Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources

Mitchell McLaren, *Member, IEEE*, and David van Leeuwen, *Member, IEEE*

***Abstract*—The recent development of the i-vector framework for speaker recognition has set a new performance standard in the research field. An i-vector is a compact representation of a speakers utterance extracted from a total variability subspace. Prior to classification using a cosine kernel, i-vectors are projected into an linear discriminant analysis (LDA) space in order to reduce inter-session variability and enhance speaker discrimination. The accurate estimation of this LDA space from a training dataset is crucial to detection performance. A typical training dataset, however, does not consist of utterances acquired through all sources of interest for each speaker. This has the effect of introducing systematic variation related to the speech source in the between-speaker covariance matrix and results in an incomplete representation of the within-speaker scatter matrix used for LDA. The recently proposed source-normalized (SN) LDA algorithm improves the robustness of i-vector-based speaker recognition under both mis-matched evaluation conditions and conditions for which inadequate speech resources are available for suitable system development. When evaluated on the recent NIST 2008 and 2010 Speaker Recognition Evaluations (SRE), SN-LDA demonstrated relative improvements of up to 38% in equal error rate (EER) and 44% in minimum DCF over LDA under mis-matched and sparsely resourced evaluation conditions while also providing improvements in the common telephone-only conditions. Extending on these initial developments, this study provides a thorough analysis of how SN-LDA transforms the i-vector space to reduce source variation and its robustness to varying evaluation and LDA training conditions. The concept of source-normalization is further extended to within-class covariance normalization (WCCN) and data-driven source detection.**

***Index Terms*—Cross-channel source variation, i-vector, linear discriminant analysis (LDA), speaker recognition, total variability.**

I. Introduction

THE recently developed i-vector framework for speaker recognition offers state-of-the-art performance [1], [2]. This framework represents a speech utterance as a vector of intermediate size between feature vector and supervector size, referred to as an *i-vector*, extracted from a low-dimensional total variability subspace. The total variability subspace defines the space in which the majority of observable utterance-to-utterance variation is captured by the i-vectors. As i-vectors are associated with utterances, one of the underlying objectives of the speaker recognition framework is to isolate the speaker-dependent information contained in the i-vectors. The subsequent application of linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) attempts to accomplish this by improving speaker discrimination prior to performing classification using a cosine kernel. As i-vectors in their raw state capture both speaker-intrinsic and speaker-extrinsic variation, it is ultimately the role of LDA to define a suitable space in which speakers can be discriminated from one another.

Linear discriminant analysis (LDA) aims to find a reduced set of axes onto which i-vectors can be projected such that the within-speaker covariance is minimized while the between-speaker covariance is simultaneously maximized. Sources of within-speaker variability include different transmission channels, microphones, acoustic environments, speaking styles, phonetic content etc., that contribute to the differences observed between utterances of the same speaker [3]. In contrast, between-speaker variation is due to the differences between the characteristics of different speakers and is the primary variation to be maximized in the LDA process. LDA, therefore, relies on the calculation of suitable covariance (or scatter) matrices from a training dataset in order to determine a set of axes optimized for *inter-session compensation*—that is, the suppression of within-speaker variability while retaining the observable between-speaker variability.

Speaker recognition using conversational telephony speech has received considerable focus in the research field, resulting in an abundance of corresponding data becoming available for system development and tuning. It is only in recent years that the NIST speaker recognition evaluations (SRE) [4] have included microphone and interview-style speech for which limited resources are available for system development. Consequently, robust speaker verification is challenging when non-telephone speech is encountered during system evaluation [5], [6]. This work is concerned with the effects of speech source-related variation which is a subset of the commonly used term *channel* variation that typically encompasses all aforementioned sources of within-speaker variation. The source variation targeted in this work is the speech acquisition method or recording scenario. This is a data collection characteristic that is labeled in recent NIST SREs. Specifically, focus is given to "telephone," "microphone." or "interview" speech sources. It should be noted, how-

Manuscript received January 20, 2011; revised April 29, 2011 and July 25, 2011; accepted July 29, 2011. Date of publication August 15, 2011; date of current version January 11, 2012. This work was supported by the European Community's Seventh Framework Program (FP7/2007–2013) under grant agreement no. 238803. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

The authors are with the Center for Language and Speech Technology, Radboud University Nijmegen, 6500HC, The Netherlands (e-mail: m.mclaren@let.ru.nl; d.vanleeuwen@let.ru.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2164533

ever, that the LDA technique described in this work could potentially be employed to counteract any of the aforementioned sources of variation.

Insufficient speech resources contribute to a reduction in the effectiveness of LDA projection due to the inaccurate estimation of the scatter matrices from the training dataset. As initially described in [7] and analyzed in depth in this study, the between-speaker scatter matrix contains adverse bias toward the different speech acquisition methods and the within-speaker scatter is typically incomplete as utterances recorded from each source of interest are not available for each speaker. Instead, the majority of a speakers' utterances in a typical LDA training dataset are representative only of a single speech source. Consequently, the scatter matrices used in LDA are often not optimized for task of cross-speech-source speaker recognition.

This paper focuses on the recently proposed source-normalized (SN) LDA algorithm used to improve the estimation of the LDA scatter matrices from a training dataset in which each speaker does not provide utterances from each source of interest. Between-speaker covariance is estimated on a source-conditioned basis, thereby, reducing the adverse bias attributed to the speech source. The within-speaker scatter is given by the residual total variability in the i-vector space that is not observed as between-speaker scatter rather than through an explicit calculation that relies on sufficient multi-source examples in the training dataset. The SN-LDA approach is evaluated using the recent NIST 2008 and 2010 SREs. Extending on [7] and [8], the robustness of SN-LDA is analyzed with respect to the varying amounts of evaluation speech, the size of the LDA training dataset, and varying LDA dimensions. The benefit of SN-LDA for SVM-based i-vector classification is also presented. Further, the potential of source-normalized WCCN is investigated along with the data-driven detection of source in the context of SN-LDA.

This paper is structured as follows. Section II reviews the i-vector framework for speaker recognition. Section III details the standard LDA algorithm and is followed by the description and analysis of SN-LDA in Section IV. The experimental protocol and corresponding results are given in Sections V and Section VI.

II. i-Vector Framework for Speaker Recognition

This section describes the i-vector framework developed by Dehak *et al.* [1], [2]. Given the centralized Baum–Welch statistics from all available speech utterances [6], the flow of the framework involves subspace training, linear discriminant analysis (LDA), within-class covariance normalization (WCCN) and classification using a cosine kernel function.

A. Total Variability Subspace

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

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{1}$$

where $\boldsymbol{M}$ consists of the 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 training regime for the total variability subspace $\boldsymbol{T}$ involves the same algorithm used to train the speaker subspace in JFA [6]. However, rather than estimating a subspace conditioned to the observable between-speaker variability, the directions of greatest between-utterance variability are estimated. Subsequently, this training regime alleviates the need for speaker-labeled utterances in the subspace training dataset.

The low-rank vector $\boldsymbol{w}$, referred to as the *i-vector*, is the point estimate of $\boldsymbol{w}$ given a set of statistics and has a standard normal distribution $\mathcal{N}(0,1)$. In order to extract an i-vector, the Baum–Welch zero and centralized first-order statistics [6] ($\boldsymbol{N}$ 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 [2]

$$\boldsymbol{w} = \left(\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{T}\right)^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{F} \tag{2}$$

where $\boldsymbol{I}$ is a $CF \times CF$ identity matrix, $\boldsymbol{N}$ is a diagonal matrix with $F \times F$ blocks $N_c\boldsymbol{I}^F$ $(c = 1, \ldots, C)$, and the supervector $\boldsymbol{F}$ is formed through the concatenation of the centralized first-order statistics. The covariance matrix $\boldsymbol{\Sigma}$ represents the residual variability not captured by $\boldsymbol{T}$. An efficient procedure for the optimization of model parameters $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ is described by [6]. In this work, gender-dependent total variability subspaces were estimated by pooling the training data from all speech sources.

