Open access • Proceedings Article • DOI:10.1109/ICASSP.2013.6639156

Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances — Source link

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

Institutions: Agency for Science, Technology and Research

Published on: 26 May 2013 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Speaker diarisation and Speaker recognition

PHONETICALLY-CONSTRAINED PLDA MODELING FOR TEXT-DEPENDENT SPEAKER VERIFICATION WITH MULTIPLE SHORT UTTERANCES

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

PHONETICALLY-CONSTRAINED PLDA MODELING FOR TEXT-DEPENDENT SPEAKER VERIFICATION WITH MULTIPLE SHORT UTTERANCES

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li
Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
{*alarcher,kalee,mabin,hli*}*@i2r.a-star.edu.sg*

ABSTRACT

The importance of phonetic variability for short duration speaker verification is widely acknowledged. This paper assesses the performance of Probabilistic Linear Discriminant Analysis (PLDA) and *i*-vector normalization for a text-dependent verification task. We show that using a class definition based on both speaker and phonetic content significantly improves the performance of a state-of-the-art system. We also compare four models for computing the verification scores using multiple enrollment utterances and show that using PLDA intrinsic scoring obtains the best performance in this context. This study suggests that such scoring regime remains to be optimized.

Index Terms— Speaker verification, Text-Dependent, *i*-vector , short duration, PLDA

1. INTRODUCTION

Text-dependent speaker verification is the task of authenticating a person based on his/her voice within a constrained "phonetic context". The phonetic constraint is imposed by requiring the speaker to pronounce certain pass-phrases, fixed or randomly generated during the authentication [1]. Text-dependent speaker verification offers several advantages. First by constraining the phonetic content of the test utterances, text-dependent speaker verification reaches higher accuracy than its text-independent counterpart; especially in dealing with short-duration utterances [2, 3]. Secondly, in cases whereby the users are assigned or are allowed to choose their personalized pass-phrases, security is reinforced as both the pass-phrase and the voice of the user have to match in order to be authenticate. In this work, we consider that users are free to choose their own pass-phrase shorter than 3 seconds.

System based on *i*-vectors [4] and Probabilistic Linear Discriminant Analysis (PLDA) are among the state-of-the-art in text-independent speaker verification. Recently, *i*-vectors have been used for text-dependent verification [5] and we show in [6] that they can take advantage of the phonetic constraint imposed on short duration utterances. Indeed, contrary to text-independent speaker verification task which aims at inferring speaker identity regardless of the text pronounced, text-dependent speaker verification takes into account both speaker characteristic and the lexical content of the test utterances. Therefore, we show in [6] that normalizing *i*-vectors according to this additional knowledge leads to significant improvement in term of accuracy in all conditions. In this work we extend the use of class definitions that include speaker and phonetic information to PLDA training. To the best of our knowledge, no work related to PLDA in the context of text-dependent verification has been reported.

One advantage of PLDA is to natively takes into account the fact that multiple enrollment *i*-vectors are generated by the same speaker [7]. Focusing on short duration, it would be easy during the enrollment phase to require several recordings from the client speaker. Typically, three occurrences of a pass-phrase would keep the recording duration below 10 seconds and improve the accuracy. Other ways of computing a verification score using multiple enrollment sessions for a speaker may be consider and we propose in this work to compare some of those scorings with the original scoring of the PLDA.

This paper is organized as follows. The speaker verification engine and the experimental set-up are described in Section 2 and Section 3 respectively. The results and analysis relative to the definition of training classes is given in Section 4 while comparison of the different verification scores is presented in Section 5. Perspective of this work are discussed in Section 6.

2. I-VECTOR BASED SPEAKER VERIFICATION

2.1. The Total Variability Paradigm

An *i*-vector is the compact representation of a variable-duration recording in a low-dimensional space called Total Variability space [4]. The *i*-vector $\boldsymbol{x}_{(n,s)}$ of the s^{th} session of the n^{th} speaker results from a probabilistic projection of a higher-dimensional vector $\mathbf{m}_{(n,s)}$ onto the total variability space spanned by the columns of the Matrix $\mathbf{T}$, as given by

$$\mathbf{m}_{(n,s)} = \mathcal{M} + \mathbf{T}\boldsymbol{x}_{(n,s)} \tag{1}$$

where $\mathbf{m}_{(n,s)}$ and $\mathcal{M}$ are the mean super-vectors of the speaker- and session-dependent Gaussian Mixture Model and the Universal Background Model respectively.

