Interspeech 2018
2-6 September 2018, Hyderabad

A Generalization of PLDA for Joint Modeling of Speaker Identity and Multiple Nuisance Conditions

Luciana Ferrer[1], *Mitchell McLaren*[2]

[1]Instituto de Investigación en Ciencias de la Computación, CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina
[2]Speech Technology and Research Lab, SRI International, Menlo Park, USA

lferrer@dc.uba.ar, mitchell.mclaren@sri.com

Abstract

Probabilistic linear discriminant analysis (PLDA) is the leading method for computing scores in speaker recognition systems. The method models the vectors representing each audio sample as a sum of three terms: one that depends on the speaker identity, one that models the within-speaker variability, and one that models any remaining variability. The last two terms are assumed to be independent across samples. We recently proposed an extension of the PLDA method, which we termed Joint PLDA (JPLDA), where the second term is considered dependent on the type of nuisance condition present in the data (e.g., the language or channel). The proposed method led to significant gains for multilanguage speaker recognition when taking language as the nuisance condition. In this paper, we present a generalization of this approach that allows for multiple nuisance terms. We show results using language and several nuisance conditions describing the acoustic characteristics of the sample and demonstrate that jointly including all these factors in the model leads to better results than including only language or acoustic condition factors. Overall, we obtain relative improvements in detection cost function between 5% and 47% for various systems and test conditions with respect to standard PLDA approaches.

Index Terms: speaker recognition, probabilistic linear discriminant analysis

1. Introduction

PLDA [1] is the leading scoring technique for speaker recognition [2, 3, 4, 5, 6]. It assumes that each sample is represented by a feature vector of fixed dimension and that this vector can be modeled as a sum of three terms: a term that depends on the sample's class, a term that models within-class variability and is assumed independent across samples, and a final term that models any remaining variability and is also independent across samples. These assumptions imply that all samples from the same class are independent of each other and also independent of the samples from other classes once the class is known. This assumption is incorrect for many training datasets where samples come from a small set of distinct conditions like microphones, languages, or speech styles. In these cases, samples corresponding to the same condition will most likely be statistically dependent.

In a recent publication [7], we proposed an extension of PLDA where the term that models the within-class variability is considered dependent on a label describing the nuisance condition of the sample. We showed in [8] that this approach gives large gains in multilingual speaker recognition when using the language as the nuisance condition. In this work, we extend the proposed method to allow for multiple nuisance terms corresponding to different conditions that are assumed to independently and additively affect the vector representing the sample. We propose a heuristic algorithm for model training that is simple to implement, effective, and computationally fast, as well as a scoring procedure that does not require knowledge of the nuisance conditions during testing.

The literature proposes a few approaches that generalize PLDA in a way that makes its parameters dependent on a condition label. The simplest approach of this family is training a separate PLDA model for each condition, as proposed by [9]. Nevertheless, in this paper, the authors show that pooling the data from all conditions, as also proposed by [10], leads to better performance than training separate models. In the tied PLDA model proposed by [11], one PLDA model is trained for each condition, but these models are tied by forcing the latent variable corresponding to each class to be the same across all conditions. The approach was shown to outperform standard PLDA with pooled training data when each class in the training data is seen under both considered conditions, frontal and profile, in a face recognition task. A similar approach is proposed by [12]; but in this case, the mixture component is not given during training but rather is dependent on a continuous metadata value. The approach is tested by adding noise to the training data at different signal-to-noise (SNR) levels, resulting in gains compared to pooling all the data to train a single PLDA model. The tied PLDA approach, though, does not work well when each speaker in the training data is seen only under a small subset of the conditions of interest (potentially, only one) or when some conditions have much less training data than others [8], which are both common training scenarios.

We show results on two multilingual speaker recognition datasets, one composed of Mixer data [13] and another composed of LASRS data [14], using three different systems to obtain the vectors that represent each sample. We show that JPLDA gives significant improvements over standard PLDA approaches when using language as the nuisance condition for all three systems. These results strenghten the conclusion obtained in [8] where only one of the three systems was used for the experiments. Further, we show that additional gains are obtained by adding nuisance terms for the microphone, noise, codec, and reverberation characteristics of the samples.

