Text-dependent speaker recognition using PLDA with uncertainty propagation

T. Stafylakis[1,2], P. Kenny[1], P. Ouellet[1], J. Perez[3], M. Kockmann[3] and P. Dumouchel[1,2]

[1]Centre de Recherche Informatique de Montreal (CRIM), Canada,
[2]Ecole de Technologie Superieure (ETS), Canada,
[3]VoiceTrust, Germany

themos.stafylakis@crim.ca

Abstract

In this paper, we apply and enhance the i-vector-PLDA paradigm to text-dependent speaker recognition. Due to its origin in text-independent speaker recognition, this paradigm does not make use of the phonetic content of each utterance. Moreover, the uncertainty in the i-vector estimates should be taken into account in the PLDA model, due to the short duration of the utterances. To bridge this gap, a phrase-dependent PLDA model with uncertainty propagation is introduced. We examined it on the RSR-2015 dataset and we show that despite its low channel variability, improved results over the GMM-UBM model are attained.

Index Terms:Text-dependent speaker recognition, Bayesian Methods, Subspace Learning

1. Introduction

Over the past few years, text-independent speaker recognition has been revolutionized by the introduction of subspace probabilistic modeling. Joint-Factor Analysis and its successor, the combination of i-vectors and Probabilistic Linear Discriminant Analysis (PLDA) have become the dominant approaches in the field, as the NIST and other contests can verify, [1], [2], [3]. The emergence of the i-vector representation of an utterance enabled us to compact a model of about 10^5 free parameters (the concatenated means of a GMM, a.k.a. supervector) into a vector of about $d = 500$ dimensions. The key idea of placing an eigenvoice prior on the supervectors rather than a set of independent priors on each gaussian component (i.e. the GMM-UBM approach, [4]), allowed us to learn the correlations between the supervector dimensions offline, using unlabeled training data. Based on this prior, we are able to make a more sophisticated estimate about the mean of those components that have not been visited for a particular utterance, rather than naively setting them equal to the population's average (i.e. letting those UBM components unchanged), [1]. Moreover, the PLDA on the i-vectors manages the decompose the total variability to speaker and channel variability more effectively compared to JFA, [2], [3].

Despite the success of the i-vector-PLDA paradigm, its applicability in text-dependent speaker recognition remains questionable. By considering the case where enrollment and test phrases are the same, one may argue that there is no need to adopt the eigenvoice approach. Due to the matched phonetic content, by using the GMM-UBM approach, the enrollment utterances will adapt exactly those components with which the likelihood ratio will be evaluated. The current research in the field has shown that conventional approaches are superior to the i-vector-based ones, [5], [6], [7], [8].

Considering the several potentials of the text-dependent setting (e.g. banking systems, voice passwords), we managed to adapt the current i-vector-PLDA paradigm in a way that makes use of the characteristics of the problem, i.e. short utterances from a set of predefined phrases. The method we propose, a phrase-dependent PLDA model with uncertainty propagation, utilizes both the phrase label and the uncertainty in the i-vector estimates and produces improved results over the GMM-UBM and the standard i-vector-PLDA approach on the RSR-2015 dataset. The rest of the paper is organized as follows. In Sect. 2, the proposed modifications to the PLDA model are defined and justified. In Sect. 3, the EM training algorithm is given, followed by way to evaluate it. Finally, experimental results are given and discussed in Sect. 4.

2. The proposed PLDA model

In this chapter, we discuss the PLDA variant that is suited to text-dependent speaker recognition with utterances of variable duration. We assume that the phrase labels are given for all three kinds of utterances, i.e. train, enrollment and test.

2.1. The baseline PLDA model

Let $\{i_r\}_{r=1}^{R}$, $i_r \in \Re^d$ be a set of i-vectors, indexed by $r = 1, \ldots, R$. The PLDA is a generative model that is described by the following equations

$$i_r = \mu + Vy + \epsilon_r \tag{1}$$

where

$$y \sim \mathcal{N}(0, I) \tag{2}$$

a vector of speaker factors,

$$\epsilon_r \sim \mathcal{N}(0, D^{-1}) \tag{3}$$

the residual and D the precision matrix (i.e. inverse covariance). This model is the dominant approach during the last two NIST contests, namely NIST-2010 and NIST-2012.

