The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/94676

Please be advised that this information was generated on 2019-06-24 and may be subject to change.
SOURCE-NORMALISED-AND-WEIGHTED LDA FOR ROBUST SPEAKER RECOGNITION USING I-VECTORS

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 recently developed 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 speaker utterance extracted from a low-dimensional total variability subspace. Prior to classification using a cosine kernel, i-vectors are projected into an 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 classification performance. A typical training dataset, however, does not consist of utterances acquired from all sources of interest (ie., telephone, microphone and interview speech sources) for each speaker. This has the effect of introducing source-related variation in the between-speaker covariance matrix and results in an incomplete representation of the within-speaker scatter matrix used for LDA.

Proposed is a novel source-normalised-and-weighted LDA algorithm developed to improve the robustness of i-vector-based speaker recognition under both mis-matched evaluation conditions and conditions for which insufficient speech resources are available for adequate system development. Evaluated on the recent NIST 2008 and 2010 Speaker Recognition Evaluations (SRE), the proposed technique demonstrated improvements of up to 31% in minimum DCF and EER under mis-matched and sparsely-resourced conditions.

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

1. INTRODUCTION

A speaker recognition framework based on i-vectors was recently developed by Dehak et al. [1, 2] offering superior recognition to the widely adopted joint factor analysis (JFA) approach [3]. This framework involves extracting i-vectors from a low-dimensional total variability subspace, enhancing discrimination using linear discriminant analysis (LDA) and within-class covariance normalisation (WCCN), and performing classification using a cosine kernel. While the total variability subspace is responsible for containing between-utterance variability in the i-vectors, it is ultimately the role of LDA to define the space in which speakers are discriminated from one another.

In the context of the i-vector framework, LDA attempts to find a reduced set of axes onto which i-vectors can be projected such that the within-speaker variability is minimised and the between-speaker variability is maximised. Within-speaker variability occurs due to the different transmission channels, microphones, acoustic environments, speaking styles, methods of speech acquisition, etc., that contribute the differences observed between utterances from the same speaker [4]. In contrast, between-speaker variation is due to the differences between true speaker characteristics and is the key variation to be maximised in the LDA process. LDA, therefore, relies on the accurate calculation of the scatter matrices in order to determine a set of axes optimised for speaker discrimination.

Speaker recognition using conversational telephony speech has long been the focus in the research field, resulting in an abundance of data available for system development and tuning. It is only in recent years that the NIST speaker recognition evaluations (SRE) [5, 6] have incorporated microphone and interview-sourced speech, both of which have limited resources available for system development. Consequently, robust speaker recognition is challenging when non-telephone speech is encountered during system evaluation, particularly in the case of mis-matched trials (ie., train on microphone speech and test on interview speech) [7, 3].

Insufficient speech resources directly reduces the effectiveness of LDA due to the inaccurate estimation of the scatter matrices from the training data. As detailed in this work, a typical training dataset, in which a speaker has only utterances available from a limited number of different sources, results in an incomplete representation of the within-speaker variability and a between-speaker scatter that is adversely influenced by source-related variation. Unless sufficient data representing this variability is available for LDA, the scatter matrices will not be optimal for task of speaker discrimination.

This paper presents a novel algorithm to robustly estimate the LDA scatter matrices from a training dataset in which few or no multi-source utterances are available per speaker, but multiple sources are available through different speaker collections. Telephone, microphone and interview-style speech sources are considered. A source-normalised-and-weighted (SNAW) average is used to estimate the between-speaker scatter, thereby, reducing the adverse bias of variation attributed to the speech source. The within-speaker scatter is given by residual of the total variability in the i-vector space. This has the effect of emphasising variation due to different speech sources in the within-speaker scatter and alleviates the need to gather a speakers’ utterances from multiple speech sources. The proposed approach is validated on the recent NIST 2008 and 2010 SRE’s.

This paper is structured as follows. Section 2 describes the i-vector framework for speaker recognition. Section 3 details the standard LDA algorithm and is followed by the proposal of SNAW-LDA in Section 4. The experimental protocol and corresponding results are given in Sections 5 and Section 6.

