Improving Short Utterance based I-vector Speaker Recognition using Source and Utterance-Duration Normalization Techniques

A. Kanagasundaram*, D. Dean*, J. Gonzalez-Dominguez†, S. Sridharan*, D. Ramos†, J. Gonzalez-Rodriguez†

* Speech Research Laboratory, Queensland University of Technology, Australia
† ATV5 Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

Abstract

A significant amount of speech is typically required for speaker verification system development and evaluation, especially in the presence of large intersession variability. This paper introduces a source and utterance-duration normalized linear discriminant analysis (SUN-LDA) approaches to compensate session variability in short-utterance i-vector speaker verification systems. Two variations of SUN-LDA are proposed where normalization techniques are used to capture source variation from both short and full-length development i-vectors, one based upon pooling (SUN-LDA-pooled) and the other on concatenation (SUN-LDA-concat) across the duration and source-dependent session variation. Both the SUN-LDA-pooled and SUN-LDA-concat techniques are shown to provide improvement over traditional LDA on NIST 08 truncated 10sec-10sec evaluation conditions, with the highest improvement obtained with the SUN-LDA-concat technique achieving a relative improvement of 8% in EER for mismatched conditions and over 3% for matched conditions over traditional LDA approaches.

Index Terms: speaker verification, i-vector, total-variability, LDA, WCCN

1. Introduction

In the presence of intersession variability conditions, significant amount of speech is required for speaker verification system development and current state-of-the-art systems still require substantial amounts of speech for training and testing. Reducing the amount of speech required for development, training and testing while obtaining satisfactory performance has been the focus of a number of recent studies in state-of-the-art speaker verification design, including joint factor analysis (JFA), i-vectors, probabilistic linear discriminant analysis (PLDA) and support vector machines (SVM) [1, 2, 3, 4]. Recently, Kenny et al. [5], have investigated how to quantify the uncertainty associated with the i-vector extraction process and propagate it into a PLDA classifier.

As an i-vector approach is based on defining only one variability space [6, 7], instead of the separate session variability and speaker spaces of the JFA approach, it is believed that i-vectors will not lose significant speaker information in session variability space [7], and is also believed that this would be additional advantage to short utterance speaker verification. Until now, several inter-session variability compensation approaches have been introduced to a long utterance cosine similarity scoring (CSS) i-vector speaker recognition system [7, 8, 9]. However, no single inter-session variability compensation approach has been specially analyzed with short utterance (~10s) based CSS i-vector speaker recognition. As our main focus of this paper is to analyze the inter-session variability compensation approaches with short utterances, we have chosen the simple CSS over PLDA classifier, as CSS is computationally more efficient approach than PLDA.

In this paper, initially, we analyze how the short utterance based i-vector system performs when the standard session variability compensation approaches, including LDA and WCCN are trained using long- and short-length utterance development data. Later, we also introduce source and utterance-duration normalized LDA (SUN-LDA) approach for the purpose of improving speaker verification performance under short utterance evaluation conditions.

This paper is structured as follows. Section 2 gives a brief introduction to speaker verification using i-vectors. Section 3 details the proposed SUN-LDA approach. The experimental protocol and corresponding results are given in Section 4 and Section 5.

2. Speaker verification using i-vectors

In contrast to the separate speaker and session dependent subspaces of JFA, i-vectors represent the GMM super-vector using a single total-variability subspace [6]. An i-vector speaker and session dependent GMM super-vector can be represented by

\[
\mu = m + Tw
\]

where \( m \) is the speaker and session independent background UBM super-vector, \( T \) is a low rank matrix representing the primary variation across a large collection of development data, \( w \) is normally distributed with parameters \( N(0, I) \), and is the i-vector representation used for speaker verification. \( R_w \) is dimension of i-vector features.

