Compensation of Intrinsic Variability with Factor Analysis Modeling for Robust Speaker Verification

Sheng Chen, Mingxing Xu

Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract

Performances of speaker verification systems are adversely affected by intrinsic variability in the real world applications. In this paper, factor analysis approaches of Joint Factor Analysis (JFA) and i-vector modeling are used to address the effects of intrinsic variations for robust speaker verification. The speaker variability and intrinsic variability are modeled with the speaker and session factors respectively in the JFA approach. In the i-vector framework, a low-dimensional space is defined to model the total variability and intrinsic variations are compensated with a variety of techniques including Linear Discriminant Analysis (LDA), Within-Class Covariance Normalization (WCCN) and Nuisance Attribute Projection (NAP). Experiments in the intrinsic variation corpus show that factor analysis approaches of JFA and i-vector framework perform much better than the GMM-UBM paradigm in modeling the intrinsic variability. Relative reductions in Error Equal Rate (EER) of around 39.85% and 36.76% are obtained respectively for JFA and i-Vector+LDA+WCCN speaker verification systems, compared to the GMM-UBM baseline system.

Index Terms: speaker verification, intrinsic variability, joint factor analysis, i-vector, LDA, WCCN, NAP

1. Introduction

Over recent years, significant progress has been made to address the effects of extrinsic variability such as the mismatch of channel or background noise with a variety of techniques including eigenchannel compensation [1], Joint Factor Analysis (JFA) [2] and Support Vector Machines (SVM) with session compensation techniques [3].

However, limited number of research has been done to address the intrinsic variation problems of speaker verification. Ghiurcau [4] found that performances of GMM-UBM based speaker verification systems decrease significantly when emotions alter the human voice. Results in [5] show that vocal effort level has a dramatic effect on speaker verification performances. It is indicated in [6] that inter-session compensation techniques originally designed for channel compensation are indeed modeling the intrinsic variation represented in the data.

In this paper, we focus on how the recent factor analysis approaches of JFA and i-vector modeling behave in addressing the effects of intrinsic variability for robust speaker verification. An intrinsic variation corpus has been designed and collected from six aspects including speaking style, speaking rate, speaking volume, emotional state, physical status, and speaking language. Speaker and session factors are used to model the speaker and intrinsic variability respectively while the i-vector framework is used to model the total variability in a low-dimensional space. LDA, WCCN and NAP are applied in the i-vector framework to compensate for intrinsic variations. A JFA based speaker verification system and four i-vector based systems, namely i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP and i-Vector+LDA+WCCN, are built up and their performances are investigated in the intrinsic variation corpus.

This paper is organized as follows. The intrinsic variation corpus used in our experiments is described in Section 2. In Section 3, we will introduce the factor analysis approaches of JFA and i-vector modeling for the compensation of intrinsic variability. Experimental results on the intrinsic variation corpus are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Intrinsic Variation Corpus

2.1. Intrinsic Variation Forms

To support the study the effects of intrinsic variations, an intrinsic variation corpus has been designed and collected from six aspects of intrinsic variations including speaking style, speaking rate, speaking volume, emotional state, physical status and speaking language. We define the neutral spontaneous speech at normal rate and volume in Chinese as the base case. Eleven different variation forms are derived from the base form and they are simply noted as reading, fast, slow, loud, soft, whispered, angry, happy, denasalized, mumbled, and English, respectively. The derivation process is presented in Figure 1.
3. Factor Analysis Speaker Verification

3.1. Joint Factor Analysis

The joint factor analysis technique proposed by Kenny [2] is based on the assumption that the Gaussian Mixture Model (GMM) supervector \( M \) of a given utterance can be decomposed into separate speaker and session dependent parts in the following equation

\[
M = S + C
\]  

where the speaker supervector \( S \) and the session supervector \( C \) can be represented as

\[
S = m + V y + D z \quad (2)
\]
\[
C = U x \quad (3)
\]

In this model, \( m \) is the speaker- and session- independent supervector which can be obtained by the Universal Background Model (UBM) training, \( V \) and \( U \) represents the primary directions of speaker variability and session variability respectively, and \( D \) models the residual variability not captured by the speaker subspace. The vectors \( x, y \) and \( z \) are the session, speaker and residual factors.