2. Standard PLDA

Standard PLDA [1] assumes that the vector m_i representing a certain sample from speaker s_i is given by

$$m_i = \mu + V y_{s_i} + U x_i + \epsilon_i, \tag{1}$$

10.21437/Interspeech.2018-1280

where μ is the global mean of the training data; y_{s_i} is a vector of size R_y, the dimension of the speaker subspace; and x_i is a vector of size R_x, the dimension of the subspace corresponding to the nuisance condition or, as usually called in speaker recognition, the channel. The model assumes that

$$y_{s_i} \sim N(0, I), \forall s_i \tag{2}$$

$$x_i \sim N(0, I), \forall i \tag{3}$$

$$\epsilon_i \sim N(0, D^{-1}), \forall i \tag{4}$$

where the matrix D is assumed to be diagonal. All these latent variables are assumed independent: speaker variables are independent across speakers, and the nuisance variable x_i and noise variable ϵ_i are independent across samples.

The model described above corresponds to the original PLDA formulation, which we will call full PLDA (FPLDA). In speaker recognition, a simplified version of PLDA (SPLDA for the purpose of this paper) is commonly preferred, where the matrix V is full rank, and the nuisance factor is absorbed into the noise factor, which is then assumed to have a full rather than diagonal covariance matrix. Sizov et al. [15] give a comprehensive explanation of the usual flavors of PLDA. The training of PLDA parameters for these two models is done using expectation-maximization (EM) algorithms [4, 16].

When using PLDA for speaker verification, the score for a trial composed of enrollment set E and test set T is computed as the following likelihood ratio (LR):

$$LR = \frac{p(E, T|H_{SS})}{p(E, T|H_{DS})}, \tag{5}$$

where H_{SS} is the hypothesis that the speakers in both sets are the same, while H_{DS} is the hypothesis that the speakers are different. This value can be computed using a closed form using the PLDA model. In our code, we use the formulation derived by [17], equation (34). Note, though, that the last term in that equation should not be there (this mistake was confirmed by one coauthor of the paper). For this work, we assume that each trial is composed of a single enrollment and test sample.

3. Joint PLDA

The joint PLDA model, originally proposed in [7] and then further developed and tested for a multilingual speaker recognition task in [8], is a generalization of PLDA where the nuisance variable is no longer considered independent across samples but instead is potentially shared (tied) across the samples that correspond to the same nuisance condition. The original work considered a single nuisance condition, deriving the EM and scoring formulas for this scenario. In this work, we further extend the model to handle multiple nuisance conditions, assuming that their effect is independent and additive.

We assume that the within-speaker variability can be decomposed into N terms corresponding to different nuisance conditions that could correspond to, for example, the language spoken in the sample, the microphone type, the noise type and level, or any other characteristic of the sample that can be considered to occur independently of all other characteristics. We propose to model vector m_i for sample i as:

$$m_i = \mu + V y_{s_i} + \sum_{j=1}^{N} U_j x^j_{c_{ji}} + \epsilon_i, \tag{6}$$

where, as before, y_{s_i} is a vector of size R_y; j is an index over the nuisance conditions; c_{ji} is the label for condition j for sample i; $x^j_{c_{ji}}$ is a vector of size R_{xj}; and

$$y_{s_i} \sim N(0, I), \forall s_i \tag{7}$$

$$x^j_{c_j} \sim N(0, I), \forall j, c_j \tag{8}$$

$$\epsilon_i \sim N(0, D^{-1}), \forall i \tag{9}$$

where, as before, all variables are assumed independent of each other.

The model's parameters to estimate are $\lambda = \{\mu, V, U_1, \dots, U_N, D\}$. The input data for the training algorithm is now required to have the nuisance condition labels for each sample as well as the speaker labels, as usual.