2.2. Spherical Nuisance Normalization

Most of the discriminative algorithms used in speaker verification are based on the assumption that observations follow a normal distribution which is not the case for *i*-vectors in practice [8]. Thus, several normalization techniques have been proposed to make the *i*-vector distribution closer to the Gaussian assumption [9, 10]. Amongst those techniques, Spherical Nuisance Normalization (SphNorm) proposed in [11] has been shown to especially improve performance of PLDA-based systems. This iterative algorithm estimates the within-class covariance matrix, $\mathbf{W}$ and mean $\boldsymbol{\mu}$ of a large background set of *i*-vectors to transform an *i*-vector $\boldsymbol{x}$ according to:

$$\boldsymbol{x} \leftarrow \frac{\mathbf{W}^{-\frac{1}{2}}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}{\left\|\mathbf{W}^{-\frac{1}{2}}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\|} \tag{2}$$

This transformation first applied to the test *i*-vectors is then used on the background set itself before re-estimating both $\mathbf{W}$ and $\boldsymbol{\mu}$ to perform another iteration.

2.3. Probabilistic Linear Discriminant Analysis

The generative modeling of Probabilistic Linear Discriminant Analysis (PLDA) [7] assumes that observations from a same speaker lie in a similar part of the space. In this work we use a simplified version of PLDA, described by Equation 3, where $\boldsymbol{\mu}$ is the overall mean of the data and $\mathbf{F}$ is a low rank matrix which column vectors form a basis of a subspace of the Total Variability space. The hidden variable $\mathbf{h}_n$ can be seen as the coordinates of the n^{th} speaker in this subspace and $\boldsymbol{\epsilon}$ is a normally distributed additive noise of full covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{x}_{(n,s)} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_n + \boldsymbol{\epsilon} \tag{3}$$

Details about the implementation can be found in [12].

2.4. Verification Score Computation

In this work, we consider a verification task with multiple enrollment utterances. Given a set of L enrollment i-vectors $\{\mathbf{x}_l\}_{l\in[1,L]}$ generated by a target speaker and a test i-vector $\mathbf{x}_{L+1}$, a verification score can be computed as the likelihood ratio of two hypotheses H_1 and H_0. Hypothesis H_1 assumes that the set of i-vectors $\{\mathbf{x}_l\}_{l\in[1,L]}$ and $\mathbf{x}_{L+1}$ have been generated by the client while H_0 assumes $\mathbf{x}_{L+1}$ has been generated by an impostor different from the client. This section describes four different manners of computing the verification score based on the i-vector/PLDA framework.

MS-LLR: PLDA intrinsically provides an elegant way of computing a multiple-segment log-likelihood ratio (MS-LLR) between hypotheses H_1 and H_0. Models for these two hypotheses are shown in Figure 1.a. In the upper part, $\{\mathbf{x}_l\}_{l\in[1,L]}$ and $\mathbf{x}_{L+1}$ come from the same speaker and thus share the same speaker-specific latent variable $\mathbf{h}_{12}$ while in the lower part, under hypothesis H_0, $\{\mathbf{x}_l\}_{l\in[1,L]}$ and $\mathbf{x}_{L+1}$ inherit from different latent variables $\mathbf{h}_1$ and $\mathbf{h}_2$. In [12], it was shown that the verification score $\mathcal{S}$ can be written as:

$$S = \frac{1}{2}\Bigg[\Big(\sum_{l=1}^{L+1}\boldsymbol{x}_l^t\Big)\mathbf{K}_{L+1}\Big(\sum_{l=1}^{L+1}\boldsymbol{x}_l\Big) - \Big(\sum_{l=1}^{L}\boldsymbol{x}_l^t\Big)\mathbf{K}_{L}\Big(\sum_{l=1}^{L}\boldsymbol{x}_l\Big) - \boldsymbol{x}_{L+1}^t\mathbf{K}_1\boldsymbol{x}_{L+1}\Bigg] + \alpha(L) \tag{4}$$

where $\mathbf{K}_L$ and $\alpha(L)$ are defined as follows.

