^{\mathrm{min}}$
G-PLDA	2.54	0.1114	0.3557
NN-PLDA	**1.88**	**0.0789**	**0.2413**

5. Results and Discussion

In this section, we summarize the results obtained with the experimental setup presented in Section 4.

5.1. NIST 2010 SRE

We began by evaluating on the core condition of SRE'10 the generatively trained PLDA (G-PLDA) which serves as our baseline with 200-dimensional speaker subspace and our proposed nearest-neighbor PLDA (NN-PLDA) on *i*-vectors without length-normalization; results are summarized in Table 2. The results indicate an improvement of more than 30% in $\mathrm{DCF}_{10}^{\mathrm{min}}$and 26% in EER relative to the standard PLDA. This indicates the non-Gaussian behavior of *i*-vectors induced by the *i*-vector extraction mechanism and the ability of NN-PLDA to generate a model more adapted to the sample space.

We continue our experiments by applying length-normalization to reduce the Gaussian behavior of *i*-vectors. Table 3 shows the results using both GMM and DNN *i*-vector systems. As can be seen, NN-PLDA outperforms G-PLDA even with length-normalization which approximately Gaussianize the *i*-vector distribution. This indicates the ability of NN-PLDA to better exploit the discriminant information in training data by explicitly emphasizing on the speakers near the target speaker. Interestingly, we observe that NN-PLDA without length-normalization in Table 2 is equivalent to G-PLDA with length normalization in low false positive rates (Table 3) which are more interested in real applications and NIST evaluation plans. We also found that the fusion (simple summation) of GMM and DNN systems leads to substantial performance improvement due to architectural differences between both systems. Results indicate more than 15% improvement relative to the G-PLDA in terms of all performance measures.

Table 3: *Performance comparison of G-PLDA and NN-PLDA on the NIST 2010 core condition (det5) obtained with both GMM and DNN i-vectors with length-normalization. The fusion of GMM and DNN systems is highlighted.*

		EER [%]	$\mathrm{DCF}_{08}^{\mathrm{min}}$	$\mathrm{DCF}_{10}^{\mathrm{min}}$
G-PLDA	GMM	1.43	0.0792	0.2917
	DNN	0.99	0.0434	0.1789
	Fusion	0.85	0.0412	0.1593
NN-PLDA	GMM	1.18	0.0686	0.2286
	DNN	0.85	0.0433	0.1511
	Fusion	**0.56**	**0.0354**	**0.1339**

5.2. NIST 2016 SRE

In this Section we present the results obtained using the protocol provided by NIST on the development and evaluation set of SRE'16. The results are shown in Table 1. The results are reported using the score fusion of MFCC- and PLP-based (a simple sum of both MFCC and PLP scores) GMM *i*-vector systems. We found that the fusion of these two systems results in a relative improvement of almost 10%, compared to MFCC-based system alone, in terms of both $\mathrm{DCF}_{16}^{\mathrm{min}}$and $\mathrm{DCF}_{16}^{\mathrm{act}}$. As already been shown in [22], our experiments with DNN *i*-vector system on non-English evaluation data resulted in unsatisfactory performance compared to the GMM *i*-vector system. The ability of NN-PLDA to exploit the unlabeled development data and also the small in-domain labeled development data to provide a more adapted model to the evaluation data is considerable and the results reflect this.

6. Conclusions

We have presented a nonparametric approach to estimate the PLDA model parameters. We evaluated the proposed nonparametrically trained PLDA model on NIST SRE'10 and found substantial performance improvement. The proposed NN-PLDA is more adapted to sample distribution especially in non-Gaussian case of *i*-vectors without length-normalization and can better exploit the discriminant information in training data. We observed a reduction of 16% in $\mathrm{DCF}_{10}^{\mathrm{min}}$on the core condition of SRE'10 (det5). We also conducted an experiment on the most recent NIST SRE'16 and also observed an improvement of 9% in $\mathrm{DCF}_{16}^{\mathrm{act}}$on the evaluation data. For future works it would be interesting to analyze the effect of speaker population size on the performance of NN-PLDA.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The nist 2014 speaker recognition i-vector machine learning challenge," in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014.

[3] A. Khosravani and M. Homayounpour, "Linearly constrained minimum variance for robust i-vector based speaker recognition," in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014, pp. 249–253.

[4] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[5] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, 2010, p. 14.

[6] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *INTERSPEECH*, 2011, pp. 249–252.

[7] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4832–4835.

[8] N. Brummer, "Em for probabilistic lda," Agnitio Research, Tech. Rep., 2010. [Online]. Available: https://sites.google.com/site/nikobrummer

[9] S. Cumani, N. Brümmer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i-vector space," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4852–4855.

[10] J. Rohdin, S. Biswas, and K. Shinoda, "Constrained discriminative plda training for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1670–1674.

[11] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[12] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009.

[13] S. O. Sadjadi, J. W. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition." in *INTERSPEECH*, 2014, pp. 1860–1864.

[14] S. O. Sadjadi, S. Ganapathy, and J. Pelecanos, "The ibm 2016 speaker recognition system," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, Bilbao, Spain, June 21-24 2016, pp. 174–180. [Online]. Available: http://www.isca-speech.org/archive/odyssey_2016/pdfs_stamped/42.pdf

[15] A. Khosravani and M. M. Homayounpour, "A plda approach for language and text independent speaker recognition," *Computer Speech & Language*, vol. 45, pp. 457–474, 2017.

[16] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7649–7653.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[18] S. O. Sadjadi, J. W. Pelecanos, and S. Ganapathy, "Nearest neighbor discriminant analysis for language recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4205–4209.

[19] M. McLaren, M. I. Mandasari, and D. A. van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, Singapore, 2012, pp. 55–61.

[20] M. Mclaren and D. Van Leeuwen, "Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

[21] N. Brümmer and E. de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, 2011.

[22] P. Mat, J. H. Cernock *et al.*, "Analysis of dnn approaches to speaker identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5100–5104.