3.1. Model Training Procedure

In [7], we derive the expectation-maximization equations for training this new model when $N = 1$. The formulation is both significantly more involved than for standard PLDA and computationally costlier. Nevertheless, in [8], we compare results obtained with different numbers of EM iterations when using a smart initialization procedure and random initialization. We show that, when using the proposed smart initialization procedure, running EM is unnecessary: the initial model leads to similar performance as those further refined after many EM iterations. Given this finding, for this work, we train our JPLDA models using a version of the smart initialization procedure proposed in [8] that is generalized to allow for more than one nuisance condition label. The procedure is as follows:

- Initialize the variables $x^j_{c_{ji}}$ and the matrices U_j to zero.
- Iterate the following steps M times:
    - For each condition $j = 1, \dots, N$:
        - Create new training vectors for each sample $\tilde{m}_i = m_i - \sum_{k!=j} U_k x^k_{c_{ki}}$, which subtracts the estimated effect of all conditions except j.
        - Estimate an SPLDA model for condition j using the $\tilde{m}_i$ vectors, the labels for condition k as classes and setting the rank of the V matrix to R_{xj}. Set U_j of the JPLDA model to be the V matrix from this SPLDA model.
        - Estimate new values for the latent variables using this SPLDA model and assign them to $\hat{x}^j_{c_{ji}}$. Note that all samples with the same label for condition j will have the same latent variable.
    - Set $x^j_{c_{ji}} = \hat{x}^j_{c_{ji}}$ for all i and j.
- Create new training vectors for each sample $\tilde{m}_i = m_i - \sum_j U_j x^j_{c_{ji}}$, which subtract the estimated effect of all nuisance conditions.
- Estimate a SPLDA model using these training vectors and the speakers as labels, setting the rank of V to R_y. Set the V and D matrices of the JPLDA model to the estimated values for the corresponding matrices of this SPLDA model. If a JPLDA model with diagonal D is required, take the diagonal part of the estimated D matrix.

In our experiments, we set $M = 10$, which was enough for the values of the latent variables to stabilize.

3.2. Score Generation

As for standard PLDA, we define the score as the likelihood ratio between the two hypotheses: that the speakers are the same and that the speakers are different. Nevertheless, in this case, we need to marginalize both likelihoods over new hypotheses:

that each nuisance condition is the same or different in the two samples. Hence, the LR is computed as follows:

$$LR = \frac{\sum_{h\in\mathcal{H}} p(E,T|H_{SS},h)P(h|H_{SS})}{\sum_{h\in\mathcal{H}} p(E,T|H_{DS},h)P(h|H_{DS})}$$

where H_{SS} is the hypothesis that the speakers for both sets are the same, H_{DS} is the hypothesis that they are different, and $\mathcal{H}$ is the set of all possible combinations of hypothesis about the nuisance conditions $\mathcal{H} = \{(H_{m_1C_1}, \ldots, H_{m_NC_N}) : m_j \in \{S, D\}, \forall j \in \{1, \ldots, N\}\}$, where S and D refer to the hypotheses that the conditions are the same and different. The priors for each combination of hypotheses are computed assuming independence as $P(h|H_{SS}) = \prod_j P(H_{m_jC_j}|H_{SS})$ and, similarly, for $P(h|H_{DS})$. Each individual prior can be set independently as a parameter of the method. This LR value can be computed using a closed form, which can be derived using similar procedures as for a single nuisance condition [18].

4. Experimental Setup

In this section, we describe the training and test datasets and the different speaker recognition systems used in our experiments.

4.1. Speaker Recognition Systems

We compare the different PLDA techniques on vectors extracted using three different procedures. In all cases, we use a speech activity detection system (described in detail in [8]) to discard non-speech frames before extracting the vectors representing each sample.

UBM i-vector system (ubmivs): This a traditional i-vector system, which uses mel-frequency cepstral coefficients (MFCCs) of 20 dimensions appended with deltas and double-deltas, a 2048-component GMM as a universal background model (UBM), and a 400-dimensional i-vector extractor. For more details on this system, see [8].

Hybrid alignment system (hybrivs): The hybrid-alignment framework [19] provides competitive speaker recognition performance across mixed conditions. This system leverages a DNN trained to predict 3450 tied tri-phone states to extract 80-dimensional bottleneck features. These phonetically rich bottleneck features are used to train a UBM of 2048 Gaussians, which is later used to generate frame occupancies or alignments for input audio. The alignments are used to generate zero-order statistics and combined with 20-dimensional MFCCs appended with deltas and double-deltas to calculate first-order statistics. The statistics are used in the training of an i-vector subspace of 400 dimensions, from which i-vectors are extracted for our PLDA experiments. Training data for the DNN included Fisher, Switchboard and Callhome data (more details on the DNN can be found in [20]), while the UBM was trained using the non-degraded signals of the PRISM training set.

