Handling Recordings Acquired Simultaneously over Multiple Channels with PLDA

Jesús Villalba[1], *Mireia Diez*[2], *Amparo Varona*[2], *Eduardo Lleida*[1]

[1] GTC, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
[2] GTTS, Department of Electricity and Electronics
University of the Basque Country UPV/EHU, Spain
villalba@unizar.es mireia.diez@ehu.es

Abstract

In some speaker recognition scenarios we find conversations recorded simultaneously over multiple channels. That is the case of the interviews in the NIST SRE dataset. To take advantage of that, we propose a modification of the PLDA model that considers two different inter-session variability terms. The first term is tied between all the recordings belonging to the same conversation whereas the second is not. Thus, the former mainly intends to capture the variability due to the phonetic content of the conversation while the latter tries to capture the channel variability. We test this approach on the NIST SRE12 core condition using multiple channels per interview to enroll the speakers. The proposed approach improves the minimum DCF by 26–29 % on telephone speech and by 1–8% on interviews compared to the standard PLDA (scored *by the book*).

Index Terms: Speaker recognition, PLDA, i-vectors, simultaneous recordings.

1. Introduction

In some speaker recognition scenarios speech is recorded simultaneously by several microphones. That is the case of the NIST speaker recognition evaluations (SRE) where interviews are recorded over 14 different microphones [1]. However, in the context of NIST evaluations, this fact has never been explicitly exploited and, in practice, simultaneous recordings are treated as recordings from different conversations [2–4].

In the context of microphone arrays, beam-forming algorithms create a direction dependent gain pattern that enhances the speech in the direction of the target speaker [5–8]. However, usually, as in NIST interviews, microphones are not configured in arrays so we cannot always apply those techniques.

In [9], we find another example of exploiting simultaneous recordings. There stereo data is used to train a linear transformation from a noisy environment to a clean environment using phoneme dependent multi-environment models based linear normalization (PD-MEMLIN). That transformation is applied to clean noisy signals.

In this work, we propose an extension of the well known PLDA model [10] that takes advantage of conversations recorded over multiple channels. We consider a PLDA with two terms of inter-session variability where the first one intends to account for inter-conversation variability and the second one for intra-conversation variability (microphone variability).

The paper is organized as follows. Section 2 introduces the baseline PLDA. Section 3 describes the extended PLDA and the mathematical formulation. Section 4 presents our experimental setup and results. Finally, section 5 shows our conclusions.

2. PLDA

PLDA [10] is a generative model that assumes that an i-vector ϕ_{ij} from the session j of speaker i can be written as:

$$\phi_{ij} = \mu + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij} + \epsilon_{ij} \tag{1}$$

where μ is a speaker independent term, $\mathbf{V}$ is a low rank matrix of eigenvoices, $\mathbf{y}_i$ is the speaker factor vector, $\mathbf{U}$ is a low rank matrix of eigenchannels, $\mathbf{x}_{ij}$ is the channel factor vector and ϵ_{ij} is an offset that accounts for the rest of channel variability not included in $\mathbf{U}\mathbf{x}_{ij}$.

Gaussian priors are assumed for the latent variables:

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i|\mathbf{0},\mathbf{I}) \tag{2}$$
$$\mathbf{x}_{ij} \sim \mathcal{N}(\mathbf{x}_{ij}|\mathbf{0},\mathbf{I}) \tag{3}$$
$$\epsilon_{ij} \sim \mathcal{N}\left(\epsilon_{ij}|\mathbf{0},\mathbf{D}^{-1}\right) \tag{4}$$

where $\mathcal{N}$ denotes a Gaussian distribution; and $\mathbf{D}$ is a diagonal precision matrix. The parameters μ, $\mathbf{V}$, $\mathbf{U}$ and and $\mathbf{D}$ are trained from a development database by ML and MD iterations [11]. We denote by $\mathcal{M}$ the set of all the model parameters.

If the $\mathbf{U}$ matrix is full rank, this model is equivalent to a simplified model (SPLDA) without $\mathbf{U}$ and with full covariance $\mathbf{D}$ [4].