2. THE I-VECTOR FRAMEWORK FOR SPEAKER RECOGNITION

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 total variability 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) mean supervector,

\[ M = m + Tw, \]

where \( M \) consists of a speaker- and session-independent mean supervector \( m \) from the universal background model (UBM) and a mean offset \( Tw \). The supervector \( M \) is assumed to be normally distributed with mean \( m \) and covariance \( TT' \), where \( T \) is the low-rank, total variability subspace. This subspace is trained via factor analysis over a set of centralised Baum-Welch statistics [3] and represents the space in which the majority of between-utterance variability is observed. The low-rank vector \( w \) has a standard normal distribution \( N(0, 1) \) and is referred to as the i-vector. Extracting an i-vector from the total variability subspace is essentially a maximum-a-posteriori adaptation of \( w \) in the space defined by \( T \). An efficient procedure for the optimisation of the total variability subspace \( T \) and subsequent extraction of i-vectors is described by [3] and [2].

2.2. Inter-session Compensation

The subspace from which i-vectors are extracted bounds both speaker-intrinsic or speaker-extrinsic. Consequently, i-vectors in their raw form are not optimised 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 to find a reduced set of axes \( A \) that minimises the within-speaker variability observed in the i-vector space while simultaneously maximising the between-speaker variability. This process is covered in detail in Section 3 with a novel source-normalised-and-weighted (SNAW) LDA algorithm proposed in Section 4.

The secondary stage, within-class covariance normalisation (WCCN) [4], normalises the residual within-speaker variance remaining in LDA-reduced i-vectors. The WCCN matrix \( B \) is found through the Cholesky decomposition of \( W^{-1} = BB' \) where the within-class covariance matrix is calculated as,

\[ W = \frac{1}{S} \sum_{i=1}^{S} \sum_{s} (A'w_i^s - \mu_i)(A'w_i^s - \mu_i)^t, \]

where \( S \) is the number of speakers that each provide \( N_s \) i-vectors in the training dataset, and the mean of the LDA-reduced i-vectors from speaker \( s \) is equated as \( \mu_i = \frac{1}{N_s} \sum_{s=1}^{N_s} A'w_i^s \).

2.3. Cosine Distance Scoring

The cosine distance score for a trial between a set of i-vectors \( w_1 \) and \( w_2 \) is given by the dot product \( \langle \hat{w}_1 \cdot \hat{w}_2 \rangle \) between the inter-session-compensated and normalised vectors,

\[ \hat{w}_1 = \frac{B'A'w_1}{\|B'A'w_1\|}. \]

In this work, normalisation of the cosine kernel is performed using the approach described by [8] in which the cosine kernel is normalised with respect to the impostor score space using,

\[ \text{score}(\hat{w}_1, \hat{w}_2) = \frac{(\hat{w}_1 - \overline{w}_\text{imp})^t(\hat{w}_2 - \overline{w}_\text{imp})}{\|C_\text{imp}\hat{w}_1\|\|C_\text{imp}\hat{w}_2\|}. \]

3. LINEAR DISCRIMINANT ANALYSIS

In the context of the i-vector framework, LDA serves the purpose of enhancing discrimination between i-vectors corresponding to different speakers. LDA minimises the within-speaker variability observed in a training dataset while maximising the between-speaker variability through the eigenvalue decomposition of \( S_Bw = \lambda S_Ww \), where the between-speaker and within-speaker covariance matrices, \( S_B \) and \( S_W \) respectively, are calculated as,

\[ S_B = \sum_{s=1}^{S} \sum_{i=1}^{N_s} (w_i^s - \mu_s)(w_i^s - \mu_s)^t \]

\[ S_W = \sum_{s=1}^{S} \sum_{i=1}^{N_s} (w_i^s - \mu_s)(w_i^s - \mu_s)^t. \]

The i-vector mean \( \mu \) is a null vector when using the same training dataset for the total variability subspace and the LDA matrix (as is done in this work) due to the factor analysis assumption of normally distributed and zero-mean factors [2, 7].