$$\mathbf{K}_L = \boldsymbol{\Sigma}\mathbf{F}(L\cdot\mathbf{F}^t\boldsymbol{\Sigma}\mathbf{F}+\mathbf{I})^{-1}\mathbf{F}^t\boldsymbol{\Sigma} \tag{5}$$

$$\alpha(L) = log\left[\frac{det((L+1)\cdot\mathbf{F}^t\boldsymbol{\Sigma}\mathbf{F}+\mathbf{I})^{-1}}{det(L\cdot\mathbf{F}^t\boldsymbol{\Sigma}\mathbf{F}+\mathbf{I})^{-1}\cdot det(\mathbf{F}^t\boldsymbol{\Sigma}\mathbf{F}+\mathbf{I})^{-1}}\right] \tag{6}$$

FUSION: a second verification score is computed by considering that all enrollment segments are statistically independent. Thus, the verification score is given by the sum of log likelihood ratios computed for each enrollment segment separately. This scoring is described by the model of Figure 1.b and can be seen as a fusion of scores [13]. An expression of this model is given below.

$$S = \frac{1}{2L}\sum_{l=1}^{L}\Bigg[\Big(\boldsymbol{x}_l^t+\boldsymbol{x}_{L+1}^t\Big)\mathbf{K}_2\Big(\boldsymbol{x}_l+\boldsymbol{x}_{L+1}\Big) - \boldsymbol{x}_l^t\mathbf{K}_1\boldsymbol{x}_l - \boldsymbol{x}_{L+1}^t\mathbf{K}_1\boldsymbol{x}_{L+1}\Bigg] + \alpha(1) \tag{7}$$

UNIQUE-IV: statistics from all enrollment sessions of a speaker are accumulated and used to extract a unique i-vector $\tilde{\boldsymbol{x}}$ that could then be used in the PLDA framework Under this configuration, hypotheses H_1 and H_0 correspond to the models given in Figure 1.c and expression of the verification score is similar to equation 4 for $L = 1$.

MEAN-IV: another expression of the verification score follows the model given in Figure 1.c except that the enrollment i-vector $\tilde{\boldsymbol{x}}$ is now the mean of all enrollment i-vectors from the target speaker such that $\tilde{\boldsymbol{x}} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}_l$. Note that the random variable $\tilde{\boldsymbol{x}}$ is the sample mean of $\boldsymbol{x}$ and thus only follows a Gaussian assumption if the number of sample L is fixed. However, even in this specific case, the variances of $\tilde{\boldsymbol{x}}$ is smaller than the variance of $\boldsymbol{x}$ modeled by the PLDA. Under this assumption, the verification score is computed as,

$$S = \frac{1}{2}\Bigg[\frac{1}{L^2}\Big(\sum_{l=1}^{L}\boldsymbol{x}_l^t + L\boldsymbol{x}_{L+1}^t\Big)\mathbf{K}_2\Big(\sum_{l=1}^{L}\boldsymbol{x}_l + L\boldsymbol{x}_{L+1}\Big) - \frac{1}{L^2}\Big(\sum_{l=1}^{L}\boldsymbol{x}_l^t\Big)\mathbf{K}_1\Big(\sum_{l=1}^{L}\boldsymbol{x}_l\Big) - \boldsymbol{x}_{L+1}^t\mathbf{K}_1\boldsymbol{x}_{L+1}\Bigg] + \alpha(1) \tag{8}$$

Comparing (4), (7) and (8) it appears that the verification scores are very similar and mainly differ by the weights given to the enrollment and the test i-vectors at different terms. Another difference comes from the fact that the first two terms of the score fusion expression, does not involve cross terms between the enrollment segments accordingly to the independence assumption. Notice that, the four models are identical for the case of $L = 1$.