2.2. Phrase-dependent PLDA model

We assume now that each speaker is constrained to say one or more phrases of a predefined set. Let $l = 1, \ldots, L$ denote the phrase label and a set of utterances with duration typically between 1 and 3 sec, [7]. Due to their short duration, the phonetic variability in such utterances is comparable or greater to the other two types of variability, namely speaker and channel variability, [6]. This is in contrast to the NIST data, where the utterances are of 2 min duration, thus the phonetic variability can be considered as negligible. Therefore, a natural question to ask is how to modify the baseline PLDA model so that it takes into account the phrase labels.

The proposed phrase-dependent PLDA model is described by the following equation

$$i_r = \mu_l + V_l y + \epsilon_r \tag{4}$$

where

$$y \sim \mathcal{N}(0, I) \tag{5}$$

is a vector of speaker factors,

$$\epsilon_r | l \sim \mathcal{N}(0, D_l^{-1}) \tag{6}$$

is the residual and D_l the precision matrix (i.e. inverse covariance).

Contrary to the baseline PLDA model, the new one is defined by a set of L phrase dependent parameters $\mathcal{P} = \{\mathcal{P}_l\}_{l=1}^{L} = \{V_l, \mu_l, D_l\}_{l=1}^{L}$. It assumes that given speaker y and phrase l, y is mapped to a random variable in the i-vector space, through (i) a phrase-dependent affine transformation (μ_l, V_l) that defines its expected value, and (ii) a phrase-dependent full covariance matrix D_l^{-1}. Therefore, we model the distribution of a single speaker as an L-component mixture PLDA model. The covariance matrices of the components are phrase-dependent yet fixed across speakers. Finally, the use of phrase-dependent V_l allows the GMM to vary non-rigidly with y, as it would have happened if a common V was deployed, [9].

2.3. Uncertainty propagation

Despite the high flexibility of the above model that is gained by the use of phrase-dependent parameters, the model still does not take into account the uncertainty in the i-vector estimate. Recall that the i-vector is the (both MAP and posterior-expectation) point-estimate of a random variable that follows a normal distribution $\mathcal{N}(i_r, B_r^{-1})$. The uncertainty in this point-estimate B_r^{-1} is a function completely defined by the UBM covariance matrices and the zero-order statistics of the utterance, [2]. Ignoring the uncertainty in datasets of long utterances (such as NIST up to 2010 contests) was reasonable, since the total variability is largely dominated by speaker and channel variability. However, when dealing with utterances of short or medium size, the estimation covariance is no longer negligible and should therefore be propagated through the PLDA model, [10], [11].

The equation that describes the generative models is as follows

$$i_r = V_l y + U_r x_r + \epsilon_r \tag{7}$$

where

$$x_r \sim \mathcal{N}(0, I) \tag{8}$$

and U_r the lower-triangle Cholesky decomposition of the estimation covariance, i.e. $B_r^{-1} = U_r U_r^t$. According to this setting the distribution of i_r given y and l is as follows

$$i_r | y, l \sim \mathcal{N}(\mu_l + V_l y, D_l^{-1} + B_r^{-1}) \tag{9}$$

Finally, in the more realistic scenario where only an estimate of y is in hand, with expected value $\hat{y}$ and precision matrix P_y, based on a set of enrollment utterances $\{i_n\}_{n=1}^{N}$, the distribution will be as follows

$$i_r | \{i_n\}_{n=1}^{N}, l \sim \mathcal{N}(\mu_l + V_l \hat{y}, V_l P_y^{-1} V_l^t + D_l^{-1} + B_r^{-1}) \tag{10}$$

The above equation is the predictive distribution and will be used for calculating the log-likelihood ratio (LLR).

3. Training and evaluating the model

In this section, we demonstrate how to train and evaluate the proposed model. The training algorithm is an extension of the standard EM algorithm, so that it takes into account (i) the phrase labels of each utterance, and (ii) the uncertainty in the i-vector estimates. Note that the complexity of the algorithm is higher that in the case of standard PLDA model, because the uncertainty propagation does not allow the acceleration tricks to be applied, [10].

Assume we have a training set of $\mathcal{D} = \{i_{s,r}, U_{s,r}, l_{s,r}\}$, where $s = 1, \dots, S$ and $r = 1, \dots, R_s$ denote the speaker and i-vector index, respectively. Let $N_{s,l}$ denote the number of training i-vectors of the lth phrase that belongs to the sth speaker. Also, let $N_{\cdot,l} = \sum_s N_{s,l}$ and $N_{s,\cdot} = \sum_l N_{s,l}$. The mean parameters of the model $\{\mu_l\}_{l=1}^{L}$ are directly calculated and subtracted from the corresponding i-vectors. The mean-normalized i-vectors are denoted by $f_{s,r}$, i.e. $f_{s,r} = i_{s,r} - \mu_r$.