3. PLDA with two types of inter-session variability

3.1. Model description

Let's suppose that we have available i-vectors from conversations that were recorded simultaneously over different channels or noisy conditions. We define a new PLDA model such as an i-vector ϕ_{ijl} of speaker i, conversation j and recorded over a channel l can be written as:

$$\phi_{ijl} = \mu + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij} + \epsilon_{ijl} \tag{5}$$

where the channel factors $\mathbf{x}_{ij}$ are tied between all the i-vectors belonging to the same conversation whereas the channel offset ϵ_{ijl} is different for each i-vector. In this case the prior for ϵ_{ijl} is chosen to be

$$\epsilon_{ijl} \sim \mathcal{N}\left(\epsilon_{ijl}|\mathbf{0},\mathbf{W}^{-1}\right) \tag{6}$$

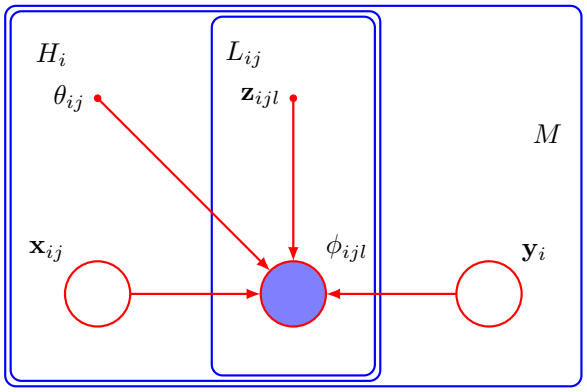

Figure 1: BN PLDA with two types of inter-session variability

where $\mathbf{W}$ is a full-rank precision matrix.

This model intends to decompose the inter-session variability into two terms that account for different types of variability. The term $\mathbf{U}\mathbf{x}_{ij}$, being tied for all the channel types, should account for variability different from the channel, mainly phonetic variability. On the contrary, the term ϵ_{ijl} should account for the channel variability.

Figure 1 depicts the Bayesian network of the model. We use the notation of [12]. The variables θ_{ij} and $\mathbf{z}_{ijl}$ indicate who is the speaker in conversation ij and which conversation corresponds to segment ijl, respectively. Each speaker has H_i conversations and each conversation is recorded over L_{ij} channels.

3.2. Posterior of the hidden variables

In order to implement this model, we need the posterior distribution of the latent factors. That is needed for the E-step of the EM algorithm as much as for the evaluation of likelihood ratios. The posterior can be decomposed into two factors:

$$P\left(\mathbf{y}_i, \mathbf{X}_i|\mathbf{\Phi}_i, \mathcal{M}\right) = P\left(\mathbf{X}_i|\mathbf{y}_i, \mathbf{\Phi}_i, \mathcal{M}\right) P\left(\mathbf{y}_i|\mathbf{\Phi}_i, \mathcal{M}\right) \quad (7)$$

where $\mathbf{\Phi}_i$ is the set of all the i-vectors of speaker i and $\mathbf{X}_i$ the set of all the channel factors.

3.2.1. *Sufficient statistics*

We find convenient to define some sufficient statistics. The conversation dependent statistics are

$$\mathbf{F}_{ij} = \sum_{l=1}^{L_{ij}} \phi_{ijl} \quad (8)$$

$$\overline{\mathbf{F}}_{ij} = \mathbf{F}_{ij} - L_{ij}\mu \quad (9)$$

where L_{ij} is the number of simultaneous channels for the conversation ij. The speaker dependent statistics are

$$N_i = \sum_{j=1}^{H_i} L_{ij} \quad (10)$$

$$\overline{\mathbf{F}}_i = \sum_{j=1}^{H_i} \overline{\mathbf{F}}_{ij} \quad (11)$$

where H_i are the number of conversations and N_i the total number recordings of speaker i.

3.2.2. *Conditional posterior of* $\mathbf{X}_i$

The conditional posterior of channel factors $\mathbf{X}_i$ is a product of Gaussians:

$$P\left(\mathbf{X}_i|\mathbf{y}_i, \mathbf{\Phi}_i, \mathcal{M}\right) = \prod_{j=1}^{H_i} \mathcal{N}\left(\mathbf{x}_{ij}|\overline{\mathbf{x}}_{ij}, \mathbf{L}_{\mathbf{x}_{ij}}^{-1}\right) \quad (12)$$

where

$$\zeta_{ij} = \tilde{\zeta}_{ij} - L_{ij}\mathbf{J}\mathbf{y}_i \quad (13)$$

$$\tilde{\zeta}_{ij} = \mathbf{U}^T\mathbf{W}\overline{\mathbf{F}}_{ij} \quad (14)$$

$$\mathbf{J} = \mathbf{U}^T\mathbf{W}\mathbf{V} \quad (15)$$

$$\mathbf{L}_{\mathbf{x}_{ij}} = \mathbf{I} + L_{ij}\mathbf{U}^T\mathbf{W}\mathbf{U} \quad (16)$$

$$\overline{\mathbf{x}}_{ij} = \mathbf{L}_{\mathbf{x}_{ij}}^{-1}\zeta_{ij} \quad (17)$$