B. Inter-Session Compensation

The total variability space bounds both speaker-intrinsic and speaker-extrinsic sources of variation. Consequently, extracted i-vectors in their raw form are not optimized for speaker discrimination and are, therefore, subject to inter-session variability compensation prior to classification. Two techniques are utilised for this purpose in the i-vector framework: LDA and WCCN.

LDA aims find a reduced set of axes $\boldsymbol{A}$ that minimizes the within-speaker variability observed in the i-vectors while simultaneously maximizing the between-speaker variability. This process is covered in detail in Section III with the recently proposed source-normalized (SN) LDA algorithm described in Section IV.

The secondary stage, within-class covariance normalization (WCCN) [3], normalizes across speakers the residual within-speaker variance in the LDA-reduced i-vectors. The WCCN matrix $\boldsymbol{B}$ is calculated through the Cholesky decomposition of $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$ where the within-class covariance matrix is given by

$$\boldsymbol{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\left(\boldsymbol{A}^t\boldsymbol{w}_i^s - \hat{\boldsymbol{\mu}}_s\right)\left(\boldsymbol{A}^t\boldsymbol{w}_i^s - \hat{\boldsymbol{\mu}}_s\right)^t. \tag{3}$$

The number of speakers (or classes) is S, while each speaker provides N_s i-vectors in the training dataset and the mean of

the LDA-reduced i-vectors from speaker s is equated as $\hat{\boldsymbol{\mu}}_s = (1/N_s)\sum_{i=1}^{N_s} \boldsymbol{A}^t \boldsymbol{w}_i^s$.

C. Cosine Distance Scoring

The detection score for a trial between a pair of i-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is given by the cosine distance $\langle \hat{\boldsymbol{w}}_1 \cdot \hat{\boldsymbol{w}}_2 \rangle$ between the inter-session-compensated vectors

$$\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{B}^t \boldsymbol{A}^t \boldsymbol{w}_i}{||\boldsymbol{B}^t \boldsymbol{A}^t \boldsymbol{w}_i||}. \tag{4}$$

In this work, normalization of the cosine kernel is performed using the approach described in [9]. This approach can be seen as an integrated derivation of common score-based normalization [10] that centres the cosine kernel with respect to the impostor score space and is implemented as

$$s(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2) = \frac{(\hat{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_{\text{imp}})^t (\hat{\boldsymbol{w}}_2 - \overline{\boldsymbol{w}}_{\text{imp}})}{||\boldsymbol{C}_{\text{imp}} \hat{\boldsymbol{w}}_1|| ||\boldsymbol{C}_{\text{imp}} \hat{\boldsymbol{w}}_2||}. \tag{5}$$

Here, a set of impostor i-vectors are subjected to (4) and used to estimate an impostor mean $\overline{\boldsymbol{w}}_{\text{imp}}$ in the cosine kernel space and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\text{imp}} = (\boldsymbol{C}_{\text{imp}})^2$.

III. Linear Discriminant Analysis

This section details the LDA algorithm and a common formula variant before analyzing and discussing the shortcomings of the current LDA procedure in the i-vector framework for speaker verification.

A. LDA Algorithm

In the context of the i-vector framework, linear discriminant analysis (LDA) serves the purpose of enhancing discrimination between i-vectors corresponding to different speakers. LDA attempts to find a set of axes that maximizes the between-speaker variation while simultaneously minimizing the within-speaker variation observed in a set of training i-vectors. This is accomplished through the eigenvalue decomposition of

$$\boldsymbol{S}_B \boldsymbol{v} = \lambda \boldsymbol{S}_W \boldsymbol{v} \tag{6}$$

where the between- and within-speaker covariance matrices $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$, respectively, are calculated as

$$\boldsymbol{S}_B = \sum_{s=1}^{S} N_s (\boldsymbol{\mu}_s - \boldsymbol{\mu})(\boldsymbol{\mu}_s - \boldsymbol{\mu})^t \tag{7}$$

$$\boldsymbol{S}_W = \sum_{s=1}^{S} \sum_{i=1}^{N_s} (\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)^t. \tag{8}$$

In the current context, however, the i-vector mean $\boldsymbol{\mu} = 0$ due to the factor analysis assumption of normally distributed and zero-mean factors. The LDA projection matrix, $\boldsymbol{A}$, is formed as the subset of eigenvectors, $\boldsymbol{v}$, from (6) having the largest eigenvalues, λ.

B. Formula Variants

Previous studies regarding i-vector speaker recognition have utilized an LDA algorithm in which all speakers have equal influence in the calculation of the between- and within-speaker scatter matrices [2], [5]. That is, a speakers' corresponding scatter is normalized by the number of utterances they provide in the training dataset. Extending on (7) and (8), this can be formulated as

$$\boldsymbol{S}_B = \sum_{s=1}^{S} (\boldsymbol{\mu}_s - \boldsymbol{\mu})(\boldsymbol{\mu}_s - \boldsymbol{\mu})^t \tag{9}$$

$$\boldsymbol{S}_W = \sum_{s=1}^{S} \frac{1}{N_s} \sum_{i=1}^{N_s} (\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)^t. \tag{10}$$

In contrast to (7), the contribution of speaker s to the final scatter $\boldsymbol{S}_B$ is no longer scaled by N_s. Likewise, a normalization factor $1/N_s$ is introduced in the calculation of $\boldsymbol{S}_W$. This normalization process has the advantage of preventing bias in the scatters from speakers that contribute a relatively large number of utterances in the training dataset. In doing so, however, the inherent reliability of a speakers' scatter when estimated from a large number of utterances is ignored. Equations (7) and (8) attempt to exploit this information by removing the normalization by N_s from (9) and (10). For completeness, a comparison of results obtained on the cross-source conditions (i.e., train on interview speech, test on telephone speech) of the NIST SRE'08 protocol (as detailed in Section V) when using the normalized and unnormalized variants of the LDA formula is provided in Table I. Results when excluding LDA from the i-vector framework are also detailed to illustrate the vital role of LDA in the i-vector framework. The benefits observed from the unnormalized algorithm over the normalized approach may be explained by the average number of utterances per speaker for each source; 6, 30, and 15 for the telephone, microphone and interview speech sources, respectively. It would seem, therefore, that scatter matrices estimated from more data per speaker provide a better estimate of the scatter matrices thus attributing to observed benefit of the unnormalized algorithm. The performance advantage when using the unnormalized approach was consistent across all algorithms and corpora evaluated in this work, thus, the unnormalized variant of LDA was the fundamental algorithm used in this study.

TABLE I
NIST SRE'08 Results Comparing Normalised and Unnormalized Variants of the Standard LDA Algorithm

	tel-mic		int-tel	
LDA Training Algorithm	DCF	EER	DCF	EER
No LDA	.0637	16.54%	.0642	15.30%
Normalised - Eq. (9) and (10)	.0252	5.38%	.0274	4.83%
Unnormalised - Eq. (7) and (8)	**.0237**	**4.69%**	**.0247**	**4.55%**

C. Shortcomings

The effectiveness of LDA relies on an accurate estimation of the scatter matrices $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$. As the total variability observed in a set of training i-vectors is given by $\boldsymbol{S}_T = \boldsymbol{S}_B + \boldsymbol{S}_W$, the estimation of the between- and within-speaker scatters becomes a suitable breakdown of the total variability. In the context of the i-vector framework, however, the current method for estimating the scatter matrices neglects a common issue with regards to speaker verification—the use of an insufficient training dataset. Here, the term *insufficient* refers to a dataset in which each speaker does not provide at least one speech sample from

TABLE II
NIST SRE'08 RESULTS USING LDA MATRICES TRAINED FROM AN INSUFFICIENT DATASET (DISJOINT-SOURCE) AND A SUFFICIENT DATASET (MULTI-SOURCE) IN WHICH SPEAKERS PROVIDE UTTERANCES FROM ONLY ONE OR ALL SOURCES OF INTEREST, RESPECTIVELY

	tel-mic		int-tel	
LDA Training Dataset	DCF	EER	DCF	EER
Disjoint-source	.0222	5.17%	.0281	6.28%
Multi-source	**.0191**	**4.09%**	**.0243**	**4.92%**

each source of interest. The following analysis and discussion highlights how an insufficient training dataset leads to the sub-optimal breakdown of the variability in the i-vector space.