3. EXPERIMENTAL SET-UP

3.1. Protocol

Experiments are reported on the male Part 1 of the RSR2015 database, a publicly available speech corpus for text-dependent speaker recognition recorded at Institute for Infocomm Research, A*STAR, Singapore [14]. RSR2015 contains audio recordings from 300 speakers, 143 female and 157 male in 9 sessions each, with a total of 151 hours of audio. In all 9 sessions of RSR2015 Part 1, each speaker pronounces 30 sentences from the TIMIT database [15] covering all English phones. Average duration of recordings over all speakers and sessions is 3.2 seconds

A set of 100 speakers pronouncing the first 15 sentences of this part is reserved for background training of the PLDA and SphNorm parameters totalizing 13,475 utterances. The remaining 15 sentences and 57 speakers are used as test set. This ensures that sentences and speakers used for testing do not overlap the background data. From the 9 recording sessions available for each of the 57 speakers, 3 are used for enrollment and 6 as test segments. Three enrollment conditions are defined using 1, 2 and 3 occurrences of a same pass-phrase. In these conditions, each speaker is used to create 45, 45 and 15 different models respectively. When considering two occurrences for the enrollment, all 3 possible pairs of i-vectors are considered for each speaker and pass-phrase. A trial involves comparison of one model with an i-vector extracted from a test-segment. All cross-pairs between models and test-segments made available with the 57 speakers of the test set are used.

Contrary to text-independent speaker verification systems which only have to differentiate between the correct user (CLIENT) and any other impostor (IMP), text-dependent systems have to consider whether the text pronounced during the test is the same as the one

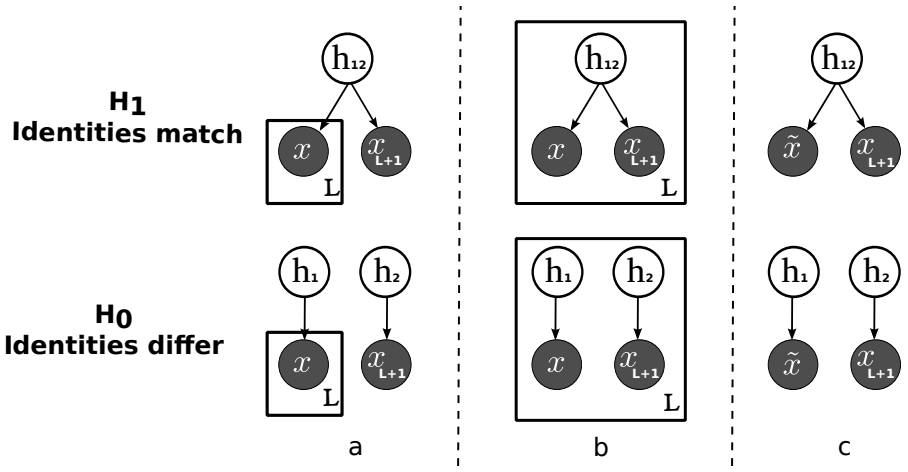

Fig. 1: Graphical models of verification task for different definitions of verification score. Hypothesis H_1 states that enrollment and test data are from the same speaker and share the same latent variable $\mathbf{h}_{12}$ while the alternative hypothesis H_0 states that they are from different speakers and thus inherit from different latent variables $\mathbf{h}_1$ and $\mathbf{h}_2$.

used for enrollment or a different one. Table 2 shows the different

Table 1: The different types of trials defined for text-dependent speaker verification. For each type of trial, the number of tests is given when considering (1/2/3) occurrences of the enrollment pass-phrase respectively.

	Same phonetic content	Different phonetic content
Target speaker	CLIENT-same (15,348 / 15,348 / 5,116)	CLIENT-diff (214,872 / 214,872 / 71,624)
Impostor	IMP-same (859,488 / 859,488 / 286,496)	

types of trials considered in this work. In the rest of this paper, we consider client pronouncing the same pass-phrase (CLIENT-same) as target trials and only focus on two types of impostures, namely CLIENT-diff and IMP-same as we observed in [6] that impostors pronouncing a different sentence are easier to reject and thus will not be considered in this work.

