IMPROVED SPEAKER RECOGNITION WHEN USING I-VECTORS FROM MULTIPLE SPEECH SOURCES

Mitchell McLaren and David van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{m.mclaren, d.vanleeuwen}@let.ru.nl

ABSTRACT

The concept of speaker recognition using i-vectors was recently introduced offering state-of-the-art performance. An i-vector is a compact representation of a speaker's utterance after projection into a low-dimensional, total variability subspace trained using factor analysis. A secondary process involving linear discriminant analysis (LDA) is then used to improve the discrimination of i-vectors from different speakers. The newness of this technology invokes the question as to the best way to train the total variability subspace and LDA matrix when using speech collected from distinctly different sources. This paper presents a comparative study of a number of subspace training techniques and a novel source-normalised-and-weighted LDA algorithm for the purpose of improving i-vector-based speaker recognition under mis-matched evaluation conditions. Results from the NIST 2010 speaker recognition evaluation (SRE) suggest that accounting for source conditions in the LDA matrix as opposed to the total variability subspace training regime provides improved robustness to mis-matched evaluation conditions.

Index Terms— speaker recognition, i-vector, total variability, source conditions, linear discriminant analysis

1. INTRODUCTION

The introduction of i-vectors as features for speaker recognition has recently amounted to a new standard in state-of-the-art technology [1, 2]. This configuration extracts i-vectors from a low-dimensional total variability subspace prior to improving speaker discrimination via linear discriminant analysis (LDA) and within-class covariance normalisation (WCCN) and performing classification using a cosine kernel. A number of highly competitive submissions to the recent NIST 2010 speaker recognition evaluations (SRE) demonstrated the potential for this new classifier with performance often exceeding that offered by the widely adopted joint factor analysis (JFA) approach to speaker verification [3].

The training regime for the total variability subspace was derived from JFA [3]. However, rather than explicitly training speaker and channel subspaces as is carried out with JFA, the i-vector approach attempts to bound all observable variation to a single, low-dimensional subspace. In using a single subspace, useful speaker discriminative information is retained that may have otherwise been captured and removed by the JFA channel subspace [1].

The developments in i-vector classification have focussed on telephony speech with few investigations into the effects of the microphone or interview speech often encountered in recent NIST SREs [2, 1, 4, 5]. The problem with these latter sources is the difficulty in estimating a suitable total variability subspace conditioned to interview or microphone speech due to the limited training data that is often available. Consequently, speaker recognition is challenging when trial segments are acquired from different sources (mis-match) or sources not well represented during system development.

A recent study used ample telephone speech to supplement sparsely available microphone speech and improve system classification performance under microphone conditions [5]. This was performed using a *tiered* approach to subspace estimation. With regards to JFA, the concatenation of source-conditioned channel subspaces has proven an effective way to improve robustness to mis-matched evaluation conditions [6]. These approaches may also be capable of reducing the effects of mis-match directly in the total variability subspace. In the subsequent LDA process, a source-normalised-and-weighted (SNAW) LDA algorithm [7] has been shown to effectively minimise errors due to mis-match.

This paper investigates the use of several source-conditioned training regimes for the total variability subspace and the LDA matrix in the i-vector framework for speaker verification. The sources of speech considered include telephone, microphone and interview. The objective of this study is to determine, firstly, which of the proposed subspace learning approaches are robust to mis-matched evaluation conditions, and secondly, whether attempts to improve robustness to mis-match should be implemented during subspace estimation or in the latter LDA process. Three different subspace training regimes are evaluated on the recent NIST 2010 SRE corpus: pooled, tiered and concatenated subspace training. Source-conditioned learning of the LDA matrix is also evaluated using the SNAW-LDA algorithm developed for the i-vector framework [7].

The layout of this paper is as follows. Section 2 details the processes involved in the i-vector framework for speaker verification. The proposed subspace learning techniques are described in Section 3 with the source-weighted-and-normalised LDA algorithm detailed in Section 4. Section 5 details the experimental protocol with corresponding results and analysis presented in Section 6.