3.1. Discussion

The effectiveness of LDA relies on the correct calculation of scatter matrices \( S_B \) and \( S_W \). The current algorithm for calculating the scatter matrices, however, neglects two major issues with regards to speaker recognition: the common use of an insufficient training dataset to completely define the within-speaker variability and the influence of source-related variation on the observed between-speaker scatter. Figure 1 depicts a graphical interpretation of these phenomena using four different speakers (defined by unique shapes) who each provide several utterances in the training dataset (depicted as repetitions of the speakers’ shape). The three speech sources of interest are depicted with respect to the sample mean \( \mu \).

The arrows in Figure 1(a) indicate the vectors used to calculate within-speaker scatter in this example. It can be observed that the lack of multi-source i-vectors from each speaker in the dataset results in a limited representation of source variation in the within-speaker scatter matrix. This desired variation is instead represented in the between-speaker scatter as shown in Figure 1(b). Although this diagram emphasises the difference between speech sources in
the i-vector space, it illustrates that speech acquisition methods have the potential to influence the observable between-speaker variation.

One approach to addressing the above issues is through the selection of a training dataset in which each speaker provides at least one sample from each source of interest. In practice, however, such a dataset is difficult to acquire. Proposed in the following section is a novel algorithm to counteract the aforementioned shortcomings of the current LDA approach for the purpose of speaker discrimination.

4. SOURCE-NORMALISED-AND-WEIGHTED LINEAR DISCRIMINANT ANALYSIS

Proposed is an approach that addresses the issues highlighted in the previous section regarding the sub-optimal estimation of LDA scatter matrices from insufficient resources in the i-vector framework.

As discussed in Section 3.1, the between-speaker scatter calculated using the standard LDA approach can be influenced by source variation. This influence can be reduced by estimating the between-speaker scatter using source-normalised vectors that are calculated with respect to their corresponding source mean. This normalisation process is depicted in Figure 1(c) such that the vectors used to calculate the scatter are considerably less affected by source variation than in Figure 1(b). The source-normalised \( S_{src} \) can then be composed of the source-dependent between-speaker scatter matrices such that

\[
S_{B} = \sum_{s=1}^{S} S_{src}^{B}
\]

(7)

\[
S_{src}^{B} = \frac{1}{N_{src}} \sum_{s=1}^{N_{src}} w_{n}^{src} (\mu_{s} - \mu_{src}) (\mu_{s} - \mu_{src})^{T},
\]

(8)

where \( \mu_{src} = \frac{1}{N_{src}} \sum_{s=1}^{N_{src}} w_{n}^{src} \) and \( N_{src} \) designates the number of speech samples taken from source \( src \). It can be noted that in this approach, a speakers’ utterances acquired from different sources are assumed to belong to disjoint speakers.

Based on the assumption that \( S_{B} \) no longer captures the within-speaker variability due to different methods of speech acquisition, the residual variability in the i-vector space should, therefore, comprise the within-speaker scatter. Specifically, the total variance in the training i-vectors is given by

\[
S_{T} = \sum_{n=1}^{N} w_{n} w_{n}^{T}
\]

(9)

The advantage to this approach is that the scatter is no longer dependent on the availability of multi-source utterances per speaker as is the case when calculated via (6). The use of \( S_{W} \) and \( S_{B} \) from equations (9) and (7) in the LDA optimisation will be referred to as source-normalised LDA (SN-LDA) for the remainder of this study.

Extending on (7), a weighting scheme can be introduced to bias the between-speaker scatter toward the most reliably estimated source-normalised covariance matrix \( S_{src}^{B} \). Motivation here comes from the inherently better representation of between-speaker scatter expected when it is calculated from a larger collection of i-vectors. The source-normalised-and-weighted (SNAW) between-speaker scatter matrix for use in LDA is calculated as

\[
S_{src}^{W} = \frac{1}{N_{src}} \sum_{s=1}^{N_{src}} w_{n}^{src} S_{src}^{B}.
\]

(10)

For the remainder of this study, SNAW-LDA will denote the use of \( S_{W} \) and \( S_{B} \) from equations (9) and (10) in the LDA optimisation.