The total-variability subspace, \( T \), training and subsequent i-vector extraction is detailed by Dehak et al. [10]. As we are investigating the telephone and microphone based speaker verification system in this paper, the total-variability subspace should be trained in a manner which exploits the useful speaker variability contained in speech acquired from both telephone and microphone sources. In this paper, the total-variability subspace (\( R_w = 500 \)) is trained on telephone and microphone speech pooled utterances, as recent studies have found that pooled total-variability approach is better than concatenated total-variability approach [11].
2.1. Inter-session variability compensation techniques

As i-vectors are defined by single variability space, containing both speaker and session variability information, there is a requirement that additional inter-session variability compensation approaches be taken before verification. These inter-session variability compensation techniques are designed to maximize the effect of between-speaker variability and minimize the effects of within-speaker variability due to differences in microphones, acoustic environment and variation in speaker’s voices.

Within this section, we will outline the LDA followed by WCCN (LDA,WCCN) approach, which is standard approach for CSS i-vector system. The WCCN and SN-LDA followed by WCCN (SN-LDA,WCCN) techniques are respectively detailed in following papers [10] and [8, 12].

2.1.1. LDA followed by WCCN

In the first stage of the LDA,WCCN approach, LDA is used to define a new spatial axes A that minimizes the within-class variance caused by channel effects and maximizes the variance between speakers in the i-vector space. WCCN is then used as an additional session variability compensation technique to scale the subspace in order to attenuate dimensions of high within-class variance.

Both LDA and WCCN calculations are based on the standard within- and between-class scatter estimations $S_w$ and $S_b$, calculated as

$$ S_b = \sum_{s=1}^{S} n_s (\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^T, $$

$$ S_w = \sum_{s=1}^{S} \sum_{i=1}^{n_s} (w_i - \bar{w}_s)(w_i - \bar{w}_s)^T, $$

where $S$ is the total number of speakers, $n_s$ is number of utterances of speaker $s$, and $N$ is the total number of sessions. The mean i-vectors, $\bar{w}_s$ for each speaker, and $\bar{w}$ is the across all speakers are defined by

$$ \bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_i, $$

$$ \bar{w} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} w_i. $$

The LDA matrix, $A$, is calculated through the eigenvalue decomposition of $S_w = AS_w A^T$. In the second stage, the WCCN transformation matrix ($B_2$) is trained using the LDA-projected i-vectors [7] from the first stage. The WCCN matrix ($B_2$) is calculated using Cholesky decomposition of $B_2B_2^T = W^{-1}$, where the within-class covariance matrix $W$ is calculated using

$$ W = \frac{1}{S} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (A^T(w_i - \bar{w}_s))(A^T(w_i - \bar{w}_s))^T. $$

The LDA,WCCN inter-session variability compensated i-vector can be calculated as follows,

$$ \hat{w}_{LDA,WCCN} = B_2^T A^T w $$

2.2. Cosine similarity scoring

Scoring of inter-session variability compensated i-vectors for speaker verification is accomplished using a cosine similarity scorer (CSS). The CSS operates by comparing the angles between a session variability compensated test i-vector, $\hat{w}_{test}$, and a session variability compensated target i-vector $\hat{w}_{target}$:

$$ \text{score}(\hat{w}_{target}, \hat{w}_{test}) = \frac{\hat{w}_{target}^T \hat{w}_{test}}{||\hat{w}_{target}|| ||\hat{w}_{test}||}. $$

3. SUN-LDA followed by WCCN

In this section, we introduce the source and utterance-duration normalized LDA inter-session variability compensation approach for the purpose of improving the short utterance i-vector speaker verification system. The influence of short utterance development data for channel estimations were detailed in Section 5.2. McLaren et al. have introduced the source-normalized between-class estimations to capture the source variation information [8]. Based upon short utterance development data analysis and the fundamentals of source-normalized estimations, we introduce the source and utterance-duration normalized between-class estimations, which can be used to capture the source variation information from full- and short-length development i-vectors. The telephone and microphone sourced utterance-duration normalized between-class scatters, $S_{b,\text{tel}}$ and $S_{b,\text{mic}}$, are defined as follows,

$$ S_{b,\text{tel}} = \alpha_{tf} S_{b,\text{full}} + \alpha_{ts} S_{b,\text{short}}, $$

$$ S_{b,\text{mic}} = \alpha_{mf} S_{b,\text{full}} + \alpha_{ms} S_{b,\text{short}}, $$