2. I-VECTOR EXTRACTION AND CLASSIFICATION

This section describes the stages involved in the i-vector framework developed by Dehak et al. [2]. Given the centralised Baum-Welch statistics from all available speech utterances [3], these stages include subspace training, LDA, WCCN and classification using a cosine kernel function.

2.1. The Total Variability Subspace

The total variability subspace training regime assumes that an utterance can be represented by the Gaussian mixture model (GMM)

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 238803.

mean supervector, $\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}$, where $\boldsymbol{M}$ consists of a speaker- and session-independent mean supervector $\boldsymbol{m}$ from the universal background model (UBM) and a mean offset $\boldsymbol{T}\boldsymbol{w}$. The supervector $\boldsymbol{M}$ is assumed to be normally distributed with mean $\boldsymbol{m}$ and covariance $\boldsymbol{T}\boldsymbol{T}^t$, where $\boldsymbol{T}$ is the low-rank, total variability subspace. The low-rank vector $\boldsymbol{w}$ has a standard normal distribution $\mathcal{N}(0,1)$ and is referred to as the *i-vector*.

The training regime for the total variability subspace $\boldsymbol{T}$ involves the same algorithm used to train the speaker subspace in the JFA approach [3]. However, rather than estimating a subspace conditioned to the observable between-speaker variability, the greatest directions of between-utterance variability are estimated. Subsequently, this training regime alleviates the need for speaker-labelled utterances in the subspace training dataset.

In order to extract an i-vector, the centralised Baum-Welch zero and first order statistics [3] ($\boldsymbol{N}$ and $\boldsymbol{F}$, respectively) are calculated for an utterance with respect to the UBM having C Gaussian components learned from features of dimension F. The i-vector representing the utterance can then be calculated as,

$$\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{N}}\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{F}, \tag{1}$$

where $\boldsymbol{I}$ is a $CF \times CF$ identity matrix, $\hat{\boldsymbol{N}}$ is a diagonal matrix with $F \times F$ blocks $N_c\boldsymbol{I}$ $(c = 1, \ldots, C)$ and $\boldsymbol{F}$ is the supervector formed through the concatenation of all centralised first-order statistics. The covariance matrix $\boldsymbol{\Sigma}$ represents the residual variability not captured by $\boldsymbol{T}$ and results from the JFA subspace training procedure. An efficient procedure for the optimisation of the model parameters $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ is described by [3].

2.2. Linear Discriminant Analysis

After an i-vector has been extracted from a speech sample, LDA is performed as a means of inter-session compensation. In the current context, LDA attempts to find a reduced set of axes that minimises the within-speaker variance while maximising between-speaker variance observed in the i-vector space. The LDA transform $\boldsymbol{A}$ consists of the eigenvectors having the largest eigenvalues when solving the eigenvalue problem $\boldsymbol{S_B}\boldsymbol{v} = \lambda\boldsymbol{S_W}\boldsymbol{v}$ where the between- and within-speaker scatter matrices, $\boldsymbol{S_B}$ and $\boldsymbol{S_W}$ respectively, are calculated,

$$\boldsymbol{S}_B = \sum_{s=1}^{S} N_s(\boldsymbol{\mu}_s - \boldsymbol{\mu})(\boldsymbol{\mu}_s - \boldsymbol{\mu})^t \tag{2}$$

$$\boldsymbol{S}_W = \sum_{s=1}^{S}\sum_{i=1}^{N_s}(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)^t. \tag{3}$$

The covariance matrices are calculated from a training dataset sourced from S speakers in which each speaker s has utterance i-vectors $\boldsymbol{w}_i^s$ $(i = 1, \ldots, N_s)$ and a speaker mean $\boldsymbol{\mu}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}\boldsymbol{w}_i^s$. The i-vector mean $\boldsymbol{\mu} = 0$ due to the factor analysis assumption of normally distributed and zero-mean factors [2, 5].