3.2.3. *Posterior of* $\mathbf{y}_i$

The marginal posterior of speaker factors $\mathbf{y}_i$ is also Gaussian

$$P\left(\mathbf{y}_i|\mathbf{\Phi}_i, \mathcal{M}\right) = \mathcal{N}\left(\mathbf{y}_i|\overline{\mathbf{y}}_i, \mathbf{L}_{\mathbf{y}_i}^{-1}\right) \quad (18)$$

where

$$\mathbf{L}_{\mathbf{y}_i} = \mathbf{I} + N_i\mathbf{V}^T\mathbf{W}\mathbf{V} - \sum_{j=1}^{H_i} L_{ij}^2\mathbf{J}^T\mathbf{L}_{\mathbf{x}_{ij}}^{-1}\mathbf{J} \quad (19)$$

$$\gamma_i = \tilde{\gamma}_i - \sum_{j=1}^{H_i} L_{ij}\mathbf{J}^T\mathbf{L}_{\mathbf{x}_{ij}}^{-1}\tilde{\zeta}_{ij} \quad (20)$$

$$\tilde{\gamma}_i = \mathbf{V}^T\mathbf{W}\overline{\mathbf{F}}_i \quad (21)$$

$$\overline{\mathbf{y}}_i = \mathbf{L}_{\mathbf{y}_i}^{-1}\gamma_i \quad (22)$$

Note that if all the conversations were recorded over only one channel ($L_{ij} = 1$ for all), we would obtain the same equations as for the standard PLDA [11].

The equations for the M-step are the same as for the standard PLDA but using equations (12) and (18) to computed the needed expectations.

4. Experiments

4.1. Development and evaluation datasets

We carried out experiments on the NIST SRE12 core condition [13]. This evaluation presented significant differences compared with previous ones. In previous evaluations, the enrollment data was released at evaluation time. In NIST SRE12, most of the target speakers were taken from evaluations SRE06 to SRE10. For enrolling each speaker, it was allowed to use all the data available. Thus, the core condition, instead of posing a *1 side against 1 side* scenario, represents a *N sides against 1 side* scenario. Besides, it was allowed to use the target speakers data for development (training of UBM, i-vector extractor and calibration, score normalization, etc.), that was also different from previous evaluations. Furthermore, the evaluation proposed new challenges like speech with artificially added noise, speech collected in noisy environments and segments of different duration.

We created our development dataset for the evaluation taking those differences into account. The dataset was divided into two parts:

- Training: This part includes all the signals from SRE04, SRE05, SRE06 and 70% of the signals of SRE08 and SRE10. We used it to train the UBM, JFA, and PLDA models. Besides, the segments in SRE08 and SRE10 parts were used to enroll the target speakers.
- Test: We reserved a 30% of the speech in SRE08 and SRE10 to create a test set for training calibration and evaluating our system. It includes short telephone calls, short and long interviews and 10 seconds calls.

The segments excerpted from the same phonecall or interview (same ldc-id) were assigned to the training part or to the test part but not to both.

Both parts of the dataset were augmented adding Babble and HVAC[1] noises of 15 and 6 dB of signal-to-noise ratio following NIST SRE12 guidelines. The Babble noises were created averaging 1000 conversations from previous evaluations. Different noise samples were added to the training and test datasets. To add the noise, the power of the noise and speech signals was estimated with a psophometric filter and a VAD. The noise added to telephone segments was filtered by a simulated telephone channel.

Adding noisy versions, our training set includes 66457 male and 87826 female segments from 982 male and 1372 female speakers.

The enrollment lists include all the telephone and interview segments of the SRE12 target speakers without noisy versions.