Speaker embeddings system (embeddings): Recent advances in speaker recognition have shown a significant improvement by using a deep neural network trained directly to target speaker classes, then extract an embedding (a low- and fix-dimensional vector) rich in speaker information, from a hidden layer in the network for use in subsequent backend classification [21, 22]. Our work in [23] was leveraged for the current study, in which an embeddings network was trained using data from approximately 3,200 speakers from 56,000 audio files sourced from the non-degraded subset of the PRISM training lists, each degraded four times with four different degradation types (16-fold degradation) consisting of noise, reverb, compression, and music. Embeddings of 512 dimensions were extracted from the first hidden layer after the statistics pooling layer. These embeddings are used for the PLDA experiments.

Table 1: *Percentage of samples in the training data for different combinations of nuisance conditions labels. The header shows the number of unique labels for each condition in parenthesis. Labels for each nuisance condition are grouped in two groups for each condition to compute the percentages. Microphone labels (mic) are grouped into telephone (tel) or other microphones (mic); codec (cod), noise (noi), and reverberation (rev) labels are grouped into "yes" or "no" to indicate degraded and non-degraded samples, respectively; and language labels (lan) are converted into English (eng) or non-English (non-eng). The total number of samples is 72,659.*

mic (23)	cod (33)	rev (10)	noi (22)	lan (17)	perc
phn	no	no	no	eng	52.8
mic	no	no	no	eng	15.2
mic	yes	no	no	eng	12.7
mic	no	no	yes	eng	7.6
mic	no	yes	no	eng	7.5
phn	no	no	no	non-eng	4.3

The vectors generated by these three systems are further transformed using linear discriminant analysis (LDA) to 300 dimensions. LDA is trained with the same data used for the PLDA methods, described below. The vectors are then mean and length normalized [24] before training or applying the PLDA models.

4.2. PLDA Training Data

The training data for all PLDA methods is given by the full PRISM training set [25], which contains simulated noisy signals created by adding babble noise to clean signals from Mixer collections at 8, 15, and 20 dB signal-to-noise ratios (SNR) and simulated reverberated signals created by convolving the same clean signals with different room impulse responses at different RT60 reverberation times of 0.3, 0.5, and 0.7. Finally, to this original PRISM training list, we added other degraded signals created by transcoding the clean signals with a number of different codecs. This is the data SRI has been using to train PLDA models for a few years. In this case, though, we discarded a small minority of training samples from languages for which only one or two speakers are available and samples where the language was unavailable or ambiguous. All the degraded data is in English and degraded with only one type of degradation: noise, reverberation, or codec distortion.

The training data is labeled with five nuisance condition labels: (1) language, (2) microphone, (3) noise, (4) reverb, and (5) codec. The language condition labels are given by the language in the sample. The microphone labels are given by a combination of the collection identifier (switchboard, Fisher, etc.) and the microphone label provided with the collection. The labels for the noise, reverb, and codec conditions are given by the type of degradation (noise signal, room type, or codec) and the level of degradation (RT or SNR), plus one label for the non-degraded signals. Table 1 shows statistics for the training data.

4.3. Test Data

We consider four testing conditions, one that uses Mixer data and three that use LASRS data.

The Mixer test data is composed of telephone samples from the Mixer collections [13] from the 2005 to 2010 NIST speaker recognition evaluations, from speakers not used for training. We include 119 samples in Arabic from 21 speakers; 200 samples in Russian from 47 speakers; 309 samples in Thai from 38 speakers; 827 samples in Chinese from 163 speakers; and 5,755 samples in English from 701 speakers. The trials are created by selecting the same number of target and impostor same-language and cross-language trials such that the final set of trials is a balanced union of both types of trials. Further, the same-language trials are created as a balanced union of English versus non-English trials. The final set of trials, which we call *Mixer Cln-mic All-lang* (Cln stands for clean, referring to the fact that samples are not degraded telephone samples, though they could have different types of "wild" degradations), contains 11,522 target trials and 858,119 impostor trials.

