SPEAKER VERIFICATION PERFORMANCE WITH CONSTRAINED DURATIONS

Pablo L Sordo Martínez, Benoît Fauve, Anthony Larcher, John Mason

SPEAKER VERIFICATION PERFORMANCE WITH CONSTRAINED DURATIONS

Pablo L. Sordo Martínez[1], Benoît Fauve[2], Anthony Larcher[3], John S.D. Mason[1]

[1]Speech and Image Research Group, Swansea University, Wales, United Kingdom
[2]ValidSoft Ltd., United Kingdom
[3]Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore

{P.L.SORDO-MARTINEZ.669345,j.s.d.mason}@swan.ac.uk
Benoit.Fauve@validsoft.com
alarcher@i2r.a-star.edu.sg

ABSTRACT

Over the last decade speaker recognition has witnessed significant advances, with successful developments in Factor Analysis (FA) and more recently i-vectors, more than halving the error rates achieved by the classical UBM/GMM approach. However when very short duration utterances are considered, it is known that these improvements are much less. This paper begins with a review of the recent developments of i-vector systems with a focus on short test duration, in the region of 10 s or less. Experimental results are then presented showing that error rates rise from approximately 5% to 18% when the test duration is systematically reduced from 30 s to just 3 s. Interestingly, with the 30 s condition the i-vector error rate is in the region of half that of the corresponding UBM/GMM system. Nevertheless, when the test segments are just 3 seconds duration then the error rates of the 2 systems systems are very similar. All experiments relate to the short-short condition of the NIST 2008 SRE, but with the test duration systematically reduced.

Index Terms— Speaker verification, GMM/UBM, i-vectors, LDA, PLDA, short duration

1. INTRODUCTION

The growth of interest for telephony based authentication in applications such as mobile banking [1] has brought new challenges to speaker verification. User convenience invariably means that systems should operate with just a few seconds of speech, certainly in the authentication (testing) phase. This is in contrast to the majority of speaker recognition research which has tended to focus on much longer durations, although in recent times significant effort has been directed towards short duration testing, including the work of [2, 3].

I-vectors have become the state-of-the-art in speaker recognition since 2010. They have improved accuracy and robustness while simplifying the classification task, bringing it into a low-dimensional space. A considerable number of works tested the performance of i-vectors in short duration conditions [4] and recently some improvements have been made [5, 6]. Given all the existing works, limitation of the speech material is still a challenging constraint.

In this paper, we first present a brief review of some of the latest developments in i-vector based systems when evaluated under short duration conditions. We then present somewhat related results with on a direct comparison with a conventional UBM/GMM as a benchmark against one using i-vectors. The main goal is to compare error rate trends of these arrangements as the quantity of speech data is reduced. Here we keep the enrolment (training) data constant and systematically reduce the amount of verification (test) data.

The paper is structured as follows. Section 2 presents a brief survey of recent developments involving i-vector schemes specifically encompassing short duration conditions while Section 3 follows with an account of the related strategies and techniques. The experimental configuration, in particular the test data shortening process, follows in Section 4, while experimental results are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. I-VECTORS IN A SHORT DURATION CONTEXT

Since its appearance [4], i-vector framework rapidly became the state-of-the-art in speaker recognition. Thanks to the intensive use of development data provided by the NIST and the Linguistic Data Consortium (LDC), i-vector systems outperformed most of the other configuration involved in the last two speaker recognition evaluations (2010 and 2012). As a consequence of this enormous support from these two institutions, i-vectors were initially only exploited in text-independent contexts [7–9].

Recently, i-vectors have proved to be efficient in some other areas such as language recognition [10] and text-dependent speaker verification [11, 12]. Regarding the latter, the inclusion and development of new text-dependent databases [13] has led to a new framework where short du-

ration is a requirement.

During time, the natural mainstream has been to evaluate the robustness of i-vectors against different sources of variability, among which duration mismatch was one of the primary options. Several works [14, 15] proved that duration mismatch between enrolment and test data is less harmful than shortening both sets equally.