4.2. Speaker recognition system configuration

As features, we used 20 short-time Gaussianized MFCC with deltas and double deltas. We trained full covariance, gender dependent UBM with 2048 components. We used a 600 dimension i-vector extractor. Both UBM and i-vector extractor were trained on telephone data from our development dataset without added noise.

We reduced the i-vector dimensionality to 400 using PLDA. That has the side effect of centering and whitening the i-vectors [4]. Then, we applied i-vector length normalization [14]. Finally, we evaluated the trials applying the standard PLDA or the proposed PLDA (PLDA_2CHT). Both PLDA, the one used for dimensionality reduction and the one used for classification, were trained on telephone and microphone data augmented with noise.

To score the trials, we compared three strategies: standard, i-vector averaging (ivavg) and i-vector statistics scaling (ivsscal). Given N enrollment i-vectors $\mathbf{\Phi}_{trn}$ of a target speaker and a test i-vector ϕ_{tst}; the standard scoring consists of computing the likelihood ratio between the probability that all the i-vectors belong to the same speaker and the probability that $\mathbf{\Phi}_{trn}$ belong to one speaker and ϕ_{tst} to another. This is, theoretically, the correct way of scoring the trial. Because of that, it is also called *by the book* or N *against* 1 scoring. It can be shown that the likelihood ratio can be computed as [15]:

$$R\left(\mathbf{\Phi}_{trn},\phi_{tst}\right) = \frac{P\left(\mathbf{\Phi}_{trn},\phi_{tst}|\mathcal{T}\right)}{P\left(\mathbf{\Phi}_{trn},\phi_{tst}|\mathcal{N}\right)} \tag{23}$$

$$= \left.\frac{P\left(\mathbf{y}_0|\mathbf{\Phi}_{trn}\right)P\left(\mathbf{y}_0|\phi_{tst}\right)}{P\left(\mathbf{y}_0\right)P\left(\mathbf{y}_0|\mathbf{\Phi}_{trn},\phi_{tst}\right)}\right|_{\mathbf{y}_0=\mathbf{0}} \tag{24}$$

where we plug-in the standard PLDA posterior $P\left(\mathbf{y}|\mathbf{\Phi}\right)$ or the PLDA_2CHT posterior given in equation (18). i-vector averaging consist of averaging the enrollment i-vectors of the target speaker and computing the likelihood ratio in a *1 against 1* fashion.

i-vector averaging proved superior performance in the systems submitted to NIST 2012 evaluation. The success of i-vector averaging could be explained because considering many enrollment i-vectors, somehow, overfits the estimation of $P\left(\mathbf{y}|\mathbf{\Phi}_{trn}\right)$, that is, produces a posterior of $\mathbf{y}$ with a too small covariance. On the contrary, having only one enrollment i-vector makes the posterior wider. Another explanation could be that having a different number of enrollment i-vectors for each target speaker, the scores produced by PLDA are in a different range of values whereas, having only one enrollment i-vector per speaker produces better aligned scores.

To combine the strengths of i-vector averaging and the PLDA with two variability terms, we propose to scale the sufficient statistics used to compute $P\left(\mathbf{y}|\mathbf{\Phi}\right)$ like this:

$$\overline{\mathbf{F}}'_{ij} = \frac{\overline{\mathbf{F}}_{ij}}{H_i L_{ij}} \tag{25}$$

$$\overline{\mathbf{F}}'_{i} = \sum_{j=1}^{H_i} \overline{\mathbf{F}}'_{ij} \tag{26}$$

$$L'_{ij} = 1/H_i \tag{27}$$

$$N'_{i} = \sum_{j=1}^{H_i} L'_{ij} = 1\ . \tag{28}$$

Doing that, we control the weight of each i-vector in the calculus of the posterior. To be precise, the weight of each conversation is $1/H_i$ and the number of effective i-vectors is $N'_i = 1$, the same as in i-vector averaging. The weight of each i-vector on its corresponding conversation is $1/L_{ij}$.

We did not explicitly calibrate the scores. The Actual DCF is computed with the scores straight out of the PLDA.

4.3. Results

Table 1 shows results on the NIST SRE12 core condition. The common conditions considered in 2012 as primary performance indicators include the following subsets of trials:

- Det1: All trials involving multiple segment training and interview speech in test without added noise in test.
- Det2: All trials involving multiple segment training and phone call speech in test without added noise in test.
- Det3: All trials involving multiple segment training and interview speech with added noise in test.
- Det4: All trials involving multiple segment training and phone call speech with added noise in test.
- Det5: All trials involving multiple segment training and phone call speech intentionally collected in a noisy environment in test.