5. EXPERIMENTAL PROTOCOL

The proposed approaches were evaluated on the recent NIST 2008 and 2010 SRE corpora. Results are reported for four evaluation conditions on each corpora with particular focus on mismatched trials. The SRE’10 conditions correspond to det conditions 2-5 in the evaluation plan [6], and include int-int, int-mic, int-tel, and tel-tel trials. Performance was evaluated using the equal error rate (EER) and a normalised minimum decision cost function (DCF) calculated using \( C_{M} = 1 \), \( C_{FA} = 1 \) and \( P_{T} = 0.001 \) for SRE’10 results. The extended evaluation protocol was used in which the number of trials ranged from 416119 (tel-tel) to more than 2 million (int-int) with 0.5-1.7% target trials. For SRE’08, det conditions 3-5 and 7 [5] were evaluated corresponding to int-int, int-tel, int-mic, and tel-tel (English-only) trials. The DCF was calculated using \( C_{M} = 10 \) and \( P_{T} = 0.01 \) for SRE’08 results.

Fully-weighted LDA was implemented as in [7] such that \( S_{B} \) and \( S_{W} \) were normalised by the number of training utterances rather than using (5) and (6) in this work. This aided performance in the under-resourced same-source conditions for which it was developed. The empirically determined weights for fully-weighted LDA were \([P_{tel}, P_{mic}, P_{int}] = [0.1, 0.45, 0.45]\). In all approaches, the number of LDA dimensions retained was evaluated in steps of 50 in order to minimise the average of \( (DCF + 10 \times \text{EER}) \) across the evaluated conditions of SRE’10 and \( (DCF + \text{EER}) \) for SRE’08.

Speech activity detection (SAD) was implemented using a 2-component GMM trained on the log-energy of a given speech file. Samples of low energy were iteratively removed from the speech signal and the GMM re-trained until the standard deviation of the speech Gaussian was less than five times that of the non-speech Gaussian. A speech threshold was then defined based on the speech Gaussian parameters. 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 the corresponding interviewer frame. The NIST supplied SAD files were utilised as a pre-processing step for SRE’08 interview segments.

Gender-dependent UBMs consisting of 512-components were trained on 60-dimensional, feature-warped MFCCs (including deltas and double-deltas) extracted 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 used as total variability subspace, LDA and WCCN training data and for cosine kernel normalisation. This was sourced from the aforementioned corpora along with additional interview data. Interview data was taken from the NIST 2008 SRE follow-up corpus for use in SRE’10 evaluations and from a subset of the 3-minute interview segments of the NIST 2010 SRE corpus for SRE’08 evaluations. The segment counts \( N_{tel}, N_{mic} \) and \( N_{int} \) were on average 15770, 2665, and 1784, respectively. Segments from each source were implicitly assumed to belong to different speakers.
Table 1 presents the SRE’08 results using standard, fully-weighted, source-normalised and SNAW LDA approaches.

Table 2. SRE’10 (extended protocol) results using standard, fully-weighted, source-normalised and SNAW LDA approaches.

### 6. RESULTS

6.1. NIST 2008 SRE Evaluations

Results from trials on the SRE’10 extended protocol are presented in Table 2. As in the SRE’08 trials, fully-weighted LDA provided no advantage over standard LDA in mis-matched evaluation conditions. Source-normalised LDA proved highly beneficial to mis-matched trials and interview-only trials such that relative improvements of up to 31% in performance statistics relative to the standard LDA approach for mis-matched trial conditions and conditions for which limited system development speech was available.

7. CONCLUSION

A novel source-normalised-and-weighted (SNAW) LDA technique was proposed for the i-vector framework for speaker recognition. The shortcomings of the standard LDA algorithm for enhancing speaker discrimination were highlighted. These included the influence of source-related variation on the between-speaker covariance matrix and an incomplete representation of the within-speaker scatter due to commonly insufficient multi-source utterances per speaker. The proposed approach reduced the influence of source variation on the between-speaker scatter through normalisation of the i-vectors with respect to the source means, along with a weighting criterion for the final scatter. The within-speaker scatter was calculated as the residual variation not captured by the source-normalised between-speaker scatter matrices, thereby improving estimation of the scatter from insufficient resources. When evaluated on recent NIST 2008 and 2010 SRE corpora, SNAW-LDA demonstrated significant improvements of up to 31% in performance statistics relative to the standard LDA approach for mis-matched trial conditions and conditions for which limited system development speech was available.

8. REFERENCES