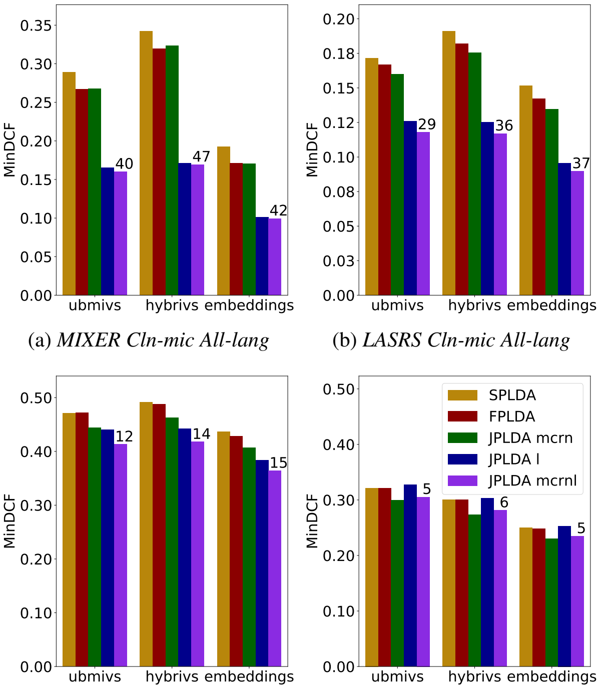
The LASRS test data is composed of samples from a bilingual, multi-model voice corpus [14]. The corpus is composed of approximately 100 bilingual speakers from each of three languages: Arabic, Korean, and Spanish. Each speaker is asked to perform a series of tasks in English and in their native language. Each task is recorded using seven recording devices (a camcorder, desktop, studio, omnidirectional, and three telephone microphones) and repeated in two separate sessions recorded on different days. For our experiments, we use the conversational data from all speakers. The trials are created by enrolling with data from the first recorded session and testing on the second recorded session in each of the two spoken languages. This results in a total of approximately 3.9 million impostor and 34 thousand (K) target trials. This is what we call the *All-mic All-lang* condition. We also subset the trials to include only three microphones that are cleaner and somewhat similar to those seen in training (two of the telephone and the studio microphones). This subset, which we call *Cln-mic All-lang*, contains approximately 715K impostor and 6.2K target trials. Finally, we create another subset including only the English versus English trials. This subset, which we call *All-mic Eng*, contains 783K impostor and 7.8K target trials.

5. Results

Figure 1 shows the minimum detection cost function (DCF) computed with a probability of target of 0.01, a cost of miss of 10, and a cost of false alarm of 1 [26] for all three vector-extraction procedures and all four test sets. The rank R_y is set to 200 for all methods. For FPLDA, the R_x rank was optimized to 40 using the Mixer data. Finally, for JPLDA, we use, in all cases, the maximum rank that can be used for each nuisance condition given by the number of labels for that condition (indicated in Table 1) minus 1. These ranks were not tuned. The prior probability for the same condition was set to 0.1 for all conditions for both speaker hypothesis (see Section 3.2). This was lightly optimized on Mixer data. Yet, values above 0.05 and below 0.5 give similar performance for all JPLDA systems.

Results show that JPLDA with language labels results in large gains for the two *Cln-mic All-lang* test sets that include all languages and microphones that are relatively clean and matched to those seen in training. This is the same conclusion we obtained in [8] for the ubmivs system. Here, we show that this conclusion holds for all three systems tested. For these two test sets, the gain from adding the other nuisance conditions is small. This is likely due to the fact that a majority of the training data matches the acoustic conditions in these test sets.

When the test set includes microphones that are noisier, more distorted, or mismatched to those in training, the gain from JPLDA with language labels becomes smaller, and the advantage of adding the other nuisance conditions during model training becomes more apparent. Overall, we see that the model that considers both language and acoustic conditions at the same time is always similar or better than the best of the two models that consider language or acoustic condition separately.