The within-speaker scatter aims to capture speaker-extrinsic sources of variation including variation attributed to different speech acquisition methods. The suppression of such variability is crucial for reducing errors in cross-source trials during system evaluation. A training dataset absent of multi-source utterances from each speaker fails to provide this information when explicitly calculating $\boldsymbol{S}_W$ via (8). In not accounting for this cross-source variability in the within-speaker scatter, it is consequently observed as between-speaker variability and is, therefore, maximized through LDA optimization.

1) Sufficient Multi-Source Training Data: To illustrate the dependency of the within-speaker scatter on a sufficient dataset, a small LDA training dataset[1] was compiled from the NIST 2010 corpus in which each speaker provided at least one telephone, microphone and 3-min interview-sourced utterance. Based on the experimental configuration detailed in Section V, two LDA matrices were used in the evaluation of the SRE'08 corpus: a *multi-source* matrix for which a speakers' multi-source utterances were recognized as belonging to the same speaker and a *disjoint-source* matrix trained such that utterances from different sources were assumed to originate from disjoint speaker sets. It should be noted that in ignoring the cross-source speaker information, the separation of observations from the same speaker class may be increased through LDA, thus implicitly disadvantaging the disjoint-source LDA matrix. This protocol, however, still provides insight into the value of cross-source speaker information through the comparison of LDA matrices trained using the same set of i-vectors. WCCN was excluded from the i-vector framework for these trials to allow for a more direct analysis of LDA. Results in Table II demonstrate that considerable benefits were found through the use of the multi-source LDA matrix as opposed to the disjoint-source matrix. This illustrates that the existence of information regarding source-related variability in the training dataset holds considerable value during LDA optimization.

2) Real Data Analysis: The relatively poor performance offered by the disjoint-source LDA matrix in Table II can be explained with the aid of Fig. 1. This figure depicts the distribution of telephone, microphone, and interview-sourced i-vectors when projected into a two-dimensional (a) PCA space and (b) LDA space trained via (7) and (8). These spaces were estimated using the female set of training i-vectors described in Section V in which each speaker provided observations from only one of the sources of interest.

[1]The dataset consisted of 3462 segments from 140 male speakers and 3852 segments from 152 female speakers.

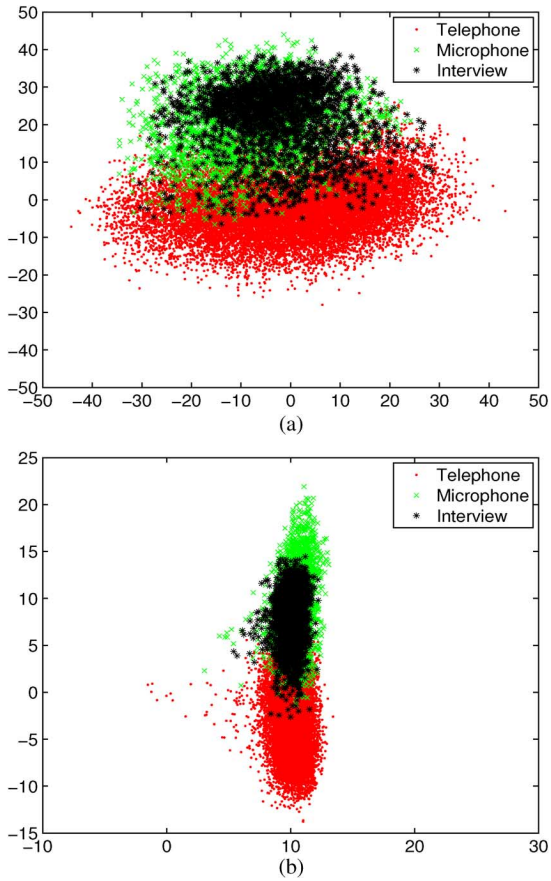

Fig. 1. Projection of female i-vectors into 2D PCA and LDA space. (a) 2D-PCA projection. (b) 2D-LDA projection.

Fig. 1(a) depicts the distribution of telephone, microphone, and interview-sourced i-vectors after 2D-PCA projection. It can be observed that the telephone i-vectors are more prominent in the lower half of the plot while the microphone and interview i-vectors fall in the upper half. It is apparent, therefore, that each speech source has a corresponding i-vector mean. This can be explained by the nature of *total* variability subspace which inherently leads to i-vectors capturing the variation attributed to the speech source. It is the role of LDA to suppress this variation. It can be observed in Fig. 1(b), however, that while attempting to maximize between-speaker variation, the standard LDA process also emphasizes the distinction between speech sources. In fact, the relative variation observed on the y-axis of the plot is far greater than the x-axis making source-related variation the greatest contributor of variability in the LDA-reduced i-vector space. As will be demonstrated, this distinction between sources is due to an inaccurate estimation of the between- and within-speaker scatters from an insufficient training dataset.

3) Graphical Interpretation: The effect of an insufficient training dataset on the scatters used in the LDA optimization can be further explained with the aid of Figs. 2 and 3. These figures depict the i-vector space as occupied by a training set of three speakers (depicted as unique shapes) taken from the three speech sources of interest along with the mean i-vector $\boldsymbol{\mu}$ and source-conditioned i-vector means $\boldsymbol{\mu}_{\{\text{tel,mic,int}\}}$.

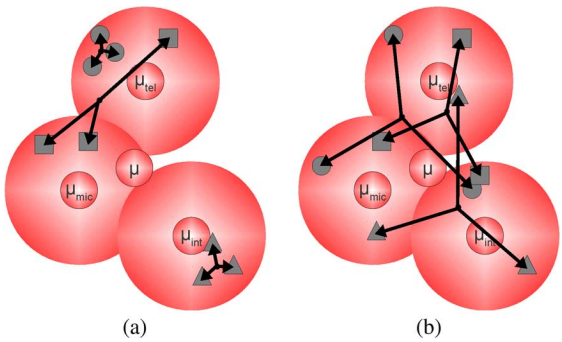

Fig. 2. Vectors used to calculate the within-speaker scatter from a typical and desired dataset. (a) Typical dataset. (b) Desired dataset.

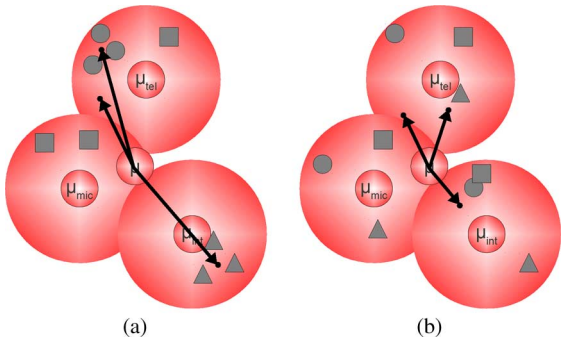

Fig. 3. Vectors used to calculate the between-speaker scatter from a typical and desired dataset. It can be observed that the vectors in (b) are a more appropriate approximation of between-speaker scatter than the vectors in (a). (a) Typical dataset. (b) Desired dataset.

Fig. 2(a) depicts as bold lines, the vectors that would be used to calculate the within-speaker scatter from a typical dataset in which cross-source variation is not sufficiently represented. In contrast, Fig. 2(b) illustrates the vectors desired from a three-speaker dataset in which each speaker provides an utterance from each source, thus, allowing the cross-source variability to be accurately estimated as part of the within-speaker scatter. As $\boldsymbol{S}_T = \boldsymbol{S}_B + \boldsymbol{S}_W$, the consequence of not observing cross-source variability as part of the within-speaker scatter results in its adverse influence on the between-speaker scatter. Fig. 3(a) illustrates this influence of source-related variation on the between-speaker scatter. It is the maximization of this variation during the LDA eigenvalue decomposition that attributes to the distinction of sources observed in Fig. 1(b). Desired is a between-speaker scatter absent of source-related variation estimated from a training dataset with sufficient multi-source speaker utterances [depicted in Fig. 3(b)].

In practice, a *desired* or sufficient LDA training dataset is difficult to acquire. Consequently, an suitable LDA algorithm must be employed that effectively counteracts the aforementioned shortcomings of the standard approach to LDA.