where $S_{b,\text{full}}$ and $S_{b,\text{short}}$ are individually estimated from telephone and microphone sourced full-length utterances using Equation 2. $S_{b,\text{tel}}$ and $S_{b,\text{mic}}$ are estimated using telephone and microphone sourced short-length utterances respectively. $\alpha_{tf}$, $\alpha_{mf}$, $\alpha_{ts}$ and $\alpha_{ms}$ are respectively weighting coefficients of telephone and microphone sourced full- and short-length between-class scatter estimations, and we have just analyzed the importance of each sources using the binary weighting coefficients.

The within-class scatter matrix ($S_w$) is calculated using standard approach as in Equation 3. Full-length utterances are used to calculate the $S_w$ as short-length i-vectors based $S_w$ affects the speaker verification performance. In this paper, two approaches are used to estimate the SUN-LDA matrix, $A$.

1. Option I (SUN-LDA-pooled)

Estimate the SUN-LDA-pooled matrix, $A$, based on summation of telephone and microphone sourced utterance-duration normalized between-class scatter and standard within-class scatter matrix. The SUN-LDA-pooled matrix can be estimated as eigenvalue decomposition of,

$$ (S_{b,\text{full}} + S_{b,\text{mic}}) \nu = \lambda S_w \nu $$

Empirically 150 eigenvectors were selected for SUN-LDA-v1 training with optimum performance.

2. Option II (SUN-LDA-concat)

Estimate the telephone and microphone sourced dependent LDA matrices separately for telephone and microphone sourced utterance-duration normalized between-class estimation. The telephone and microphone sourced dependent LDA matrices, $A_{tel}$ and $A_{mic}$, can be estimated as eigenvalue decomposition of,

$$ S_{b,\text{tel}} \nu = \lambda S_{w} \nu $$

$$ S_{b,\text{mic}} \nu = \lambda S_{w} \nu $$
Table 1: Performance comparison of baseline systems on truncated 10sec-10sec evaluation conditions. The best performing systems by both EER and DCF are highlighted across each column.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Interview-interview</th>
<th>Interview-telephone</th>
<th>Telephone-microphone</th>
<th>Telephone-telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCCN</td>
<td>19.81% 0.0798</td>
<td>24.33% 0.0896</td>
<td>19.76% 0.0821</td>
<td>17.37% 0.0695</td>
</tr>
<tr>
<td>LDA-WCCN</td>
<td>18.10% 0.0767</td>
<td>22.67% 0.0861</td>
<td>19.03% 0.0817</td>
<td>16.46% 0.0679</td>
</tr>
<tr>
<td>SN-LDA-WCCN</td>
<td>18.01% 0.0771</td>
<td>21.57% 0.0858</td>
<td>18.94% 0.0813</td>
<td>16.56% 0.0683</td>
</tr>
</tbody>
</table>

The SUN-LDA-concat matrix is formed by concatenating the telephone and microphone sourced LDA matrices, \( A_{tel} \) and \( A_{mic} \), which can be estimated as follows,

\[
A = [A_{tel} A_{mic}]
\]

Empirically 100 and 50 eigenvectors were respectively selected for telephone and microphone sourced LDA estimations.

4. Methodology

The proposed methods were evaluated using the NIST 2008 SRE corpora [13]. The shortened evaluation utterances were obtained by truncating the NIST2008 short2-short3 conditions to the specified length of active speech for both enrolment and verification. Prior to the evaluation and development utterance truncation, the first 20 seconds of active speech were removed from all utterances to avoid capturing similar introductory statements across multiple utterances. For NIST SRE 2008, the performance was evaluated using equal error rate (EER) and minimum decision cost function (DCF), calculated using \( C_{miss} = 10 \), \( C_{FA} = 1 \), and \( P_{target} = 0.01 \).