2.3. Within-Class Covariance Normalisation

The final stage undertaken before i-vectors are used for classification is within-class covariance normalisation (WCCN) [8]. Although originally developed for SVM-based speaker verification, WCCN has proven useful in normalising the within-speaker variance remaining in LDA-reduced i-vectors. The WCCN matrix $\boldsymbol{B}$ is found through the Cholesky decomposition of $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$ where the within-class covariance matrix is calculated as,

$$\boldsymbol{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}(\boldsymbol{A}^t\boldsymbol{w}_i^s - \hat{\boldsymbol{\mu}_s})(\boldsymbol{A}^t\boldsymbol{w}_i^s - \hat{\boldsymbol{\mu}_s})^t. \tag{4}$$

Distinguishing a difference from (3), the mean of the LDA-reduced i-vectors from speaker s is equated as $\hat{\boldsymbol{\mu}_s} = \frac{1}{N_s}\sum_{i=1}^{N_s}\boldsymbol{A}^t\boldsymbol{w}_i^s$.

2.4. Cosine Similarity Scoring

The classification score for a given trial between two i-vectors is given by the cosine distance. Specifically, a cosine kernel incorporating the LDA and WCCN matrices gives the similarity score $\langle\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2\rangle$ where,

$$\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{B}^t\boldsymbol{A}^t\boldsymbol{w}_i}{\|\boldsymbol{B}^t\boldsymbol{A}^t\boldsymbol{w}_i\|}. \tag{5}$$

As both training and testing i-vectors undergo the same transformations and there is no explicit modelling or enrolment stage, the cosine distance can be seen as a symmetric classification method. Cosine kernel normalisation [4] takes advantage of this symmetry and alleviates the need for common score-based normalisation. The normalised cosine kernel is given by,

$$\text{score}(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2) = \frac{(\hat{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_{\text{imp}})^t(\hat{\boldsymbol{w}}_2 - \overline{\boldsymbol{w}}_{\text{imp}})}{\|\boldsymbol{C}_{\text{imp}}\hat{\boldsymbol{w}}_1\|\|\boldsymbol{C}_{\text{imp}}\hat{\boldsymbol{w}}_2\|}. \tag{6}$$

Here, a set of impostor i-vectors are subjected to (5) and used to estimate an impostor mean $\overline{\boldsymbol{w}}_{\text{imp}}$ in the cosine kernel space and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\text{imp}} = (\boldsymbol{C}_{\text{imp}})^2$.

3. TOTAL VARIABILITY SUBSPACE TRAINING APPROACHES

The total variability subspace is responsible for defining a suitable space from which i-vectors are extracted. The subspace should, therefore, be trained in a manner that best exploits the useful speaker variability from speech acquired from a variety sources. To date, most literature has focussed on training the subspace from telephony speech with little investigation into alternate sources of speech [2, 1, 4, 5].

This section presents three different subspace training approaches to deal with the speech acquisition methods observed in the recent NIST SREs: telephone, microphone and interview. While telephony speech is widely available, resources for microphone and interview speech are limited and typically taken from previous NIST SREs. Finding a training regime that is able to robustly estimate the total variability subspace from such limited resources is desired.

3.1. Pooled

The *pooled* approach to subspace training involves compiling a training dataset from all telephone-, microphone- and interview-sourced speech. The training regime outlined in Section 2.1 is then used to train a single subspace. In this approach, no attempt is made to normalise for the heavy bias toward telephone speech in the training dataset. A major advantage to this approach, however, lies in it's relative simplicity and computational efficiency compared to the approaches presented below. In this work, the pooled approach used a pooled training dataset to train a single 400 dimensional subspace.

3.2. Tiered Training

A recent study investigated the use of limited microphone speech to complement a subspace trained on telephone speech [5]. In this approach a telephone subspace was trained on ample speech prior to estimating a secondary subspace to capture any variation from microphone-based speech that was not found in the telephone subspace. A single total variability subspace was then formed through

the concatenation of the larger telephone subspace and the smaller microphone subspace. This approach is extended here to account for additional interview-sourced speech.