In the classical UBM/GMM configuration [16] it has long been known [17] that, when the two speech segments under consideration are of meaningfully different durations then it proves beneficial to generate the GMM on the longer segment and carry out the UBM/GMM scoring on the shorter segment. Of course in the context of the i-vector approach, where two i-vectors are derived, one from test segment GMM and one from training segment GMM, then no such separation exists and the models presented to the classification process are structurally identical (and interchangeable) for the two segments. This leads to the challenging question of how to address the imbalance of data in the i-vector context when one speech segment is meaningfully different in duration to the second segment, see for example the recent work of [5].

3. SPEAKER VERIFICATION SYSTEMS

In the experimental work presented here speaker verification scores for two i-vectors configurations are compared directly with those derived from a conventional UBM/GMM. The latter is now well understood [16, 18] with research and application applied over almost two decades. Thus here we address the much more recent i-vector approach which is then to be directly compared with the UBM/GMM when tested against different speech durations.

3.1. The i-vectors framework

The i-vector paradigm has been motivated by inconsistencies found in the Joint Factor Analysis (JFA) [19] framework. Indeed, it was shown in [4] the the assumption of speaker and session variabilities laying in different subspaces is not true and that the session subspace contains information on the speaker identity.

In this paradigm, it is assumed that a speech segment can be represented by a single vector, the i-vector, in a low-dimensional space referred to as total variability space [4]. Then, a GMM super-vector can be decomposed as:

$$\mathbf{m_{h,s}} = \boldsymbol{\mu} + \mathbf{T} \cdot \mathbf{w_{h,s}} \tag{1}$$

where $\boldsymbol{\mu}$ is the speaker-and-session-independent component, i.e. the mean super-vector coming of the UBM. A basis of the total-variability subspace is given by the rows of $\mathbf{T}$, which is a rectangular matrix of low rank and $\mathbf{w_{h,s}}$ is a vector normally distributed with parameters $N(\mathbf{0}, \mathbf{I})$. Extracting an i-vector $\mathbf{w_{h,s}}$, is essentially a Maximum a-Posteriori adaptation (MAP) in the subspace defined by $\mathbf{T}$.

In this paper, a pooled total-variability approach is utilized for convenience, as considered by McLaren and van Leeuwen [20]. Therefore, all available training speech has been compiled into a dataset regardless its source (telephone, microphone or interview).

3.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), was used in [4] to minimize the intra-class variance and maximize the between-class variance in the total variability space. These techniques attempt to project i-vectors onto a new set of orthogonal axes, so that those which belong to the same speaker lay into the same region, and apart from others. This problem is defined according to the Rayleigh coefficient:

$$J(\mathbf{v}) = \frac{\mathbf{v}^t \cdot \mathbf{S}_b \cdot \mathbf{v}}{\mathbf{v}^t \cdot \mathbf{S}_w \cdot \mathbf{v}} \tag{2}$$

where $\mathbf{v}$ represents a space direction, $\mathbf{S}_b$ is the between-class variance and $\mathbf{S}_w$ is the within-class variance. Therefore, $J(\mathbf{v})$ is proportional to the quality of the LDA performance. To calculate the variances:

$$\mathbf{S_b} = \sum_{s=1}^{S} (\mathbf{w_s} - \overline{\mathbf{w}}) \cdot (\mathbf{w_s} - \overline{\mathbf{w}})^t \tag{3}$$

$$\mathbf{S_w} = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{ns} (\mathbf{w_i^s} - \overline{\mathbf{w}}_\mathbf{s}) \cdot (\mathbf{w_i^s} - \overline{\mathbf{w}}_\mathbf{s})^t \tag{4}$$

where $\overline{\mathbf{w}}_\mathbf{s}$ is the centroid of the class s, i.e. the mean of the i-vectors of each speaker, S is the number of speakers involved, and n_s is the number of sessions per speaker s. LDA seeks a projection matrix which consists in the eigenvectors whose eigenvalues are the highest from the general equation:

$$\mathbf{S_b} \cdot \mathbf{v} = \lambda \cdot \mathbf{S_w} \cdot \mathbf{v} \tag{5}$$