IV. Source-Normalized Linear Discriminant Analysis

Source-normalized LDA addresses the issues highlighted in the previous section regarding the poor estimation of LDA scatter matrices from insufficient resources in the i-vector framework for speaker recognition.

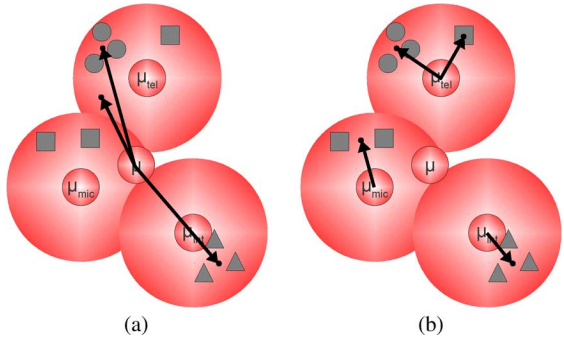

Fig. 4. Vectors used to calculate the between-speaker scatter from an insufficient dataset in the standard and source-normalized LDA approaches. (a) Standard LDA. (b) SN-LDA.

A. Algorithm

As discussed in Section III-C, the between-speaker scatter calculated using the standard LDA approach is adversely influenced by source-related variation when estimated from an insufficient training dataset. This influence can be reduced by estimating the between-class scatter using vectors that are calculated with respect to their corresponding source mean. The process is depicted in Fig. 4 such that the vectors used to calculate the scatter in the source-normalized approach are considerably less affected by source variation than the standard LDA algorithm. The source-normalized $\widehat{\boldsymbol{S}}_B$ is given by the sum of the source-conditioned between-class covariance matrices such that

$$\widehat{\boldsymbol{S}}_B = \sum_{\text{src}} \boldsymbol{S}_B^{\text{src}} \tag{11}$$

$$\boldsymbol{S}_B^{\text{src}} = \sum_{s=1}^{S_{\text{src}}} N_s (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})^t \tag{12}$$

where $\boldsymbol{\mu}_{\text{src}} = (1/N_{\text{src}}) \sum_{n=1}^{N_{\text{src}}} \boldsymbol{w}_n^{\text{src}}$ and N_{src} designates the number of utterances taken from source src, and similarly, S_{src} the number of speakers from source src. It can be noted that in this approach, a speakers' utterances from different sources are assumed to belong to disjoint speakers. For instance, segments for the speaker represented by the square in Fig. 4(b) are treated as though they originate from two different speakers. This assumption implicitly places a requirement on the dataset used for SN-LDA—i-vectors corresponding to an individual speaker should be representative of a single speech source. As previously stated, this can be viewed as a typical dataset in the field of speaker recognition. Not adhering to this assumption will lead to increased separation between same-speaker segments from different sources (this aligns with the "disjoint-source" discussion in Section III-C1).

The main novelty of this approach comes from the manner in which the within-speaker scatter is calculated. Section III-C highlighted the fact that a training dataset without cross-source speaker utterances results in the sub-optimal representation of the within-speaker scatter. In the SN-LDA approach, the within-speaker covariance matrix is estimated as the variability not captured by the between-speaker covariance matrices rather than through the explicit calculation (8). Assuming $\widehat{\boldsymbol{S}}_B$ no longer captures the within-class variation due to different

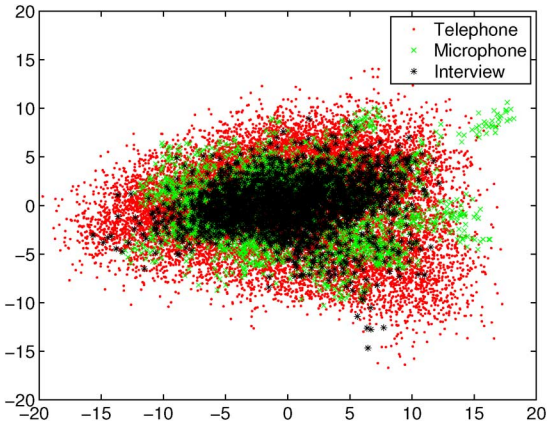

Fig. 5. Projection of female i-vectors into 2D SN-LDA space.

speech sources, it can be assumed that this variation is represented by the remaining variation observable in the i-vector space. Specifically, estimation of the within-speaker scatter is based on the knowledge that the total variance in the training i-vectors is given by $\boldsymbol{S}_T = \sum_{n=1}^{N} \boldsymbol{w}_n \boldsymbol{w}_n^t$ (the i-vector mean $\boldsymbol{\mu}$ is not specified here due to the zero-mean assumption for factor analysis) and is composed such that $\boldsymbol{S}_T = \boldsymbol{S}_W + \widehat{\boldsymbol{S}}_B$. Thus, the within-class scatter is given by

$$\boldsymbol{S}_W = \boldsymbol{S}_T - \widehat{\boldsymbol{S}}_B. \tag{13}$$

The use of $\boldsymbol{S}_W$ and $\widehat{\boldsymbol{S}}_B$ from (13) and (11) in the LDA optimization will be referred to as source-normalized LDA (SN-LDA) for the remainder of this study.

B. Analysis

The ability of the SN-LDA algorithm to suppress source-related variation was analyzed following the same procedure outlined in Section III-C-2. The same female set of telephone, microphone and interview-sourced i-vectors was projected into a two-dimensional SN-LDA space prior to being plotted in Fig. 5. In contrast to the analogous plot when using standard LDA in Fig. 1(b), it is apparent from Fig. 5 that SN-LDA is effective in reducing the source-related variation observable in the i-vector space.

C. Source-Normalized-and-Weighted LDA

SN-LDA was originally proposed along with a variant termed source-normalized-and-weighted (SNAW) LDA [7]. SNAW-LDA extended on SN-LDA by weighting the source-normalized between-speaker scatter matrices, $\boldsymbol{S}_B^{\text{src}}$, by the proportion of total LDA training i-vectors that corresponded to speech source src via

$$\overline{\boldsymbol{S}}_B = \sum_{\text{src}} \frac{N_{\text{src}}}{N} \boldsymbol{S}_B^{\text{src}}. \tag{14}$$

It was determined in a later study, however, that this weighting procedure was not conducive to the generalization of the resulting LDA transform to unseen datasets [8]. Specifically, the heuristic-based weighting found via (14) provided few significant improvements over SN-LDA and, even when optimized through an exhaustive search, selected weights were highly dependent on the evaluation dataset for which they were tuned. For this reason, SNAW-LDA is not considered in this work and, instead, the more straightforward SN-LDA is utilized.

D. Comparison to Fully Weighted LDA

An LDA algorithm was recently presented in [5] in which the authors calculated $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ as an empirically weighted average of within- and between-speaker scatter matrices individually estimated from telephone and microphone speech. To aid in discussion, this approach will be termed *fully weighted (FW) LDA*. The SN approach used in this study differs in several aspects. Most significantly is that $\boldsymbol{S}_W$ is not calculated explicitly via (8) nor is it composed of weighted within-speaker scatter matrices. Further, the microphone and telephone between-speaker covariance matrices in [5] were calculated under the assumption of the i-vector mean $\boldsymbol{\mu} = 0$ thus potentially capturing variation due to different speech sources. It should also be noted that compared to (7) and (8) in this work, the LDA algorithm in [5] normalizes $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ with respect to the number of utterances via (9) and (10). This normalization, however, was empirically found to harm FW-LDA performance based on the protocol defined in Section V. Section VI-A compares the performance of the standard, FW- and SN-LDA techniques.