We have used 13 feature-warped MFCC with appended delta coefficients and two gender-dependent UBM containing 512 Gaussian throughout our experiments. UBMs were trained on telephone and microphone data from NIST 2004, 2005, and 2006 SRE corpora for telephone and microphone i-vector experiments. These gender-dependent UBMs were used to calculate the Baum-Welch statistics before training two gender-dependent total-variability subspaces of dimension \( R_v = 500 \). Total variability representation and session variability compensation matrices were trained using telephone and microphone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II. Randomly selected telephone and microphone utterances from NIST04, 05 and 06 were pooled to form the ZT score normalization dataset. 150 eigenvectors were selected for LDA and SN-LDA training by performance on a development dataset. The short-length development data was obtained by truncating the telephone and microphone speech data from NIST 2004, 2005 and 2006 SRE corpora as well as Switchboard II to the specified length of active speech, and it was truncated into two interval periods in order to increase the amount of development data.

5. Results and discussion

Initially, we have analyzed the baseline i-vector performance with short utterance evaluation conditions. After that we have progressively analyzed the i-vector performance, when the inter-session variability compensation approach was trained using full- and short-length utterance development data. Finally, the newly proposed both SUN-LDA-pooled and SUN-LDA-concat techniques were analyzed.

5.1. Baseline performance

To serve as a performance baseline, standard inter-session variability compensation approaches, including WCCN, LDA-WCCN and SN-LDA-WCCN were investigated with truncated 10sec-10sec evaluation condition as shown in Table 1. The NIST standard condition development data (full-length) was used for inter-session variability compensation approaches training. As it had been previously shown by Dehak [7], we have confirmed that the LDA-WCCN provides an improvement over WCCN. As it had been previously shown by McLaren [8], the results also confirm that SN-LDA-WCCN provides an improvement over LDA-WCCN on mis-matched conditions.

5.2. Inter-session variability compensation training using full- and short-length utterances

In this section, short utterance i-vector performance is analyzed when the inter-session variability approach, LDA-WCCN, is trained using full- and short-length development data. The NIST standard development data is used for full-length training. For matched-length, the development data is truncated into similar length to the evaluation condition. Full- and matched-length utterances are pooled together to create the mixed-length development data.

Table 2 presents the results comparing the LDA-WCCN based i-vector performance with full-, matched- and mixed-length WCCN training on NIST 2008 truncated 10sec-10sec evaluation conditions. The results suggest when short utterances are added to development set for WCCN training, it considerably affects the speaker verification performance as it de-

Table 2: Performance comparison of LDA_WCCN based i-vector systems on truncated 10sec-10sec evaluation conditions with full-, matched- and mixed-length utterance based WCCN training. The best performing systems by both EER and DCF are highlighted across each column.

<table>
<thead>
<tr>
<th>WCCN training</th>
<th>Interview-interview</th>
<th>Interview-telephone</th>
<th>Telephone-microphone</th>
<th>Telephone-telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-length</td>
<td>18.10% 0.0767</td>
<td>22.67% 0.0861</td>
<td>19.03% 0.0817</td>
<td>16.46% 0.0679</td>
</tr>
<tr>
<td>Matched-length</td>
<td>20.39% 0.0822</td>
<td>25.34% 0.0894</td>
<td>21.05% 0.0842</td>
<td>17.87% 0.0720</td>
</tr>
<tr>
<td>Mixed-length</td>
<td>19.57% 0.0804</td>
<td>24.44% 0.0879</td>
<td>20.38% 0.0823</td>
<td>17.30% 0.0708</td>
</tr>
</tbody>
</table>

\[
A = [A_{tel} A_{mic}]
\]
Table 3: Performance comparison of SUN-LDA approach based i-vector systems on truncated 10sec-10sec evaluation conditions. The best performing systems by both EER and DCF are highlighted across each column.