3.3. Mahalanobis distance scoring

The Mahalanobis distance originates from the Euclidean metric concept and has been proved to outperform the classical cosine distance [21]. As explained by Bousquet et al. [7], given a new observation $\mathbf{w}$ representing an i-vector, the goal of a statistical classifier is to identify to which class it belongs. Assuming equality of class covariances and Gaussian conditional density models, an i-vector $\mathbf{w}$ is assigned to that particular class which minimizes:

$$(\mathbf{w} - \overline{\mathbf{w}}_\mathbf{s})^\mathbf{t} \cdot \mathbf{W^{-1}} \cdot (\mathbf{w} - \overline{\mathbf{w}}_\mathbf{s}) = \|\mathbf{w} - \overline{\mathbf{w}}_\mathbf{s}\|^2_{\mathbf{W^{-1}}} \tag{6}$$

where $\mathbf{W}$ is the within class covariance matrix. $\overline{\mathbf{w}}_\mathbf{s}$ represents the same as in (4). Note that this score is proportional to the log - probability that $\mathbf{w}$ belongs to the class s. Therefore, the Mahalanobis metric between two i-vectors $\mathbf{w_1}$ and $\mathbf{w_2}$ is:

$$score(\mathbf{w_1}, \mathbf{w_2}) = -\|\mathbf{w_1} - \mathbf{w_2}\|^2_{\mathbf{W^{-1}}} \tag{7}$$

with $\mathbf{W}$ being the within-class covariance matrix of any class of interest. In this work, the Mahalanobis metric is used after LDA.

3.4. Probabilistic Linear Discriminant Analysis

The aim of the approach is to define a set of factors which directly model session and speaker variability in that subspace. The generative model [22] is:

$$\mathbf{w_r} = \overline{\mathbf{w}} + \mathbf{U_1} \cdot \mathbf{x_1} + \mathbf{U_2} \cdot \mathbf{x_{2r}} + \epsilon_\mathbf{r} \quad (8)$$

where $\mathbf{w_r}$ is a feature vector with $r = 1, ..., R$, being R the number of recordings of a speaker; $\mathbf{U_1}$ is the eigenvoice matrix and $\mathbf{U_2}$ is the eigenchannel matrix. $\mathbf{x_1}$, $\mathbf{x_{2r}}$ and $\epsilon_\mathbf{r}$ are respectively the speaker, channel and residual factors.

4. SYSTEM CONFIGURATIONS AND EXPERIMENTS

4.1. Data and evaluation protocol

Unlike previous editions, NIST SRE 2012 [23] core task involved duration mismatch [5]. This increasing interest has led us to choose data which include a more constrained task in terms of duration.

Here we have taken data from a previous NIST evaluation (2008) and the "short-short" condition for which the speech durations are typically 180 s. The shortened test segments are obtained by systematically and successively utilising a lower and lower percentage of the original test segment. Follow the procedure proposed in [2, 24], which takes portions of active speech. This keeps an $x\%$ of frames from the original test excerpts, being 100% the actual NIST 2008 SRE short2 - short3 condition. The mean duration and mode are shown in the Table 1 and the distributions of the utterances' length for each subset are presented in the Figure 1. The standard deviation is in all cases approximately just over 30% of the mode value.

Performance was evaluated using the equal error rate (EER) and the minimum decision cost function (minDCF) with the values proposed by NIST SRE 2008, i.e. $C_{miss} = 10$, $C_{FA} = 1$ and $P_{target} = 0.01$, and for the evaluation condition DET6. Evaluation involved for 12,511 trials employing 1,270 enrolled speakers and 2,528 test segments.

4.2. System configuration

The baseline configuration used for all experiments utilizes an energy-based voice activity detector (VAD) [2, 24] with 19 dimensional feature-warped linear frequency cepstral coefficients (LFCC) and appended delta (19), double delta (11) and the delta energy coefficients, for an overall sum of 50. A gender-dependent UBM of 512 Gaussians has been used, trained on NIST 2004 SRE corpus, specifically 219 male speakers. The total variability matrix and the PLDA estimation involved utterances from NIST-SRE 2004, 2005 and 2006 as well as Switchboard II, Phases II and III and Switchboard Cellular, Parts I and II. In total, 1,289 male speakers and 16,969 sessions have been used. The rank of the total variability matrix (i.e. the dimension of the i-vectors) has been set to 400.