3.1. Training: E-step

In the E-step, we condition on the current estimate of the model parameters and estimate the posterior of the hidden variables, i.e. of the speaker factors $\{y_s\}$ and $\{x_{s,r}\}$. The posterior is decomposed as follows

$$p(y_s, \{x_{s,r}\}_{r=1}^{R_s} | \mathcal{D}) = p(\{x_{s,r}\}_{r=1}^{R_s} | y_s, \mathcal{D}) p(y_s | \mathcal{D}) \tag{11}$$

where the conditioning on the current estimate of $\mathcal{P}$ is implied. We define and precalculate the following matrices, $J_{s,r} = U_{s,r}^t D_{l_{s,r}} V_{l_{s,r}}$ and $K_{s,r} = I + U_{s,r}^t D_{l_{s,r}} U_{s,r}$. The posterior of the speaker factors, namely its posterior expectation $\hat{y}_s$ and precision P_s is calculated using the following formulae

$$P_s \hat{y}_s = \sum_{r=1}^{R_s} \left(V_{l_{s,r}}^t - J_{s,r}^t K_{s,r}^{-1} U_{s,r}^t \right) D_{l_{s,r}} f_{s,r} \tag{12}$$

and

$$P_s = I + \sum_{r=1}^{R_s} V_{l_{s,r}}^t D_{l_{s,r}} V_{l_{s,r}} - J_{l_{s,r}}^t K_{s,r}^{-1} J_{l_{s,r}} \tag{13}$$

from which we obtain $\hat{y}_s$. In the case where the eigenvalues of $U_{s,r} U_{s,r}^t$ are large due to its short duration, the contribution of the rth i-vector in estimating y_s would be small as well. Moreover, the formulae of the PLDA algorithm without uncertainty propagation is a special case where the eigenvalues of all i-vectors are zero, which is only feasible with utterances of infinite duration. Finally, the posterior expectation of $\{x_{s,r}\}_{r=1}^{R_s}$ is calculated as follows

$$K_{s,r} \hat{x}_{s,r} = D_{l_{s,r}} U_{s,r}^t \left(f_{s,r} - V_{l_{s,r}} \hat{y}_s \right) \tag{14}$$

Note that $K_{s,r}$ is the precision matrix of $x_{s,r}$.

3.2. Training: M-step

In the M-step, the posterior expectations and precision matrices are used to update the point-estimates of the model parameters. To do so, we define and calculate the following expressions for each y_s and $x_{s,r}$

$$\left\langle y_s y_s^t \right\rangle = \hat{y}_s \hat{y}_s^t + P_s^{-1} \tag{15}$$

and

$$\left\langle x_{s,r} x_{s,r}^t \right\rangle = K_r^{-1} U_{s,r}^t D_r T_{s,r} D_r U_{s,r} K_r^{-1} + K_r^{-1} \tag{16}$$

where

$$T_{s,r} = f_{s,r}f_{s,r}^t - V_r\hat{y}_s f_{s,r}^t - f_{s,r}\hat{y}_s^t V_r^t + V_r \left\langle y_s y_s^t \right\rangle V_r^t \quad (17)$$

Moreover,

$$\left\langle x_{s,r} y_s^t \right\rangle = K_r^{-1} U_{s,r}^t D_r \left(f_{s,r}\hat{y}_s^t - V_r \left\langle y_s y_s^t \right\rangle \right) \quad (18)$$

Based on these expressions, we calculate the correlation matrices

$$R_{l,yy} = \sum_{s=1}^{S} N_{s,l} \left\langle y_s y_s^t \right\rangle \quad (19)$$

and

$$R_{l,fy} = \sum_{s=1}^{S} \sum_{r:l_r=l} f_{s,r}\hat{y}_s^t - U_{s,r} \left\langle x_{s,r} y_s^t \right\rangle \quad (20)$$