3.2. Configuration

Due to the limited amount of data recorded at 16kHz available for our system development, all data from RSR2015 have been down-sampled to 8kHz. The *i*-vector system uses acoustic features of 13 PLP coefficients with their first and second derivatives. RASTA filtering, VAD detection [16], CMS and Gaussianization were applied. A gender dependent 2048-distribution Universal Background Model is trained on NIST-SRE 2004 and 2005 data. The Total Variability matrix of rank 400 is trained using 28,727 session from NIST-SRE 2004, 2005, 2006, SwitchBoard II phases 2 and 3 and Fisher English parts 1 and 2 male speakers. The *i*-vector extractor and *i*-vector normalization are realized using the ALIZE toolkit [17]. When applied, 3 iterations of Spherical Nuisance normalization are used.

4. CLASS DEFINITION FOR PLDA AND SPHNORM

Here we propose to compare two definitions of classes used for PLDA training and *i*-vector normalization. In the first configuration we consider that each speaker defined exactly one class. This definition, similar to the one used for text-independent speaker verification is referred as *Spk*. For the second configuration, referred as *Spk+Phon*, each class takes advantage of the phonetic content and is defined by a unique couple speaker/pass-phrase. The background set of data used for Spherical Nuisance Normalization and PLDA training includes 100 speakers pronouncing 15 sentences which gives 100 classes for the *Spk* configuration and 1500 classes for *Spk+Phon*. It is worth mentioning that the 15 sentences used for PLDA training are not the same as those used for the verification tests.

For each configuration, two experiments are performed by normalizing or not the *i*-vectors with SphNorm. Results of the four experiments are given in Table 2. Observation of rows 1 and 3 of Table 2 shows that training the PLDA on the *spk+Phn* instead of *spk* classes provides a substantial improvement as the Equal Errors Rate (EER) drops by 87% and 30% relative when considering CLIENT-diff (from 16.58% to 3.68%) and IMP-same (from 10.35% to 7.20%) trials respectively. Moreover, PLDA model trained on *spk+Phn* classes reject CLIENT-diff better that IMP-same as EER become respectively 3.68% and 7.20% when the tendency is the opposite for training on *spk+Phn* classes. According to our previous observation in [6], this result suggests that phonetic variability is more important than speaker information for the short pass-phrases of RSR2015 database.

Table 2: Performance of the PLDA system for *spk* and *spk+Phn* definitions of classes with and without Spherical Nuisance Normalization. The results are given in percentage of EER for two types of non-target trials: client pronouncing a pass-phrase different from the enrollment one (CLIENT-diff) and impostor speakers pronouncing the enrollment pass-phrase (IMP-same). Enrollment using only one occurrence of the pass-phrase.

Configuration	Normalization	CLIENT-diff	Imp-same
Spk	-	16.58	10.35
	SphNorm	15.15	9.06
Spk+Phon	-	3.68	7.20
	SphNorm	3.44	6.96

Rows 2 and 4 of Table 2 show that SphNorm consistently improves the performance across class definitions and for the two types of impostures. The effect of SphNorm appears to be slightly less for *Spk+Phon* class definition than for *Spk*. This may be explain by the fact that PLDA modeling already takes advantage of the phonetic information.

5. VERIFICATION SCORES FOR MULTIPLE ENROLLMENT SEGMENTS

This section compares the four models, as presented in Section 2.4, for computing the verification scores involving multiple uterrances of enrollment in PLDA framework. Performance of the different scorings are given in Table 3 for the two types of impostures considered in this works and for 1,2 and 3 enrollment pass-phrases.

As expected, increasing the number of enrollment pass-phrases reduces the error rate in all conditions and for the four verification scores. The native scoring from PLDA, MS-LLR, reaches the lowest EER in all configurations. Surprisingly, MEAN-IV scoring obtains similar performance despite the fact that the PLDA model is not adapted to the distribution of sample-mean *i*-vectors and that we expect to loose information by taking the mean of the observations. Considering FUSION score for verification also leads to similar performance when dealing with CLIENT speakers pronouncing the correct pass-phrase (CLIENT-diff). However, EER are slightly higher for IMP-same impostures, i.e. 4.64% against 4.09% and 3.82% against 3.27% for 2 and 3 enrollment pass-phrases respectively. Finally, extracting a unique *i*-vector by accumulating statistics over all the enrollment sessions doesn't perform as well as the other verification scores in any condition except when dealing with CLIENT-diff and 3 enrollment pass-phrases. The bad performance of this scoring is probably due to the fact that information about the between-utterance variability is lost when accumulating the statistics across all uterrances. Similar performances obtained for MS-LLR, MEAN-IV and FUSION suggest that PLDA's native scoring is not optimal. As mentioned previously, the scoring described in (4), (7) and (8) show some similarities and share the same characteristic, i.e. all enrollment segments are considered as equally important. Thus PLDA's native scoring does not take any benefit from the multiple enrollment segment as it is suppose to do.