(a) SUN-LDA-pooled vs LDA

<table>
<thead>
<tr>
<th>System</th>
<th>Interview-interview</th>
<th>Interview-telephone</th>
<th>Telephone-microphone</th>
<th>Telephone-telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
</tr>
<tr>
<td>Baseline approach (LDA, WCCN)</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
</tr>
<tr>
<td>Source and utterance-duration normalized approach (SUN-LDA-pooled, WCCN)</td>
<td>0.0767 22.67% 0.0861 19.03% 0.0817 16.46% 0.0679</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 1.0</td>
<td>0.0764 21.37% 0.0857 18.75% 0.0790 16.56% 0.0670</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 1.0</td>
<td>0.0766 21.47% 0.0859 18.66% 0.0811 16.56% 0.0681</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 0.0</td>
<td>0.0759 21.22% 0.0858 18.60% 0.0783 16.31% 0.0667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 0.0</td>
<td>0.0761 21.47% 0.0859 18.87% 0.0807 16.48% 0.0674</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 0.0 1.0 0.0</td>
<td>0.0774 21.21% 0.0856 17.66% 0.0776 16.31% 0.0665</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) SUN-LDA-concat vs LDA

<table>
<thead>
<tr>
<th>System</th>
<th>Interview-interview</th>
<th>Interview-telephone</th>
<th>Telephone-microphone</th>
<th>Telephone-telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
<td>α_{ts} α_{tf} α_{ms}</td>
</tr>
<tr>
<td>Baseline approach (LDA, WCCN)</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
</tr>
<tr>
<td>Source and utterance-duration normalized approach (SUN-LDA-concat, WCCN)</td>
<td>0.0767 22.67% 0.0861 19.03% 0.0817 16.46% 0.0679</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 1.0 1.0</td>
<td>0.0760 20.19% 0.0852 18.06% 0.0852 16.14% 0.0706</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 1.0</td>
<td>0.0767 20.36% 0.0846 17.59% 0.0831 16.06% 0.0694</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 0.0</td>
<td>0.0764 20.39% 0.0848 17.79% 0.0832 16.31% 0.0692</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 1.0 0.0 0.0</td>
<td>0.0760 20.91% 0.0841 17.39% 0.0814 15.82% 0.0677</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0 0.0 1.0 0.0</td>
<td>0.0773 21.84% 0.0848 18.13% 0.0794 16.64% 0.0654</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3. SUN-LDA analysis

We hypothesize the short utterance development set i-vectors may not affect the quality of inter-speaker variation as they do not depend on phonetic contents. The influence of short utterance i-vectors for inter-speaker variance estimation is analyzed in this section. In order to capture the source variation information from full- and short-length i-vectors, two different types of SUN-LDA approaches, SUN-LDA-pooled and SUN-LDA-concat, are analyzed.

In this paper, we have introduced two different types of source and utterance-duration normalized LDA approaches, SUN-LDA-pooled and SUN-LDA-concat, to compensate the session variability in short utterance i-vector speaker verification systems. By capturing the source variation information from short- and full-length development i-vectors, the SUN-LDA-concat technique has shown over 8% improvement in EER for mismatched conditions and over 3% improvement in EER for matched conditions over traditional LDA approaches in NIST 08 truncated 10sec-10sec evaluation conditions. In future, fine tuning of weighting will be analyzed to improve the performance of SUN-LDA based i-vector approach, and SUN-LDA and length-normalized PLDA combined system will be also analyzed to improve the speaker verification performance.

6. Conclusion

In this paper, we have introduced two different types of source and utterance-duration normalized LDA approaches, SUN-LDA-pooled and SUN-LDA-concat, to compensate the session variability in short utterance i-vector speaker verification systems. By capturing the source variation information from short- and full-length development i-vectors, the SUN-LDA-concat technique has shown over 8% improvement in EER for mismatched conditions and over 3% improvement in EER for matched conditions over traditional LDA approaches on NIST 08 truncated 10sec-10sec evaluation conditions. In future, fine tuning of weighting will be analyzed to improve the performance of SUN-LDA based i-vector approach, and SUN-LDA and length-normalized PLDA combined system will be also analyzed to improve the speaker verification performance.

7. Acknowledgements

This project was supported by the European Commission Marie Curie ITN “Bayesian Biometrics for Forensics” (BBfor2) network and the Spanish Ministerio de Economia y Competitividad under the project TEC2012-37585-C02-01.

8. References

ances,” in Interspeech 2008, no. 853-856, (Brisbane, Australia), September 2008.