In order to show the sensitivity of the configurations to short duration excerpts, the number of remaining dimensions after LDA has been varied from 50 to 400 in steps of 50. The number of eigenvoices for PLDA has been changed in the same manner while the number of eigenchannels have fixed to 400 throughout. For both, LDA and PLDA, 3 iterations of Eigen Factor Radial (EFR) [25] have been applied for i-vector normalization. Each of these iterations is equivalent to length normalization [26].

Table 1: *The percentage of speech frames from the original and the mean duration of the utterances involved.*

Percentage of remaining frames	Utterances mode duration
2%	2.14 s
5%	5.12 s
10%	10.3 s
20%	20.8 s
30%	30.6 s
40%	40.7 s
50%	50.8 s
75%	74.5 s
100%	99.9 s

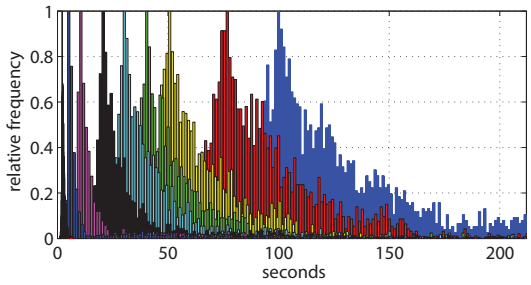
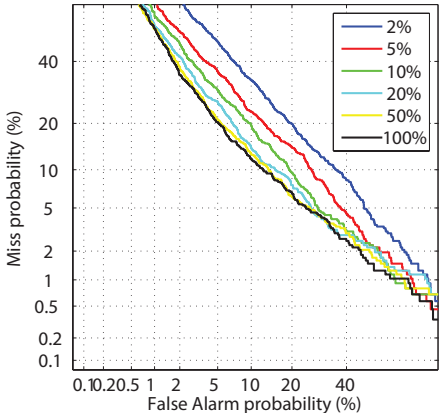


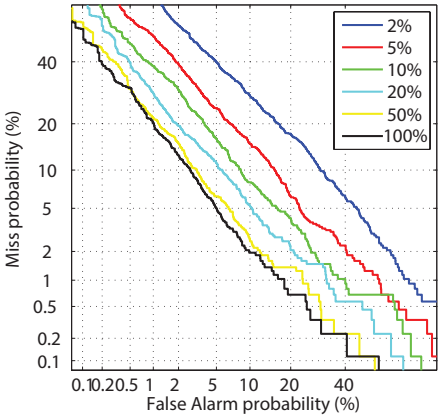
Fig. 1: *Distributions of the given utterance sets after shortening in 8 stages down to approximately 2% of original durations.*

5. PERFORMANCE COMPARISONS

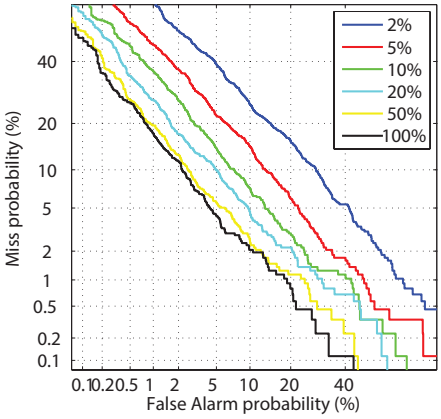
Figure 2 shows results in the form of DET plots for the 3 systems, namely the UBM/GMM and the 2 i-vector systems, LDA and PLDA. Each plot has 6 profiles reflecting the different durations of the test segments, from 2% upto 100%, which is the original recording in full. The immediate difference in the three plots is the bunching of the profiles in the case of the UBM/GMM. This claims that the initial performance on the full original segments is far superior for both the LDA and



(a) *UBM/GMM*



(b) *LDA, k = 300*



(c) *PLDA, nv = 300*

Fig. 2: *DET plots for three systems: (a) a standard UBM/GMM providing a benchmark (b) an i-vector system and LDA, and (c) an i-vector system and PLDA. In all 3 cases the profiles relate to data reduction in the test segment only, from 100% (original duration) down to 2% of the original. Note the profiles for the 2% condition are similar in all 3 cases, with an EER just below 20%*