V. Experimental Protocol

Evaluations were performed on the recent NIST 2008 and 2010 SRE corpora. Speech segments in these corpora have an average duration of 2.5 minutes with the exception of a portion of the SRE'10 interview speech segments which contain 8 minutes of audio. Results are reported for four evaluation conditions on each corpora with particular focus on mis-matched trials. The SRE'10 conditions correspond to det conditions 2–5 in the evaluation plan [11], and include *int-int* (different microphones), *int-mic*, *int-tel*, and *tel-tel* (English-only) source conditions, respectively. Performance was evaluated using the equal error rate (EER) and a normalized minimum decision cost function (DCF or C_{det}) calculated using $C_{\text{miss}} = 1$, $C_{\text{FA}} = 1$ and $P_{\text{tar}} = 0.001$ for the core SRE'10 results. The *extended* trial set was used to ensure sufficient impostor trials to estimate the minimum DCF. The number of trials ranged from 416 119 (tel-tel) to more than 2.8 million (int-int) with 0.5–1.7% target trials. For SRE'08, det conditions 3–5 and 7 [12] were evaluated corresponding to *int-int* (different microphones), *int-tel*, *tel-mic*, and *tel-tel* (English-only) trials, respectively. The DCF was calculated using $C_{\text{miss}} = 10$ and $P_{\text{tar}} = 0.01$ for SRE'08 results and the short utterance conditions of SRE'10. The number of trials ranged from 8444 for the *tel-mic* condition to 32 318 in the interview-only condition with 7% target trials in the telephone-only condition ranging up to 34% in the interview-only trials. Gender-pooled results are reported throughout.

In all approaches, the number of LDA dimensions retained was evaluated in steps of 50 in order to minimize the average of $(C_{\text{det}}^{\min} + 10 \times \text{EER})$ across the evaluated conditions of SRE'10 and $(C_{\text{det}}^{\min} + \text{EER})$ for SRE'08. The weights for fully weighted LDA were determined in a similar manner resulting in $[P_{\text{tel}}, P_{\text{mic}}, P_{\text{int}}] = [0.2, 0.1, 0.7]$ for SRE'10 trials and [0.3, 0.1, 0.6] for SRE'08 trials. Due to the overlap of interview data between impostor datasets (used for total variability subspace

TABLE III
COMPARISON OF STANDARD, FULLY WEIGHTED (FW) AND SOURCE-NORMALIZED (SN) LDA ON THE SRE'08 AND SRE'10 CORE CONDITIONS

Corpus	LDA Algorithm	Optimised LDA Dim.	tel-mic		int-tel		int-mic		int-int		tel-tel	
			DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER
SRE'08	Standard LDA	200	.0237	4.69%	.0247	4.55%	—	—	.0120	3.24%	.0144	2.69%
	FW-LDA	200	.0211	4.56%	.0207	3.82%	—	—	**.0114**	3.38%	.0140	2.83%
	SN-LDA	200	**.0132**	**3.27%**	**.0148**	**2.82%**	—	—	.0122	**3.17%**	**.0135**	**2.61%**
SRE'10	Standard LDA	200	—	—	.5651	3.81%	.3699	2.20%	.5179	3.53%	.5107	3.17%
	FW-LDA	200	—	—	.5553	3.59%	**.3214**	**2.06%**	**.4914**	**3.18%**	**.5079**	3.11%
	SN-LDA	200	—	—	**.4287**	**2.86%**	.3439	**2.06%**	.4959	**3.18%**	.5160	**3.07%**

training, LDA, WCCN and cosine kernel normalization) and evaluation datasets (see below), these weights were optimized directly on the evaluation dataset.

Speech activity detection (SAD) was implemented using a two-component Gaussian mixture model (GMM) trained on the log energy observed in the speech file. If the variance of the speech component was more than fives times that of the non-speech component, energy samples below the mean of the non-speech component were removed and the GMM retrained. This ensured that, in the presence of considerable silence, the speech component was representative of the higher energy samples and not of nearby samples corresponding to noise. No speech was found in the segment if the difference in component means was less than 4 or the speech mean was below 30. Alternatively, a speech threshold for the whole speech segment was set to the mean of the speech component minus 1.3 standard deviations. Parameters for this algorithm were determined empirically by listening to a wide selection of speech from multiple speech sources. Dual-SAD was used for the SRE'10 interview conversations such that an interviewee speech frame was retained if its normalized gain was more than 5 dB louder than the corresponding audio frame from the interviewer. Gender-dependent 2048-component UBMs 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 400-dimensional total variability subspace was trained for each gender from the same data used in UBM training. A gender-dependent dataset was compiled to serve the purpose of training the LDA and WCCN transforms as well as the impostor dataset used for cosine kernel normalization. This dataset was formed from the same corpora used to train the UBMs with additional interview data. Interview data for SRE'10 evaluations was taken from the follow-up corpus of the NIST 2008 SRE while interview data for SRE'08 evaluations was taken from a subset of the 3-minute interview segments of the NIST 2010 SRE corpus. There was no overlap between the speakers of interview segments used for system development and the intended evaluation corpus. The average number of utterances from telephone, microphone and interview speech samples was 15 178, 2665, and 1776, respectively, and consisted of an average of 2605, 88, and 119 speakers, respectively. Utterances from each speech source were selected to belong to disjoint speaker sets.

VI. RESULTS

This section evaluates the SN-LDA approaches described in Section IV across both NIST 2008 and 2010 SRE corpora. The robustness of SN-LDA is analyzed with respect to utterance duration, LDA dimensionality, LDA training dataset size and its potential in SVM-based classifiers. Finally, source-normalized WCCN is investigated as is data-driven source detection as a means of removing the requirement of source labeled data for SN-LDA. To aid in discussion, *under-resourced* conditions refers to conditions involving interview or microphone speech while *mis-matched* refers to conditions in which the training speech is sourced in a different manner to the testing speech (for instance, *tel-mic*).

A. Evaluation of SN-LDA

Table III presents the results from the evaluation of the source-normalized (SN) on the SRE'08 and SRE'10 corpora alongside those obtained when using standard and fully weighted (FW) LDA. The SRE'08 results show that FW-LDA, while developed for under-resourced conditions, provided improvements over standard LDA in the mis-matched evaluation conditions. SN-LDA further reduced verification errors relative to standard LDA, offering improvements of up to 44% in minimum DCF and 38% in EER under mis-matched conditions. SN-LDA also offered the best performance in *tel-tel* trials. This demonstrates that the within-speaker scatter can be reliably estimated as the total variation not represented by the source-conditioned between-speaker covariance matrices when speakers in the training dataset provide no multi-source examples.

Results from trials on the SRE'10 offered similar trends to those observed from the SRE'08 results. Namely, significant performance improvements were offered through the SN-LDA under mis-matched evaluation conditions compared to standard LDA. In contrast to the SRE'08 findings, however, FW-LDA provided marginally better minimum DCF statistics than SN-LDA with the exception of the *int-tel* condition. In this case, SN-LDA demonstrated a relative improvement of greater than 22% and 20% in minimum DCF and EER over both standard and FW-LDA. In general, FW-LDA offered similar benefits to SN-LDA but only when telephone data was not trialled in combination with microphone or interview-sourced speech—in this case, SN-LDA significantly outperformed the alternate approaches to LDA. The reason for this observation lies in the ability of SN-LDA to remove the severe source variation observed in the i-vector space (see Figs. 1 and 5) while this variation is not suppressed via FW-LDA.

TABLE IV
COMPARISON OF STANDARD AND SOURCE-NORMALIZED (SN) LDA APPROACHES EVALUATED USING 10-SECOND TRAIN AND TEST SEGMENTS ON BOTH SRE'08 AND SRE'10 CORPORA

Corpus	LDA Algorithm	Optimised LDA Dim.	tel-mic		int-tel		int-mic		int-int		tel-tel[1]	
			DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER
SRE'08	Standard LDA	250	.0655	18.54%	.0751	19.95%	—	—	**.0594**	13.24%	**.0572**	12.58%
	SN-LDA	100	**.0618**	**17.33%**	**.0706**	**16.64%**	—	—	.0608	**12.64%**	.0576	**12.09%**
SRE'10	Standard LDA	300	—	—	.0687	17.40%	.0595	14.08%	.0666	16.84%	.0625	15.00%
	SN-LDA	300	—	—	**.0656**	**16.12%**	**.0573**	**13.72%**	**.0642**	**16.57%**	**.0603**	**14.37%**

TABLE V
SVM PERFORMANCE ON NIST SRE'10 COMPARING STANDARD LDA AND SN-LDA

LDA Training Regime	int-tel		int-mic		int-int		tel-tel	
	DCF	EER	DCF	EER	DCF	EER	DCF	EER
Standard LDA	.4472	**3.43%**	.4042	2.56%	.4990	3.91%	.4527	**3.26%**
SN-LDA	**.4249**	**3.43%**	**.3509**	**2.36%**	**.4675**	**3.83%**	**.4399**	3.27%

The above findings demonstrate that SN-LDA significantly improved performance over standard LDA when subject to under-resourced and mis-matched conditions while also offering some benefit to the commonly encountered *tel-tel* conditions.