Table 3: Performance of PLDA system for different verification score definitions. Results are given in term of Equal Error Rate (%) for different numbers of enrollment sessions for non-target trials considering client speakers pronouncing a pass-phrase different from the enrollment one (CLIENT-diff) and impostor speakers pronouncing the same pass-phrase (IMP-same). All *i*-vectors are normalized using 3 iterations of Spherical Nuisance Normalization.

Impostor	Verification Scores	Number of Enrollment *i*-vectors		
		1	2	3
CLIENT-diff	MS-LLR	3.44	**1.67**	**1.35**
	MEAN-IV		1.82	1.36
	FUSION		1.91	1.50
	UNIQUE-IV		2.30	1.36
IMP-same	MS-LLR	6.96	**4.09**	**3.27**
	MEAN-IV		4.15	3.35
	FUSION		4.64	3.82
	UNIQUE-IV		6.20	4.50

Indeed, all these three score definitions consider all enrollment segments as equally important, which seems highly sub-optimal. Future work will focus on including prior knowledge about the different enrollment segments into the PLDA scoring in order to increase the benefit of multiple enrollment utterances. Further experiments following this work show that the behavior of the proposed scorings vary depending on the number of enrollment utterances available raising the question of the balance between enrollment and test segments in the PLDA scoring.

6. CONCLUSION

This work assesses the effect of adding phonetic information to speaker classes into PLDA training for text-dependent speaker verification. Substantial improvements are reported for two types of impostures as the EER is reduced by 87% relative when considering the client speaker pronouncing a wrong pass-phrase and by 30% relative in case of attack from an impostor speaker who pronounces the correct pass-phrase. Additional improvement can be brought by normalizing the *i*-vectors with Spherical Nuisance Normalization trained on speaker and phonetic classes.

Comparing four definitions of verification score defined in the PLDA framework we show that, when using multiple enrollment utterances, the native PLDA scoring obtains the best performance. However, the fact that this scoring obtains similar performance as a score based on the mean of enrollment *i*-vectors and a score based on a fusion of scores show that this score definition is not optimal.

7. REFERENCES

[1] Matthieu Hébert, *Text-dependent speaker recognition*, Springer-Verlag, Heidelberg, 2008.

[2] Robert J. Vogt, Christopher J. Lustri, and Sridha Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Odyssey Speaker and Language Recognition Workshop*. 2008, IEEE.

[3] Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, and Michael Mason, "i-vector Based Speaker Recognition on Short Utterances," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011.

[4] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] Hagai Aronowitz, "Text-Dependent Speaker Verification Using a Small Development Set," in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[6] Anthony Larcher, Pierre-Michel Bousquet, Kong Aik Lee, Driss Matrouf, Haizhou Li, and Jean-Francois Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2012.

[7] Simon J.D. Prince, *Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012, In press.

[8] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, 2010.

[9] Pierre-Michel Bousquet, Driss Matrouf, and Jean-Francois Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 485–488.

[10] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 249–252.

[11] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-Francois Bonastre, and Oldrich Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[12] Ye Jiang, Kong Aik Lee, Zhenmin Tang, Bin Ma, Anthony Larcher, and Haizhou Li, "PLDA Modeling in I-vector and Supervector Space for Speaker Verification," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[13] P. Verlinde, G. Chollet, and M. Acheroy, "Multi-modal identity verification using expert fusion," *Information Fusion*, vol. 1, no. 1, pp. 17–33, 2000.

[14] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[15] William M. Fisher, Georges R. Doddington, and Kathleen M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[16] Lori Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.

[17] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve, and John S.D. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Odyssey Speaker and Language Recognition Workshop*, 2008, http://mistral.univ-avignon.fr/.