The updating rules are as follows

$$V_l = R_{l,fy} R_{l,yy}^{-1} \quad (21)$$

and

$$D_l^{-1} = \left(\sum_{s=1}^{S} N_{s,l} \right)^{-1} (F_l - V_l R_{l,fy}^t) \quad (22)$$

where

$$F_l = \sum_{s=1}^{S} \sum_{r:l_r=l} f_{s,r}f_{s,r}^t - f_{s,r}\hat{x}_{s,r}^t U_{s,r}^t - U_{s,r}\hat{x}_{s,r}f_{s,r}^t + U_{s,r} \left\langle x_{s,r} x_{s,r}^t \right\rangle U_{s,r}^t \quad (23)$$

3.3. Training: Minimum divergence step

As in the standard PLDA training algorithm, we apply a Minimum divergence (MD) step, so that we enforce the prior of y to be a standard normal one. This can be achieved by transforming the current estimates of $\{V_l\}_{l=1}^L$ in a way so that the covariance of the speaker factors

$$C_{yy} = \frac{1}{S} \sum_{s=1}^{S} \left\langle y_s y_s^t \right\rangle \quad (24)$$

becomes equal to the identity matrix. To do so, the Cholesky decomposition of C_{yy} is calculated, i.e. $H^tH = C_{yy}$ and the transformed $\{V_l\}_{l=1}^L$ are calculated as $V_l \leftarrow V_l H^t$.

3.4. Training: Evaluating the evidence

The log-evidence of the model is a very useful criterion in order to track the convergence and debug the code. After every single E-step, M-step and MD-step, its value should be non-decreasing. It is calculated with the following formula

$$\mathcal{L}(\mathcal{D}) = \frac{1}{2} \sum_{l=1}^{L} \left[N(\cdot, l) \log \left| \frac{1}{2\pi} D_l \right| - \text{tr}(D_l G_l) \right] - \frac{1}{2} W \quad (25)$$

where

$$G_l = \sum_{s=1}^{S} \sum_{r:l_{s,r}=l} f_{s,r} \left(f_{s,r} - U_{s,r}\hat{x}_{s,r} - V_l \hat{y}_s \right)^t \quad (26)$$

and

$$W = \sum_{s=1}^{S} \log |P_s| + \sum_{r=1}^{R_s} \log |K_{s,r}| \quad (27)$$

3.5. Evaluating: Calculating the LLR

To evaluate the model, we assume a set of enrollment utterances from the target speaker s, $\mathcal{D}_s = \{i_{s,r}, U_{s,r}, l_{s,r}\}_{r=1}^{R_s}$ and a test utterance $\{i_t, U_t, l_t\}$. Note that we assume that l_t is given, otherwise we should sum over the phrase-labels. The LLR is formed as follows

$$\text{LLR}_{s,t} = \frac{p(i_t, \{i_{s,r}\}_{r=1}^{R_s})}{p(i_t)p(\{i_{s,r}\}_{r=1}^{R_s})} = \frac{p(i_t | \{i_{s,r}\}_{r=1}^{R_s})}{p(i_t)} \quad (28)$$

where the probability density functions (pdf) are assumed to be conditioned on the model parameters $\mathcal{P}$ and on $\{U_{s,r}, l_{s,r}\}_{r=1}^{R_s}$ and $\{U_t, l_t\}$. To calculate the numerator of the LLR (known as the predictive distribution) we first need to enroll the model, that consists of estimating the first and second order statistics of y_s i.e. $(\hat{y}_s, P_s)$ given in (12) and (13). The numerator is a normal pdf evaluated at i_t with mean and covariance matrix equal to $(\mu_{l_t} + V_{l_t}\hat{y}_s, V_{l_t}P_s^{-1}V_{l_t}^t + D_{l_t}^{-1} + U_tU_t^t)$. The denominator is again a normal pdf evaluated at i_t with parameters $(\mu_{l_t}, V_{l_t}V_{l_t}^t + D_{l_t}^{-1} + U_tU_t^t)$, i.e. with $(\hat{y}_s, P_s)$ set to their prior values $(0, I)$. Note that the expression of the standard-PLDA LLR can be recovered by setting the entries of all U matrices equal to zero.

4. Experiments

4.1. Dataset and experimental set-up

We evaluate the performance of the proposed method using the RSR-2015 speaker recognition dataset, [7]. It consists of 299 speakers, divided into background (bkg), development (dev) and evaluation ($eval$) subsets. We used the part I of the dataset, where each of the speakers is speaking 30 different phrases, into 9 different sessions. We trained a gender-independent PLDA model using bkg and dev subsets, as explained in Sect. 3. At recognition time, a speaker is enrolled with 3 utterances of the same phrase. The test utterance is also of the same phrase, however all utterances in a trial come from different sessions and taken from the $eval$ set.