Results are reported in terms of Equal Error Rate (EER), Minimum Detection Cost Function (MinDCF) and Actual Detection Cost Function (ActDCF) as defined by NIST [13]. The new primary DCF is the average of the classical DCF in two operating points ($P_{\mathcal{T}} = 0.01$ and 0.001).

For clean interviews (det1), standard PLDA is better in terms of EER and PLDA_2CHT in term of minDCF. The versions with i-vector averaging and stats scaling clearly outperform the versions scored *by the book*. For noisy interviews (det3), the PLDA_2CHT presents slightly better performance. However, stats scaling is superior in terms of EER and, standard

[1] We downloaded HVAC noises from Freesound.org

Table 1: EER, minDCF and actDCF of PLDA approaches on the SRE12 core condition.

Cond.	System	EER(%)	MinDCF	ActDCF
Det1	PLDA	5.51	0.350	0.441
	PLDA 2CHT	5.62	0.322	**0.418**
	PLDA ivavg	**4.31**	0.333	1.296
	PLDA 2CHT iv-sscal	4.54	**0.318**	0.527
Det2	PLDA	8.93	0.550	0.604
	PLDA 2CHT	5.97	0.390	0.426
	PLDA ivavg	1.89	0.243	0.321
	PLDA 2CHT iv-sscal	**1.70**	**0.212**	**0.235**
Det3	PLDA	5.32	0.278	**0.345**
	PLDA 2CHT	5.26	**0.275**	0.356
	PLDA ivavg	5.16	0.345	2.162
	PLDA 2CHT iv-sscal	**4.75**	0.312	0.913
Det4	PLDA	10.14	0.632	0.724
	PLDA 2CHT	7.42	0.470	0.545
	PLDA ivavg	2.73	0.273	0.278
	PLDA 2CHT iv-sscal	**2.51**	**0.237**	**0.242**
Det5	PLDA	10.04	0.587	0.660
	PLDA 2CHT	6.85	0.430	0.477
	PLDA ivavg	2.30	0.264	0.490
	PLDA 2CHT iv-sscal	**2.05**	**0.195**	**0.213**

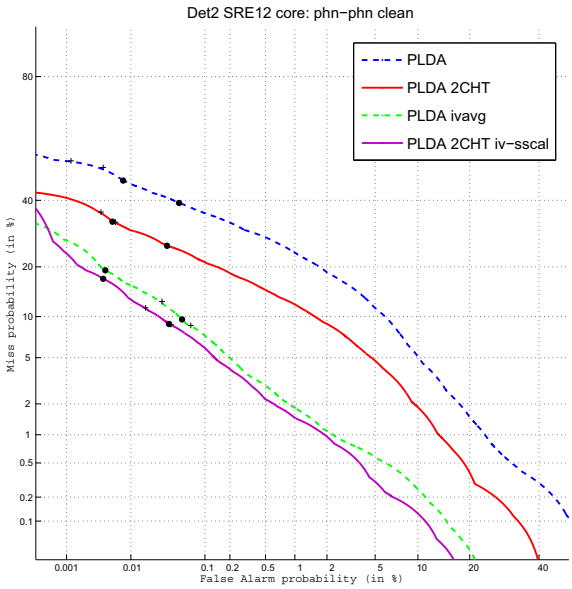

Figure 2: DET curves for the condition with telephone speech without added noise on test (det2).

scoring in terms of minDCF. The standard scoring produces better naturally calibrated scores than i-vector averaging and stats scaling. Nevertheless, that can be solved by a calibration step.

Regarding conditions involving telephone speech in test (det2,4,5), the differences between PLDA and PLDA_2CHT are more significant. Using standard scoring, PLDA_2CHT outperforms the PLDA achieving a relative improvement of 27–33% in terms of EER and 26–29% in terms of minDCF. Furthermore, i-vector averaging and stats scaling clearly outperform the standard scoring. The PLDA_2CH with stats scaling improves by 8–10% in terms of EER and by 12–26% in terms of minDCF with regard to PLDA with i-vector averaging. The PLDA_2CHT with scaling produces very well calibrated scores. Figure 2 shows DET curves [16] for condition det2. The curves prove that the behavior of the systems is consistent over all operating points.