Specifically, a 400-dimensional telephone subspace is firstly trained based on the assumption that $\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}_{\text{tel}}\boldsymbol{w}_{\text{tel}}$. A secondary 100-dimensional subspace $\boldsymbol{T}_{\mathbf{mic}}$ is then trained from microphone data based on $\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}_{\text{tel}}\boldsymbol{w}_{\text{tel}} + \boldsymbol{T}_{\text{mic}}\boldsymbol{w}_{\text{mic}}$. Similarly, the interview training dataset is then used to train the final 100-dimensional subspace $\boldsymbol{T}_{\text{int}}$ assuming,

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}_{\text{tel}}\boldsymbol{w}_{\text{tel}} + \boldsymbol{T}_{\text{mic}}\boldsymbol{w}_{\text{mic}} + \boldsymbol{T}_{\text{int}}\boldsymbol{w}_{\text{int}}. \quad (7)$$

Both $\boldsymbol{T}_{\text{mic}}$ and $\boldsymbol{T}_{\text{int}}$ consist of relatively few dimensions as they need only capture the variability not already found in the preceding subspaces. The resulting subspaces are concatenated to form the total variability subspace $\boldsymbol{T} = [\boldsymbol{T}_{\text{tel}}\boldsymbol{T}_{\text{mic}}\boldsymbol{T}_{\text{int}}]$ which is subject to a singular value decomposition. This approach will be referred to as *tiered* subspace training in this work. For a more detailed description of the tiered training procedure, readers are directed to [5].

3.3. Concatenation

The concept of concatenating total variability subspaces extends from the common approach used to estimate the channel subspace in the joint factor analysis (JFA) framework for speaker verification [6]. Typically telephone, microphone and interview channel subspaces are trained prior to their concatenation into a single, larger subspace. In the context of JFA, these subspaces capture the within-speaker variability attributed to each source. The independent estimation of total variability subspaces is likely to result in overlap in the true speaker variation observed between the spaces. It is hypothesised that this bias toward speaker variation may aid in suppressing the source-related variation observed in the i-vectors. The following approach was taken for subspace concatenation in this work.

Three 400-dimensional total variability subspaces $\boldsymbol{T}_{\text{src}}$ $(\text{src} \in \{\text{tel}, \text{mic}, \text{int}\})$ were trained from each source-dependent subset of the training data. A single subspace was then formed through the concatenation of the individual subspaces such that $\boldsymbol{T} = [\boldsymbol{T}_{\text{tel}}\boldsymbol{T}_{\text{mic}}\boldsymbol{T}_{\text{int}}]$. A singular value decomposition was performed on $\boldsymbol{T}$ to ensure orthogonality of directions in the new subspace. Finally, the residual covariance matrix $\boldsymbol{\Sigma}$ in (1) was re-estimated using the pooled training dataset and the new subspace $\boldsymbol{T}$ of 1200 dimensions according to the procedure described in [3].

4. SOURCE-NORMALISED-AND-WEIGHTED LDA

Source-normalised-and-weighted (SNAW) LDA [7] is an effective means of reducing classification errors attributed to mis-matched evaluation conditions directly in the i-vector space. The development of this technique was motivated by [5] in which a weighted LDA algorithm exploited widely availability telephone speech to improve the classification of scarcely-resourced microphone speech. SNAW-LDA aims to remove the influence of variation attributed to different speech acquisition methods on the between-speaker scatter while providing an accurate estimation of the within-speaker scatter from a training dataset void of multi-source utterances per speaker.

Extending on the between-speaker covariance calculation (2), the source-normalised-and-weighted $\boldsymbol{S_B}$ is calculated as,

$$\boldsymbol{S}_B = \sum \frac{S_{\text{src}}}{S} \boldsymbol{S}_B^{\text{src}} \quad (8)$$

$$\boldsymbol{S}_B^{\text{src}} = \sum_{s=1}^{S_{\text{src}}} N_s (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})^t, \quad (9)$$

where $\boldsymbol{\mu}_{\text{src}}$ and N_{src} are the source-conditioned i-vector mean and speech sample count, respectively. In this approach, i-vectors are normalised with respect to their corresponding source mean. Consequently, speaker utterances acquired from different sources are implicitly assumed to belong to disjoint speaker sets. Further, the weight assigned to each scatter matrix $\boldsymbol{S}_B^{\text{src}}$ is the proportion of total training data used in it's calculation.