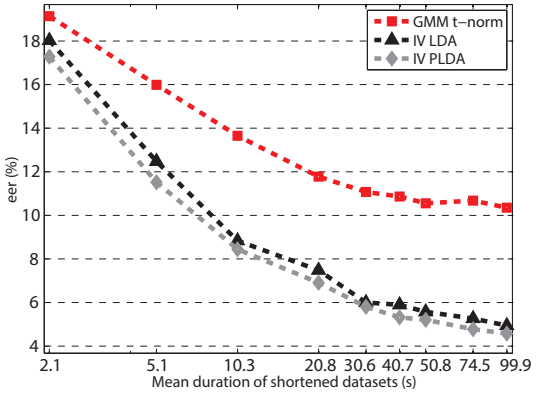


Fig. 3: *EER vs % of active speech retained for each test segment in the original set.*

PLDA i-vector systems.

A second observation is that the upper profile in each case (the one representing just 2% of the original test data) is similar in all 3 cases at just below 20%. So the performance of these three systems in terms of EER, certainly tends to converge when just 2 to 3 seconds are available at the test stage. In contrast, for the full duration (1 to 2 minutes of speech), the performances of the i-vector systems are far better than that of the UBM/GMM, with EER's of approximately 5% for the latter 2 compared with 10% for the UBM/GMM. These duration performances are shown more clearly in Figure 3, with the convergence of the i-vector schemes with the UBM/GMM at 2.5 seconds well illustrated. Following we consider variations in the i-vector parameters, with particular attention to the shortest duration performance, specifically to see if there are any simple gains to be made in this area.

Figure 4 shows a series of profiles for the i-vector LDA and PLDA configurations where each one reflect variations in i-vector subspace dimension, from 50 up to 400. For PLDA, dimension of the speaker factor (represented by $\mathbf{x_1}$ in equation 8) is taken as a parameter. In the case of LDA, the variable to be changed is the number of eigenvectors which form the projection matrix, eq. 5. The profiles show minimal variations, certainly for the higher dimensions, 150 up to 400. This changes at 5% profile and is quite marked at the shortest duration of 2%. Here, in the case of LDA the error rates fall slightly and in the case of PLDA increase. This is perhaps an interesting trend, worthy of further investigation.

Finally, Figure 5 compares both i-vector systems, when only 2% of test data remains. The parameter involved from each scheme is the same as the one varied in the previous figure, respectively. We can observe a knee point in the middle of both graphs, which marks a sustainable change. Considering the profiles, it can be stated that LDA shows a more consistent performance than PLDA, which deteriorates abruptly when only a few dimensions are present.

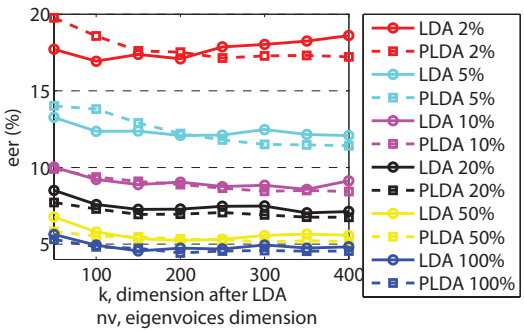
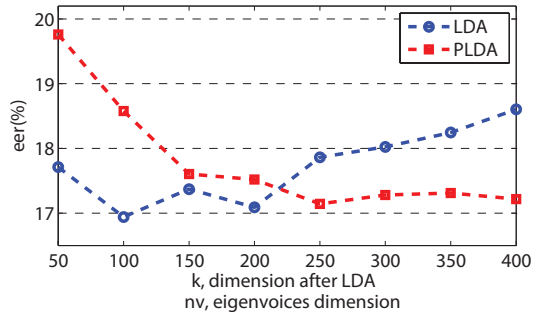
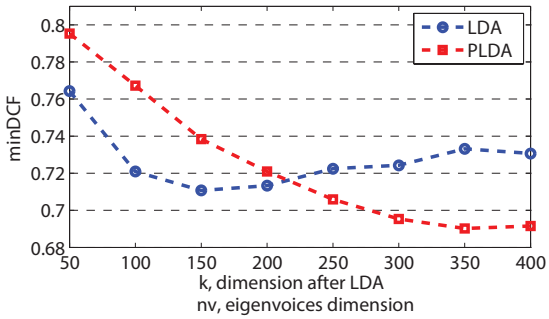