(a) *MIXER Cln-mic All-lang*

(b) *LASRS Cln-mic All-lang*

(c) *LASRS All-mic All-lang*

(d) *LASRS All-mic Eng*

Figure 1: *MinDCF results for five PLDA approaches. The JPLDA names include the nuisance conditions considered in each case: l (language), m (microphone), c (codec), r (reverberation), and n (noise). The numbers on top of the last bar in each group are the relative gains with respect to FPLDA.*

6. Conclusions

We presented a generalization of PLDA that enables modeling of dependencies between samples due to common nuisance conditions. This new model, which we termed Joint PLDA (JPLDA) for its ability to jointly model speaker identity and a nuisance condition of interest, was recently shown to outperform PLDA in a multilingual speaker recognition task when using language as the nuisance condition. In this work, we further generalize the approach to allow for multiple nuisance conditions and propose a simple and fast training algorithm, as well as a scoring procedure that does not require knowledge of the nuisance condition labels at test time. Results show that joint modeling of language and a set of acoustic conditions leads to the best results compared to standard PLDA and to JPLDA using only language or acoustic conditions. Further work includes investigating ways of automatically estimating the condition labels when the training data does not have these labels available, generalizing the scoring formulation to multiple enrollment samples, and deriving the EM algorithm for multiple nuisance conditions.

7. Acknowledgements

This material is based upon work supported partly by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement A: Approved for Public Release, Distribution Unlimited.

8. References

[1] S. Prince, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proceedings of the International Conference on Computer Vision*, 2007.

[2] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in *Proc. Odyssey-10*, Brno, Czech Republic, Jun. 2010, keynote presentation.

[3] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *Proc. ICASSP*, Prague, May 2011.

[4] N. Brümmer, “EM for probabilistic LDA,” Available at https://sites. google.com/site/nikobrummer/EMforPLDA.pdf, Tech. Rep., 2010.

[5] M. Senoussaoui, P. Kenny, N. Brümmer, N.mmer, E. De Villiers, and P. Dumouchel, “Mixture of PLDA models in i-vector space for gender-independent speaker recognition.” in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 25–28.

[6] P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance UBM and heavy-tailed PLDA in i-Vector speaker verification,” in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[7] L. Ferrer, “Joint probabilistic linear discriminant analysis,” *arXiv:1704.02346v2*, 2017.

[8] L. Ferrer and M. McLaren, “Joint PLDA for simultaneous modeling of two factors,” *accepted for publication in Journal of Machine Learning Research*, 2018.

[9] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, “Multi-condition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition,” in *Proc. ICASSP*. Kyoto: IEEE, Mar. 2012, pp. 4257–4260.

[10] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, “Towards noise robust speaker recognition using probabilistic linear discriminant analysis,” in *Proc. ICASSP*, Kyoto, Mar. 2012.

[11] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince, “Probabilistic models for inference about identity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012.

[12] M.-W. Mak, X. Pang, and J.-T. Chien, “Mixture of plda for noise robust i-vector speaker verification,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 130–142, 2016.

[13] C. Cieri, L. Corson, D. Graff, and K. Walker, “Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora,” in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[14] S. D. Beck, R. Schwartz, and H. Nakasone, “A bilingual multi-modal voice corpus for language and speaker recognition (LASR) services,” in *Odyssey: The Speaker and Language Recognition Workshop*, 2004.

[15] A. Sizov, K. A. Lee, and T. 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.

[16] N. Brümmer, “EM for simplified PLDA,” Available at https://sites. google.com/site/nikobrummer/EMforSPLDA.pdf, Tech. Rep., 2010.

[17] S. Cumani, O. Plchot, and P. Laface, “On the use of i-vector posterior distributions in probabilistic linear discriminant analysis,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 846–857, 2014.

[18] L. Ferrer, “Scoring formulation for multi-condition joint plda,” *arXiv:1803.03684*, 2018.

[19] M. McLaren, D. Castan, L. Ferrer, and A. Lawson, “On the issue of calibration in DNN-Based speaker recognition systems,” in *Proc. Interspeech*, San Francisco, September 2016.

[20] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Proc. ICASSP*, Florence, Italy, May 2014.

[21] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[22] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Proc. Interspeech*, Stockholm, August 2017.

[23] M. McLaren, D. Castan, M. Nandwana, L. Ferrer, and E. Yilmaz, “How to train your speaker embeddings extractor,” in *Proc. of Speaker Odyssey*, Les Sables d'Olonne, France, June 2018.

[24] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[25] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, “Promoting robustness for speaker modeling in the community: The PRISM evaluation set,” in *Proceedings of SRE11 Analysis Workshop*, Atlanta, USA, Dec. 2011.

[26] “NIST SRE10 evaluation plan,” http://www.itl.nist.gov/iad/mig/tests/sre/ 2010/NIST_SRE10_evalplan.r6.pdf.