B. Detection Performance for Short-Utterances

The previous section demonstrated the effectiveness of SN-LDA when using speech of more than 2 minutes in duration. In order to determine whether the technique is effective when dealing with short utterances, standard LDA was compared to SN-LDA in the evaluation of 10-second train and 10-second test conditions of SRE'08 and SRE'10 data. All conditions were evaluated by extracting features from 10 seconds of active speech from segments in the common evaluation protocols.[2] The first minute of audio was removed prior to feature extraction so as to reduce overlap in introductory speech between segments. It should be noted that short utterances were utilized only in the evaluation data while the total variability subspace, LDA, WCCN, and normalization datasets remained the same as used in Section VI-A. Results from these evaluations are detailed in Table IV. It can be observed from these results that SN-LDA provided improved performance compared to standard LDA in all evaluation conditions of both SRE'08 and SRE'10. This was particularly evident in mis-matched conditions that involved both telephone speech and interview or microphone speech in which the average relative minimum DCF and EER improvements were 6% and 10%, respectively. These findings demonstrate the robustness of the SN-LDA algorithm to reduced speech length.

C. Detection Performance Using Support Vector Machines

Support vector machines (SVMs) are commonly employed in the i-vector framework as a substitute for cosine kernel distance scoring [1]. This framework involves training an SVM for each enrolled client using their corresponding i-vectors as features and obtaining a detection score as the distance that a testing i-vector lies from the SVM hyperplane of the claimed identity. The SVM system in this work was based on the libSVM package and the cosine kernel which was effectively implemented by the transformation of i-vectors via (4). Experiments were performed using SRE'10 to determine whether SN-LDA could provide benefit to SVM-based classification performance. The background dataset used for SVM training was compiled using one i-vector per speaker from the LDA training dataset. Results from the SVM trials are presented in Table V. It can be noted that when compared to results in Section VI-A, SVMs outperformed the cosine distance measure in terms of minimum DCF in most conditions but this often came at the cost of a higher EER. Nonetheless, SN-LDA provided improved results over standard LDA in almost all performance metrics when used in conjunction with SVM-based classification. While these improvements were somewhat limited in terms of EER, the minimum DCF of the *int-mic* trials found a 13% relative improvement. These results demonstrate that SVM-based speaker recognition using i-vectors can be improved using SN-LDA rather than standard LDA.

D. Case of Sufficient Training Data

Section III-C1 illustrated that in order for standard LDA to robustly estimate the within-speaker scatter from a training dataset, each speaker in the dataset must provide utterances from each source of interest. This was analyzed by recognizing or ignoring the cross-source speaker information during LDA training from such an ideal dataset that was sourced from SRE'10. The resulting LDA matrices were then evaluated on SRE'08. Using the same LDA training dataset and again removing WCCN from the i-vector framework, SN-LDA was evaluated to determine whether it could offer the same performance as standard LDA when trained using an ideal dataset. This objective is expected to be difficult to achieve, however, due to the implicit assumption that segments from different sources correspond to disjoint speaker sets in SN-LDA.

Table VI details the performance offered by standard LDA when using a disjoint-source matrix (learned by assuming speaker sets for each source were disjoint) and a multi-source matrix (i.e., the cross-source speaker information was retained) compared to that offered by SN-LDA. As can be expected, the multi-source LDA matrix provided the best performance. Nonetheless, SN-LDA offered improvements

[2]Note that the 10-second *tel-tel* results were not found using the official 10 sec-10 sec protocols but truncated features from the core conditions.

TABLE VI
NIST SRE'08 RESULTS COMPARING STANDARD LDA MATRICES TRAINED FROM A SUFFICIENT DATASET, IN WHICH A SPEAKERS MULTI-SOURCE INFORMATION WAS IGNORED (DISJOINT-SOURCE) OR ACKNOWLEDGED (MULTI-SOURCE), TO SN-LDA IN WHICH MULTI-SOURCE INFORMATION IS IMPLICITLY IGNORED

	tel-mic		int-tel	
LDA Training Regime	DCF	EER	DCF	EER
Disjoint-source	.0222	5.17%	.0281	6.28%
Multi-source	**.0191**	**4.09%**	**.0243**	**4.92%**
SN	.0193	4.63%	.0245	5.38%

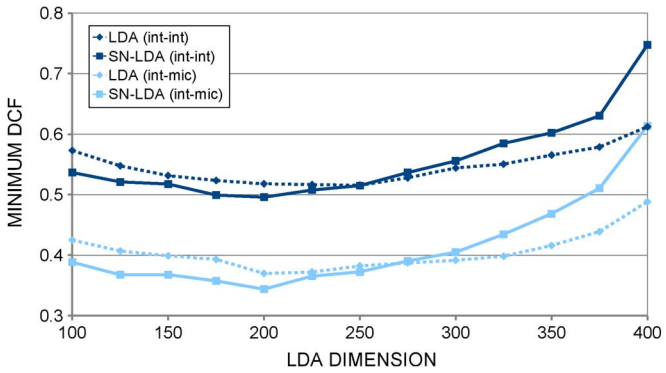

Fig. 6. Sensitivity of $C_{\text{det}}^{\text{min}}$ to LDA matrix dimension in under-resourced trial conditions of SRE'10 when using Standard and SN-LDA.

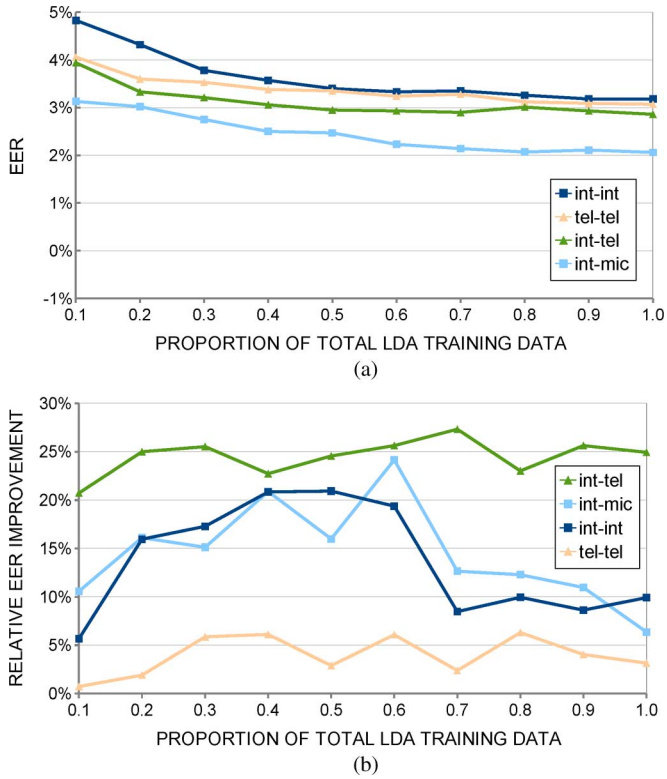

Fig. 7. Effect of limited LDA training data in terms of SRE'10 EER. (a) EER when using SN-LDA. (b) Relative EER improvement from SN-LDA over standard LDA.

over disjoint-source results with the minimum DCF being comparable to the ideal multi-source trials. As noted in Section III-C-1, ignoring the available cross-source information from the dataset implicitly handicaps the disjoint-source and SN-LDA matrices. This is because the variation observed between different sources of the same speaker consequently contribute to the estimated between-speaker scatter and is therefore maximized in the LDA optimization. It is worth reiterating, therefore, that a suitable dataset for SN-LDA has *no* cross-source segments from speakers—the most commonplace dataset in speaker recognition research.