The within-speaker scatter in the standard LDA approach can be formulated as $\boldsymbol{S}_W = \boldsymbol{S}_T - \boldsymbol{S}_B$ where to the total variance in the pooled training data $\boldsymbol{S}_T = \sum_{n=1}^{N} \boldsymbol{w}_n \boldsymbol{w}_n^t$ since the source-independent sample mean is zero. A derivation of this formula is used to calculate the source-normalised $\boldsymbol{S}_W$:

$$\boldsymbol{S}_W = \boldsymbol{S}_T - \textstyle\sum_{\text{src}} \boldsymbol{S}_B^{\text{src}}. \quad (10)$$

Interested readers can find a more detailed description of the SNAW-LDA algorithm in the accompanying paper [7].

5. EXPERIMENTAL PROTOCOL

The proposed approaches were evaluated on the recent NIST 2010 SRE corpus. Results are reported for four evaluation conditions with particular focus on mis-matched conditions. Corresponding to det conditions 2-5 in the SRE'10 evaluation plan [9], these include *int-int*, *int-mic*, *int-tel*, and *tel-tel*. Performance is evaluated using the equal error rate (EER) and a normalised minimum decision cost function (DCF) calculated using $C_{\text{miss}} = 1$, $C_{\text{FA}} = 1$ and $P_{\text{tar}} = 0.001$ [9]. The *extended* evaluation protocol is used to ensure sufficient impostor trials to estimate the minimum DCF. The number of trials for a given condition range from 416119 (*tel-tel*) to more than 2.8 million (*int-int*) with 0.5-1.7% belonging to target trials. Gender-pooled results are reported throughout. In all configurations, the number of LDA dimensions retained was evaluated in steps of 50 in order to minimise the average of (normalised min. DCF$+10\times$EER) across the evaluated conditions.

Speech activity detection (SAD) involved training a 2-component GMM from the log-energy of the speech signal and setting a threshold based on the resulting speech Gaussian. Low energy frames were iteratively removed from the signal until the standard deviation of the speech Gaussian was less than five times that of the non-speech Gaussian. Dual-SAD was used for SRE'10 interview segments such that an interviewee speech frame was retained if it's normalised energy was at least 5dB greater than than the corresponding interviewer frame. Gender-dependent UBMs consisting of 512-components trained on 60-dimensional, feature-warped MFCCs (including deltas and double-deltas) were used to calculate the Baum-Welch statistics. UBM training data was taken from the NIST 2004, 2005, and 2006 SRE corpora and LDC releases of Fisher English, Switchboard II: phase 3 and Switchboard Cellular (parts 1 and 2).

A single, gender-dependent dataset was compiled for total variability subspace, LDA and WCCN training and normalisation of the cosine kernel. This dataset was formed from the same corpora used to train the UBMs with additional interview data taken from the follow-up corpus of the NIST 2008 SRE. The distribution of speech samples for the male and female datasets was, on average, $[N_{\text{tel}}, N_{\text{mic}}, N_{\text{int}}] = [15770, 2665, 1268]$ acquired from $2670, 88,$ and 75 speakers, respectively. SRE'05 and SRE'06 speakers of microphone segments also provided telephony speech samples.

6. RESULTS

The following experiments aim to address two specific questions: (1) how should the total variability subspace be trained to maximise

LDA Matrix	TV Space	Optimised LDA Dim.	int-int		int-tel		int-mic		tel-tel	
			DCF	EER	DCF	EER	DCF	EER	DCF	EER
Standard	Pooled 400D	250	**.5743**	**4.72%**	**.6211**	**4.79%**	**.4598**	**3.55%**	.6124	**4.62%**
	Tiered 600D	400	.6339	5.98%	.6545	5.32%	.4988	4.21%	**.6009**	4.97%
	Concatenated 1200D	600	.6404	6.89%	.6748	6.45%	.5020	4.90%	.6162	5.50%
SNAW	Pooled 400D	150	**.5345**	**3.58%**	**.5578**	4.32%	**.4121**	**2.68%**	.6123	4.37%
	Tiered 600D	150	.5484	3.80%	.5722	**4.21%**	.4165	2.86%	.5964	**4.34%**
	Concatenated 1200D	200	.5673	4.31%	.5867	4.64%	.4353	3.00%	**.5947**	4.49%

Table 1. Comparing Pooled, Tiered, and Concatenated subspace training methods when incorporating standard LDA or source-normalised-and-weighted (SNAW) LDA evaluated on the SRE'10 (extended protocol).

robustness to mis-matched evaluation conditions, and (2) whether it is better to account for such mis-match during subspace training or in the latter LDA process.