We used 60-dimensional MFCC with short-term mean and variance normalization, an energy-based voice-activity detector and a gender-independent UBM of 1024-components, trained on all parts of the bkg data. The gender-independent i-vector extractor was trained on the bkg and dev subsets. Length normalization is applied, i.e. $i \leftarrow \sqrt{d}\frac{i}{\|i\|}$, where the multiplicative term $\sqrt{d}$ is placed in order to preserve the scaling of the original i-vector space, [12]. A scaling factor $\rho = 1/3$ is used, to compensate for inter-frame correlation (see [13]). We use diagonal covariance matrices to encode the uncertainty in the i-vector estimates, thus U_r is also diagonal. Despite the use of length normalization, no transform on the covariance matrices is applied so that we preserve its diagonal form. The use of the Jacobian matrix of the length normalization transform seems to be a reasonable choice in order to transform U_r, however we did not attempt it in this paper. For the same reason, no LDA is applied either. Finally, the dimensionality of $x_{s,r}$ and y_s is also d.

4.2. Experimental results

4.2.1. Baseline system

The first experiment is with a baseline 512-component, gender independent GMM-UBM approach. No score normalization has been applied. We present the error metrics, namely Equal Error Rate (EER) and minimum normalized Detection

Cost Function, both NIST-08 and NIST-10 (DCF$_o$ and DCF$_n$, respectively) for UBM-training sets in Table 1. The error rates that the GMM-UBM approach attained are very low, even with an unmatched training set (S_1, trained on CSLU Multilingual Telephone) and constitute a baseline performance that is hard to improve. Two are the main reasons for its success. The first is the matched phonetic content between enrollment and test utterances. In such cases, an independent prior on each component on the UBM (i.e. the GMM-UBM approach) seems to be well justified, since those components that would be adapted during enrollment are the ones that will be activated during testing. The other main reason is the relatively low channel variability of the dataset. In such cases, the front-end normalization techniques that we apply can remove most of the channel variability. Hence, an explicit modeling of channel variability with PLDA might not be required.

Table 1: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, using the UBM-GMM approach for several different training sets. UBM trained on CSLU (S_1) and RSR bkg (S_2).*

	female			male		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
S_1	1.77	0.107	0.396	2.28	0.136	0.472
S_2	1.53	0.089	0.332	2.17	0.125	0.466

4.2.2. PLDA systems

We now examine the performance of the i-vector-PLDA model. We have included the standard phrase-independent PLDA models (M_1), a phrase-dependent PLDA model (M_2) and a phrase-dependent PLDA model with uncertainty propagation (M_3). The results are given in Table 2. A comparison between these two tables shows that the standard PLDA version is inferior to the GMM-UBM. Moreover, the performance of the phrase-dependent PLDA model is comparable to the one the GMM-UBM model attains. It is only after combining phrase-dependence and uncertainty propagation that the i-vector-PLDA paradigm showed its superiority against GMM-UBM.

Table 2: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, using several i-vector-PLDA approach. M_1: phrase-independent model, no UP, M_2: phrase-dependent model, no UP, M_3: phrase-dependent model, with UP. The models starting with C are the corresponding ones, with the inclusion of the eval set in i-vector extractor training. Finally, S_3: GMM-UBM trained on RSR bkg, dev & eval*

	female			male		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
M_1	2.99	0.158	0.544	2.37	0.134	0.528
M_2	2.93	0.134	0.423	1.86	0.098	0.366
M_3	1.71	0.082	0.294	1.21	0.069	0.314
C_1	2.30	0.119	0.423	2.05	0.119	0.495
C_2	2.41	0.112	0.338	1.58	0.080	0.299
C_3	1.31	0.061	0.226	0.93	0.056	0.266
S_3	1.45	0.089	0.324	2.07	0.118	0.433

The number of speaker in bkg+dev is only 194, so we were also interested to know what would be the best performance we could get, in the case where a much greater number of training utterances was in hand for training the i-vector extractor. To do so, we created a cheating experiment, where the i-vector extractor has been trained on RSR bkg, dev and $eval$ sets. The results of the corresponding PLDA approaches are given in Table 2, denoted by $C_i, i = 1, 2, 3$ and show a significant gain in performance. Finally, the same cheating experiment is repeated for the GMM-UBM approach (S_3), with a minor gain. Although this set of results should not be regarded as official, it serves as an indicator for the potential of the proposed method in the case of a sufficiently large training set. Finally, the DET curves for several of the approached discussed are depicted in Fig. 1 for the two genders combined. The results are worse that those in the table due to the gender independent threshold.