E. Reducing LDA Matrix Dimensionality

Classification performance in the i-vector framework varies along with the dimensionality of the LDA matrix [2]. It was noted during system development that the performance from trials involving only under-resourced speech conditions (i.e., non-telephone speech) varied considerably along with the LDA dimension. Fig. 6 illustrates this in a plot of minimum DCF as a function of LDA dimensionality for the *int-int* and *int-mic* conditions of the SRE'10 when using both standard and SN-LDA. Fig. 6 shows the preference for lower dimensionality in the case of SN-LDA compared to standard LDA. Similar trends were observed in terms of EER while this was not as obvious in trials involving telephone speech. Instead, more consistent improvements to DCF and EER were observed by SN-LDA over standard LDA throughout the range of LDA dimensions. This relatively high dependency of performance from under-resourced trials on LDA dimension is expected to be attributed to the relatively few speakers from which to accurately estimate the between-speaker scatter. It is hypothesized that a training dataset more evenly distributed across sources may result in less sensitivity to LDA dimension when using SN-LDA.

F. Effect of Limited LDA Training Data

The following experiments aim to determine how dependent SN-LDA is on the quantity of training data and whether it maintains a consistent relative improvement over standard LDA as this quantity is reduced. Fig. 7 illustrates (a) the effect of limited LDA training data on EER performance when using SN-LDA and (b) the relative EER improvement offered by SN-LDA over standard LDA as the size of the training data is reduced from all available data (right-hand side of the plot) down to a proportion of 0.1 (left-hand side). Each of the four conditions from SRE'10 are represented in the Figure. Removal of segments from the dataset was done using equal proportions from telephone, microphone and interview speech sources on a per speaker basis. This differs from the analysis performed in [8] where only telephone-sourced speech was removed.

Fig. 7(a) illustrates, first, that as LDA training data was reduced, the EER incurred a subtle increase. It is interesting to observe that the EER of the cross-source conditions was consistently lower than in same-source conditions. There are a number of reasons for this observation. SN-LDA in this case has explicitly targeted differences between speech sources such that intra-source differences, such as handset, microphone and transmission channel, may subsequently have become relatively more dominant. The use of the SRE'08 followup dataset in LDA training is likely to have aided in reducing this variation in the interview-based trials since this dataset contains utterances from every speaker-microphone combination. It could be expected, then, that the *int-int* trials would provide the lowest EER, however, this was not the case. A difference between

SRE'08 and SRE'10 interview segments that may help explain this phenomena is the speech activity detection (SAD) process. In the case of SRE'08, SAD indices were provided by NIST while a dual-SAD algorithm was utilized for SRE'10 segments (see Section V). The *int-mic* condition consistently surpassed performance from alternate conditions, which may be attributed to the improved speech signal compared to telephone-sourced speech and the fact that interview and microphone i-vectors appeared to reside in a similar area in the i-vector space prior to LDA [Fig. 1(a)].

Fig. 7(b) indicates that SN-LDA consistently provided EER improvements over the use of standard LDA irrespective of training dataset size. Further, it can be noted that the relative EER improvement was considerably robust to variation in dataset size for trials involving telephone speech. When moving from a dataset proportion of 0.2 to 0.1, the relative improvement of trials involving only microphone and/or interview speech dropped sharply. This is likely due to the lack of speakers from these sources in the training dataset (the dataset proportion of 0.1 provided less than 12 speakers from each of these sources on average). Finally, the relative EER improvement doubled for the *int-int* and *int-mic* conditions when the proportion was reduced from 0.7 to 0.6. This was due to a considerable drop in performance from standard LDA. It was hypothesized that i-vectors from a particularly useful set of speakers were pruned from the training data at this point. These observations suggest that the relative EER improvement of SN-LDA over standard LDA is robust to variation in dataset size assuming the number of speakers representing each speech source is not severely limited.

G. Source-Normalised WCCN

Until now, the source-normalized regime for estimating the between and within-speaker scatter matrices has been incorporated only during the LDA process of the i-vector framework. The effectiveness of the WCCN process that follows LDA in this framework also relies on the accurate estimation of the within-speaker scatter from LDA-reduced i-vectors. The following experiments utilize algorithm (13) to calculate this within-speaker scatter in the same manner that $\boldsymbol{S}_W$ was calculated for the SN-LDA approach. This process will be termed SN-WCCN. The aim here was to determine first, whether scatter source-normalization was feasible in the WCCN process and second, the degree that WCCN assists classification performance after standard LDA and SN-LDA. Results on the mis-matched conditions of SRE'10 when using no WCCN, WCCN, and SN-WCCN are detailed in Table VII.

The results in Table VII show a 5%–19% relative improvement in performance statistics when introducing WCCN to a framework based on standard LDA, thus illustrating the importance of WCCN in the i-vector framework. Source-normalized (SN) WCCN in conjunction with standard LDA provided a relative improvement of up to 20% over the standard WCCN approach, thus illustrating that source-normalization is highly suited to the calculation of the WCCN matrix. As in the case of standard LDA, the use of WCCN along with SN-LDA appears necessary in order to minimize detection errors. In contrast, however, there was no observable gain from SN-WCCN over WCCN when used in conjunction with SN-LDA. It is hypothesized that this is due to SN-LDA already having removed the effects of source-related variation from the LDA-reduced i-vectors used in the estimation of the WCCN matrix. Comparing the best performance statistics from standard and SN-LDA, it can be seen that employing source-normalization in the LDA space offers some advantage over its use in the later WCCN process.

TABLE VII
NIST SRE'10 CROSS-SOURCE TRIALS USING NO WCCN, WCCN, AND SOURCE-NORMALIZED (SN) WCCN IN CONJUNCTION WITH STANDARD LDA AND SN-LDA

	int-tel		int-mic	
LDA Training Regime	DCF	EER	DCF	EER
LDA (without WCCN)	.6060	4.31%	.4562	2.53%
LDA + WCCN	.5651	3.81%	.3699	2.20%
LDA + SN-WCCN	**.4537**	**3.18%**	**.3537**	**1.84%**
SN-LDA (without WCCN)	.5303	3.16%	.4949	2.45%
SN-LDA + WCCN	.4287	**2.86%**	.3439	2.06%
SN-LDA + SN-WCCN	**.4264**	**2.86%**	**.3436**	**2.04%**

H. Data-Driven Segmentation of LDA Training Data

The effectiveness of SN approach to LDA has been demonstrated in preceding experiments. This approach does not require source-labeled enrolment or test speech, and instead requires only that segments in the LDA training dataset be source-labeled. Although defining only three common sources in this paper has proven effective, it seems plausible that further segmentation of these source subsets may lead to improved classification performance (for example, treating cellular and landline telephone speech as originating from different sources). Depending on the datasets on hand, the manual source-labeling of data to this extent may not always be plausible. It would be beneficial, therefore, to alleviate the need for source-labeled LDA training data while still benefiting from the advantages of SN-LDA. One method of accomplishing this is through the data-driven detection of the source, or more accurately, the data-driven segmentation of the LDA training dataset. An investigatory experiment was performed to determine the validity of data-driven dataset segmentation and whether corresponding performance statistics were able to maintain an improvement over the standard approach to LDA.

Data-driven segmentation of the LDA training dataset was performed in the following manner. A five-dimensional PCA space was learned using all i-vectors in the training dataset. These same i-vectors were then projected into this PCA space. Finally, k-means clustering was employed to segment the PCA-projected i-vectors into four clusters. Each i-vector in the original LDA training dataset was then assigned a source [1]–[4] based on the cluster to which they were assigned. This configuration was empirically found to provide best performance when varying the number of PCA dimensions and k-means centroids. Table VIII details the SRE'10 results from mis-matched evaluation conditions when using standard LDA and SN-LDA trained using a dataset source-labeled in a *supervised* manner (as used in preceding experiments) and a dataset segmented in a *unsupervised*, data-driven manner.