The improvement of the PLDA_2CHT, larger in phonecalls than in interviews can be explained because, cross-channel (interview vs telephone) compensation is not good enough and, for most speakers, there are more enrollment interviews than phonecalls.

5. Conclusions

In this paper, we presented an extension of the standard PLDA model that considers two different terms of inter-session variability (channel terms). This model takes advantage of scenarios that include conversations recorded simultaneously over different channels. To do that, the first channel term is tied between all the recordings belonging to the same conversation while the second is allowed to be different for every recording. Thus, we intend that the former captures the variability between conversations, mainly phonetic variability, and the latter, the variability between channels.

The approach was tested on the core condition of the recent NIST 2012 speaker recognition evaluation. In this evaluation, we count with interviews recorded simultaneously over several channels to train the PLDA and to enroll the target speakers. For conditions with interview speech on test, the differences between the approaches evaluated were not very significant. However, the proposed PLDA achieved a clear gain compared to standard PLDA on phonecalls. The minDCF improves by around 27% if we compare both PLDA scored *by the book* and by around 17% if we use i-vector averaging and stats scaling scorings.

6. Acknowledgments

The work of GTC is supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000. The work of GTTS is supported by the University of the Basque Country under grant GIU10/18 and Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country. We would like to thank Brno University of Technology for hosting the 2012 Bosaris workshop where this work began.

7. References

[1] 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 *Interspeech 2007*, Antwerp (Belgium), Aug. 2007.

[2] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards Noise-Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis," in *International Conference on Acoustics, Speech and Signal Processing ICASSP 2012*, Kyoto (Japan), Mar. 2012, pp. 4253–4256.

[3] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition Training of Gaussian PLDA Models in i-Vector Space for Noise and Reverberation Robust Speaker Recognition," in *International Conference on Acoustics, Speech and Signal Processing ICASSP 2012*, Kyoto (Japan), Mar. 2012, pp. 4257–4260.

[4] J. Villalba and E. Lleida, "Handling i-Vectors from Different Recording Conditions Using Multi-Channel Simplified PLDA in Speaker Recognition," in *International Conference on Acoustics, Speech and Signal Processing ICASSP 2013*, Vancouver (Canada), May 2013.

[5] Q. L. Q. Lin, E.-E. J. E.-E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *Ieee Transactions On Speech And Audio Processing*, vol. 2, no. 4, 1994.

[6] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *1997 IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2. IEEE Comput. Soc. Press, 1997, pp. 1107–1110. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=596135

[7] I. A. Mccowan, J. Pelecanos, and S. Sridharan, "Robust Speaker Recognition using Microphone Arrays," in *Odyssey Speaker and Language Recognition Workshop*, no. 1, Crete (Greece), 2001.

[8] J. W. Stokes, J. C. Platt, and S. Basu, "Speaker Identification using a Microphone Array and a Joint HMM with Speech Spectrum and Angle of Arrival," in *2006 IEEE International Conference on Multimedia and Expo*, 2006.

[9] L. Buera, E. Lleida, J. D. Rosas, J. Villalba, A. Miguel, A. Ortega, and O. Saz, "Speaker verification and identification using Phoneme Dependent Multi-Environment Models based LInear Normalization in adverse and dynamic acoustic environments," in *Summer School for Advanced studies on Biometrics for Secure Authentication Multimodality ans System Integration*, 2005.

[10] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," *IEEE International Conference on Computer Vision*, no. iii, pp. 1–8, 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4409052

[11] N. Brummer, "EM for Probabilistic LDA," Agnitio Research, Cape Town (South Africa), Tech. Rep. February, Feb. 2010. [Online]. Available: https://sites.google.com/site/nikobrummer/EMforPLDA.pdf

[12] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[13] "The NIST Year 2012 Speaker Recognition Evaluation Plan," NIST, Tech. Rep., 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Interspeech 2011*, Florence, 2011, pp. 249–252.

[15] N. Brummer and E. De Villiers, "The Speaker Partitioning Problem," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[16] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Fifth European Conference on Speech Communication and Technology*, vol. 97, ISCA. Citeseer, 1997, pp. 1895–1898. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.4489\&rep=rep1\&type=pdf