The experiments show that a clear improvement can be attained when the proposed model is used, showing the strength of the i-vector-PLDA paradigm also in text-dependent setting. Moreover, considering the low channel variability of the particular dataset and the superiority of PLDA to perform under high channel variability, we assume that the increase in performance should normally be even larger on other datasets.

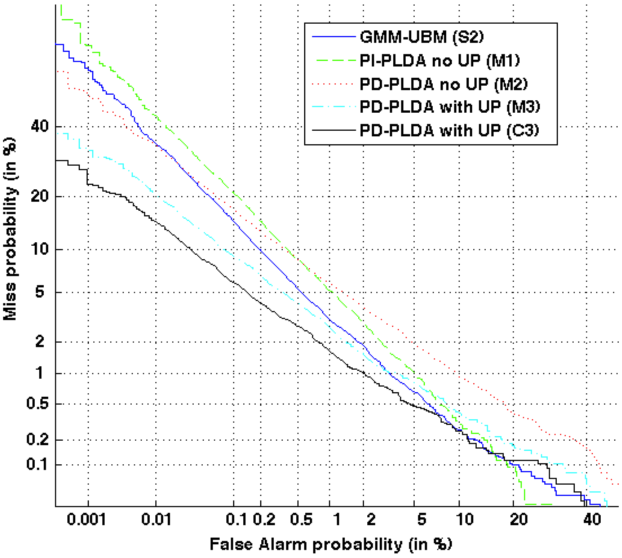

Figure 1: *DET curves on both genders of RSR part I.*

5. Conclusions

In this paper, we examined the use of the i-vector-PLDA paradigm to text-dependent speaker recognition. The field had many potentials since the knowledge of the phonetic content enables systems to deal with much shorter utterances for the same error rates, compared to text-independent. We showed how the phrase label can be used in order to define phrase-dependent model parameters and how the uncertainty in the i-vector estimates can be propagated though the PLDA model. The experimental results showed that the proposed modifications were necessary for the PLDA to perform better than a conventional GMM-UBM, especially for a dataset of low channel variability. Finally, we showed that the performance could be further improved when enough utterances are available for training the i-vector extractor.

6. References

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, A study of inter-speaker variability in speaker verification, in *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, Front-End Factor Analysis for Speaker Verification, in *IEEE Transactions on Audio, Speech & Language Processing*, 2011.

[3] P. Kenny, Bayesian Speaker Verification with Heavy-Tailed Priors, in *Proc. Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models, in *Digital signal processing*, 2000.

[5] H. Aronowitz, Text-Dependent Speaker Verification Using a Small Development Set, in *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Singapore, June 2012.

[6] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Y. Li, J.-F. Bonastre, I-vectors in the context of phonetically-constrained short utterances for speaker verification, in *Proceedings of ICASSP*, Kyoto, Japan, March 2012.

[7] A. Larcher, K.-A. Lee, B. Ma and H. Li, The RSR2015: database for text-dependent speaker verification using multiple pass-phrases, in *Proceedings of Interspeech*, Portland (Oregon), USA, Sept. 2012.

[8] A. Larcher, K.-A. Lee, B. Ma and H. Li, Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances, to appear at *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[9] J. Villalba, E. Lleida, A. Ortega and A. Miguel, I3A SRE12 System Description, on *SRE12 Speaker Recognition Workshop*, Oct. 2012.

[10] P. Kenny, T. Stafylakis, P. Ouellet, M.J. Alam and P. Dumouchel, PLDA for speaker verification with utterances of arbitrary duration, to appear at *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[11] B. J. Borgstrom and A. McCree, Supervector Bayesian speaker comparison, to appear at *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[12] D. Garcia-Romero and C. Y. Espy-Wilso, Analysis of i-vector length normalization in speaker recognition systems, in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[13] T. Stafylakis, P. Kenny, V. Gupta and P. Dumouchel, Compensation for inter-frame correlations in speaker diarization and recognition, to appear at *Proceedings of ICASSP*, Vancouver, Canada, May 2013.