Fig. 4: *Systems' performance against subspace dimension after LDA(k) and PLDA(nv) across all the shortened subsets.*



(a) *EER*



(b) *minDCF*

Fig. 5: *Comparison of LDA and PLDA systems at 2% of active speech. Performances vary according to the values of the space dimension after LDA and the number of eigenvoices in PLDA*

6. CONCLUSIONS

The work described here focuses on short duration segments in the test stage of speaker recognition; it assumes well trained models are available from adequate quantities of enrolment speech.

In this paper we have compared the behaviour or two state-of-the-art i-vector systems and a classical GMM/UBM engine when varying the speech duration. Experimental results confirm the fact that i-vector based systems are indeed markedly better than the UBM/GMM, when sufficient test data are available. Here for instance, the first scores 5% EER against the UBM/GMM of 10% EER with full length utterances. However, such a huge improvement decreases consistently when short duration utterances are involved. In this regard, it is observed a knee point in the region of 10 sec. to 20 sec. and when data is reduced further to the region of 2 sec. to 3 sec., the performance of the two type of systems (i-vector system and UBM/GMM) converge to give very similar scores just below 20% EER. Furthermore, if we consider the most restrictive test condition represented by the blue solid line on 2a, it is remarkable that UBM/GMM gives a performance not very far from the i-vectors one. This observation has led to consider that i-vector does not bring the same improvement for all durations and hereby, it is more sensitive to the lack of data than the UBM/GMM. In the same way, LDA has proved to be more consistent than PLDA when very short utterances are evaluated and dimension reduction is carried out in both approaches.

7. ACKNOWLEDGEMENTS

The results in this paper have been produced thanks to the ALIZE/SpkDETopen source toolkit [21], and in collaboration with the University of Avignon (LIA), France.

8. REFERENCES

[1] Kong Aik Lee, Bin Ma, and Haizhou Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.

[2] B. Fauve, N. Evans, N. Pearson, J.F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker verification," *Annual Conference of the International Speech Communication Association (Interspeech)*, p. 4, 2007.

[3] 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*, Stellenbosch, South Africa, 2008, IEEE.

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

[5] T. Hasan, R. Saeidi, J.H.L. Hansen, and D.A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013, pp. 7663–7667.

[6] Achintya Kumar Sarkar, Driss Matrouf, Pierre-Michel Bousquet, and Jean-Franois Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Ut-

terance Duration for Speaker Verification.," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[7] P.-M. Bousquet, D. Matrouf, and J.-F. 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.

[8] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, "Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011.

[9] 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.

[10] David Martínez González, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka, "Language Recognition in iVectors Space," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 861–864.

[11] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013, pp. 7673–7677.

[12] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011.

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

[14] A. Kanagasundaram, R. J. Vogt, David B. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in *Odyssey Speaker and Language Recognition Workshop*, Singapore, June 2012.

[15] A. Kanagasundaram, R. J. Vogt, D. B. Dean, S. Sridharan, and Michael W. Mason, "i-vector based speaker recognition on short utterances," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 2341–2344.

[16] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using Adapted Gaussian mixture models," in *Digital Signal Processing*, 2000.

[17] B. Fauve, *Tackling Variabilities in Speaker Verification with a Focus on Short Durations*, Ph.D. thesis, Swansea University, 2009.

[18] Frederic Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meigner, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz, and Douglas A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, April 2004.

[19] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., CRIM, 2005.

[20] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011, pp. 5460–5463.

[21] Anthony Larcher, Jean-François Bonastre, Benoit G. B. Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John S. D. Mason, and Jean-Yves Parfait, "ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 2768–2772.

[22] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," *Odyssey Speaker and Language Recognition Workshop*, June 2010.

[23] *The NIST year 2012 speaker Recognition evaluation plan*, 2012.

[24] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification," *Odyssey Speaker and Language Recognition Workshop*, p. 7, 2008.

[25] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[26] 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.