6.1. Comparison of Subspace Training Regimes

Section 3 described three methods for training the total variability subspace: pooled, tiered and concatenated. Results when evaluating each of these techniques on SRE'10 using the standard LDA algorithm are presented in the top of Table 1. These results indicate that the pooled approach to subspace training provided, in general, better classification performance than the tiered and concatenated approaches. This was particularly evident in terms of EER for trials involving microphone or interview speech. It is apparent, therefore, that i-vectors extracted from the tiered or concatenated subspace provided less robustness to mis-matched evaluation conditions compared to the subspace learning from a pooled dataset. Further, it can be observed that a larger total variability subspace did not necessarily bring improved performance. These findings suggest that source-conditioned learning of the total variability subspace does not improve robustness to mis-matched conditions and that the additional complexity and computation associated with the tiered and concatenated approaches may not be worthwhile.

6.2. Standard LDA versus SNAW-LDA

Each of the subspace training approaches were evaluated in conjunction with source-normalised-and-weighted (SNAW) LDA. The objective here was to determine whether accounting for source-related variation was more appropriate in the i-vector space relative to the GMM supervector space. Results from trials incorporating SNAW-LDA are presented in the bottom of Table 1.

A comparison of results obtained when using standard LDA and SNAW-LDA indicate that the latter approach provided a significant performance advantage to cross-source trials (*int-mic* and *int-tel*). In these conditions, relative improvements of 10-16% in minimum DCF and 10–39% in EER were observed over the standard LDA approach across the subspace training regimes. In contrast to standard LDA, the number of dimensions retained in the SNAW-LDA matrix were similar across each of the subspace training approaches while also requiring fewer directions for speaker discrimination. Surprisingly, the worst performing configuration (concatenated) in conjunction with SNAW-LDA provided superior performance to the pooled subspace with standard LDA. These findings suggest that improved robustness to mis-matched source conditions in the i-vector framework can be more readily achieved by dealing with such mis-match in the i-vector space as opposed to the GMM supervector space.

In the case of same-source trials, the *int-int* condition found significant advantage through SNAW-LDA while this was only apparent in the EER statistic of the *tel-tel* trials. Evidently, SNAW-LDA provided considerable benefits to classification performance for those conditions with limited development data. It can also be noted that, for each evaluated condition, less variation was observed in performance from the different subspace training approaches when using SNAW-LDA as opposed to the standard LDA approach. This demonstrates the effectiveness of SNAW-LDA to exploit the true speaker characteristics observed in i-vectors extracted from different total variability subspaces.

7. CONCLUSION

This study compared the robustness of source-conditioned total variability subspace and LDA learning approaches to mis-matched evaluation conditions in the i-vector framework for speaker verification. Pooled, tiered and concatenated subspace training approaches were evaluated on the recent NIST 2010 SRE. Results demonstrated that the pooled approach held a performance advantage over the alternate, source-conditioned approaches while also having lower complexity and computational cost. In contrast to the total variability space, introducing source-conditioned learning in the LDA optimisation using SNAW-LDA provided improved robustness to mis-matched evaluation conditions and reduced minimum DCF and EER statistics by up to 16% and 39%, respectively.

8. REFERENCES

[1] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, 2009, pp. 1559–1562.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *In print IEEE Trans. Audio, Speech and Language Processing*, 2010.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, , and P Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 980–988, 2008.

[4] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010.

[5] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recogntion Workshop*, 2010, pp. 28–33.

[6] N. Scheffer, R. Vogt, S. Kajarekar, and J. Pelecanos, "Combination strategies for a factor analysis phone-conditioned speaker verification system," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2009.

[7] M. McLaren and D. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *accepted into IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011.

[8] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Ninth Int. Conf. on Spoken Language Processing*, 2006, pp. 1471–1474.

[9] National Institute of Standards and Technology, *NIST 2010 SRE Evaluation Plan*, Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/.