It can be observed from the results in Table VIII that unsupervised segmentation of the LDA training dataset provided signifi-

TABLE VIII
USE OF DATA-DRIVEN SEGMENTATION OF THE LDA TRAINING DATASET FOR THE PURPOSE OF SN-LDA IN THE NIST SRE'10 CROSS-SOURCE TRIALS

	int-tel		int-mic	
LDA Training Regime	DCF	EER	DCF	EER
Standard LDA	.5651	3.81%	.3699	2.20%
SN-LDA (unsupervised)	.4305	3.23%	**.3276**	2.36%
SN-LDA (supervised)	**.4287**	**2.86%**	.3439	**2.06%**

cant improvements over standard LDA for *int-tel* trials. In terms of minimum DCF and EER these relative improvements were 24% and 15%, respectively. It is interesting to note that the performance offered through unsupervised SN-LDA was broadly comparable to a supervised approach while there remains room for improvement in terms of EER. These results indicate that data-driven segmentation of the LDA training dataset based on the observation of clusters in the i-vector space is a viable alternative to manually source-labeling data.

VII. DISCUSSION

This section aims to review some historical developments that relate to the work presented here and outlines the direction of subsequent research paths. In the early years of the NIST text-independent speaker recognition evaluations, the strong influence of the handset microphone type (electret or carbon-button) on the detection score was observed. To counteract this influence, score normalization techniques such as H-norm [13] were developed. Because of the large difference in performance in same-number and different-number trials [14], more general compensation techniques in feature space were investigated, such as feature mapping [15]. Common to both H-norm and feature mapping was the requirement of channel-labeled conversations for training, where the channels were chosen from a discrete set. New techniques then attempted to generalize the channel variability found in training conversations to a continuous space, of which the most celebrated ones are joint factor analysis (JFA) pioneered by Kenny [16] and Brümmer [17], and nuisance attribution projection (NAP) [18], [19]. While NAP models the "nuisance" directions of session variability in SVM supervector space, JFA models the spaces in which the majority of speaker variability and session variability are observed-either in GMM supervector space or feature space [20]—alleviating the need for explicit channel labels. In the i-vector approach, which was pioneered during the JHU workshop following the NIST 2008 SRE by Dehak *et al.* [1], the explicit channel/speaker space difference was removed by modelling all possible variation using JFA techniques, and session compensation was left to the discriminating powers of LDA and WCCN. However, with the introduction of source variation—a new type of supervised "channel"—in NIST SRE 2005, namely microphone (instead of telephone) recordings of both telephone conversation and interview (SRE 2008) speaking styles, a plethora of evaluation conditions were possible. The new conditions received "core condition" status in NIST SRE 2008, drawing the attention of a larger group of researchers. A trivial approach, using the explicit source information in train and test segments by making detectors for all combinations under evaluation, proved laborious in SRE 2010 where again new source combinations were introduced.

Although our preference has always been for minimal dependence of the speaker recognition system on source information [21], the work presented in this paper attempts to treat utterances from all speech sources equally in the system at run time, thus making it possibly more robust against new source combinations. As a result, the adverse effect of the source that was observed in the i-vector space (see Fig. 1) was significantly reduced by introducing SN-LDA. SN-LDA still uses explicit source labeling during system development (although it was demonstrated in Section VI-H that an unsupervised clustering approach also provided improvements over the standard LDA approach), and the source labels are discrete. In a different space, a parallel may be drawn with feature mapping [15] in which channels are also explicit and discrete in training. Historical developments suggest therefore that, as JFA can be viewed as a continuous generalization of feature mapping, a JFA-style approach to the source variation in the i-vector framework may lead to a continuous representation of the source variation, thus leading to more robust speaker recognition systems.

VIII. CONCLUSION

The source-normalized LDA technique for the i-vector framework for speaker recognition was analyzed in this study. The shortcomings of the standard LDA algorithm for enhancing speaker discrimination were illustrated. These included the influence of variation related to speech source on the between-speaker covariance matrix and an incomplete representation of the within-speaker scatter due to commonly insufficient cross-source utterances per speaker in the LDA training dataset. SN-LDA was shown (with the aid of real data analysis) to reduced the influence of source variation on the between-speaker scatter through normalization of the i-vectors with respect to the source-conditioned i-vector means. The within-speaker scatter was calculated as the residual variation not captured by the source-normalized between-speaker scatter matrices thus improving estimation of the scatter from insufficient resources.

When evaluated on recent NIST 2008 and 2010 SRE corpora, SN-LDA demonstrated significant improvements of 44% in minimum DCF and 38% in EER relative to the standard LDA approach for mis-matched trial conditions and conditions for which limited system development speech was available. SVM classification of i-vectors projected into the SN-LDA space was also found to outperform the use of standard LDA. When subject to adverse conditions such as limited training and testing speech duration as well as reduced LDA training dataset quantity, SN-LDA was found to consistently outperform the standard approach to LDA. Source normalized WCCN provided significant benefit when used in conjunction with standard LDA. Finally, data-driven segmentation of the LDA training dataset provided a means of alleviating source-labels while still allowing SN-LDA to provide considerable improvements over the standard approach to LDA.

ACKNOWLEDGMENT

The authors would like to thank O. Glembek from BUT for the making the Joint Factor Analysis Matlab demo available.

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," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 1471–1474.

[4] National Institute of Standards and Technology, *NIST Speaker Recognition Evaluation site* [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/

[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 Lang. Recognition Workshop*, 2010.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 980–988, Aug. 2008.

[7] M. McLaren and D. van Leeuwen, "Source-normalized-and-weighted LDA for robust speaker recognition using i-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5456–5459.

[8] M. McLaren and D. van Leeuwen, "To weight or not to weight—Source-normalized LDA for speaker recognition using i-vectors," in *Proc. Interspeech'11*, 2011, accepted for publication.

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

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, no. 1, pp. 42–54, 2000.

[11] "NIST 2010 SRE evaluation plan," National Inst. of Standards and Technol. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/, ,

[12] "NIST 2008 SRE Evaluation Plan," National Inst. of Standards and Technol. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[13] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[14] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective," *Speech Commun.*, vol. 31, pp. 225–254, 2000.

[15] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2003, vol. 2, pp. 53–56.

[16] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 37–40.

[17] N. Brümmer, "Spescom DataVoice NIST2005 SRE system description," in *Proc. NIST Speaker Recognition Eval. Workshop*, 2004.

[18] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, pp. 97–100.

[19] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 629–632.

[20] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.

[21] D. A. van Leeuwen, "The TNO SRE-2008 speaker recognition system," in *Proc. NIST Speaker Recognition Eval. Workshop*, 2008.

Mitchell McLaren (M'10) received the BCompSysEng and the Ph.D. degree with the Speech, Audio, Image, and Video Technologies (SAIVT), Queensland University of Technology (QUT), Brisbane, Australia, in 2006 and 2010, respectively. His Ph.D. research has concentrated on speaker verification using support vector machine techniques.

He has been with the Centre for Language and Speech Technology (CLST), Radboud University, Nijmegen, The Netherlands, since 2010 where he is currently in a post-doctoral role. In 2007, he was a visiting intern within the Laboratoire Informatique D'Avignon in Avignon, France.

Dr. McLaren was awarded the "Best Student Paper Award" at Interspeech 2008 and the "IEEE 2009 Spoken Language Processing Student Grant" at ICASSP 2009.

David A. van Leeuwen received the Ir. (M.S.) degree from the Delft University of Technology, Delft, The Netherlands, in 1984 and the Dr. (Ph.D.) degree from the University of Leiden, Leiden, The Netherlands, in 1993.

He was with TNO Human Factors since 1994 and is with Radboud University, Nijmegen, The Netherlands, since 2008. He has worked in various areas in speech technology, with special interests in evaluation of systems. Examples of evaluations he organized are the EU FP5 SQALE project (1995), evaluating large-vocabulary speech recognition systems in four languages, the NFI-TNO Forensic Speaker Recognition Evaluation with 11 international participants in 2003, and the N-Best evaluation of Dutch speech recognition systems in 2008. He has also successfully participated in various NIST Rich Transcription, Speaker and Language Recognition Evaluations. In 2008, he was appointed Professor at Radboud University and he coauthored the EU FP7 Marie Curie ITN project "Bayesian Biometrics for Forensics" (BBfor2), which he is currently leading. In recent years he has been focusing on speaker diarization and automatic speaker and language recognition, with special interest in highly accurate and efficient systems, forensic application scenarios, and calibration.