Speaker enrollment is accomplished by extracting the full speaker dependent GMM supervector while discarding the session dependent component. In the verification process, the session dependent component of the testing utterance is estimated and the linear dot-product approach is used to compute the likelihood of testing utterance against the session compensated speaker model \( M - U x \).

3.2. i-Vector Modeling

3.2.1. Total Variability

A more recent approach to front-end factor analysis, termed i-vectors, has been proposed by Dehak [8] with the motivation of JFA. In the i-vector framework, the speaker- and session-dependent GMM supervector \( M \) is defined as

\[
M = m + Tw
\]  

where \( T \) is rectangular matrix of low rank representing the total variability space and \( w \) representing the total variability factors is referred to as identity vector or i-vector. Cosine similarity is calculated as the detection score for a trial between the target speaker i-vector \( w_{\text{target}} \) and the testing i-vector \( w_{\text{test}} \) in the following equation

\[
\text{score}(w_{\text{target}}, w_{\text{test}}) = \frac{w_{\text{target}}^T w_{\text{test}}}{\| w_{\text{target}} \| \| w_{\text{test}} \|}
\]  

3.2.2. Inter-session Compensation

After i-vectors are extracted from utterances, the session variability can be compensated with a number of existing approaches borrowed from SVM speaker verification such as Linear Discriminant Analysis (LDA), Within-Class Covariance Normalization (WCCN) and Nuisance Attribute Projection (NAP).

LDA aims to find a reduced set of axes that minimize the within-speaker variability observed in the i-vectors while simultaneously maximizing the between-speaker variability. The LDA projection matrix, \( A \), is formed as the subset of eigenvectors having the largest eigenvalues from equation

\[
S_B v = \lambda S_W v
\]  

where \( S_B \) and \( S_W \) are the between-speaker covariance matrix and the within-speaker covariance matrix respectively.

WCCN [9] is used as a channel compensation technique to attenuate dimensions of high with-class variance by scaling the total variability space. The i-vectors are normalized with the inverse of the within-speaker covariance matrix, which is equivalent to scaling the i-vector space with the projection matrix \( B \), where \( S_W^{-1} = B B^T \).

NAP [10] attempts to find an appropriate projection matrix to remove the nuisance direction. The NAP transformation matrix \( P \) is defined as

\[
P = I - V V^t
\]  

where \( I \) is the identity matrix and \( V \) is a rectangular matrix of low rank which can be obtained by taking the top
$k$ eigenvectors having the best eigenvalues of the within-class covariance matrix.

4. Experiments and Discussion

4.1. Experiment data

The intrinsic variation corpus introduced in Section 2 is used in our experiments to investigate the performance of factor analysis techniques in addressing the effects of intrinsic variability. The duration of each enrollment utterance and testing utterance is about 18 seconds. Table 1 presents the data partitions of the intrinsic variation corpus for training and testing in the experiments.

Table 1: Data partitions in the intrinsic variation corpus

<table>
<thead>
<tr>
<th>Function</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM</td>
<td>30 speakers</td>
<td>18 hours 12 variation forms</td>
</tr>
<tr>
<td>U, V, D in JFA</td>
<td>30 speakers</td>
<td>18 hours 12 variation forms</td>
</tr>
<tr>
<td>T in i-Vector</td>
<td>12 hours 12 variation forms</td>
<td></td>
</tr>
<tr>
<td>LDA, WCCN, NAP</td>
<td>20 speakers</td>
<td>2400 utterances 12 variation forms</td>
</tr>
<tr>
<td>Testing Data</td>
<td>20 speakers</td>
<td></td>
</tr>
</tbody>
</table>

4.2. Speaker Verification Systems

GMM-UBM based speaker verification system is chosen as the baseline system while a JFA system and four i-vector based systems (i-Vector+LDA, i-Vector+WCCN, i-Vector+NAP, i-Vector+LDA+WCCN) are built up to investigate the effects of factor analysis approaches in modeling intrinsic variability. Mel Frequency Cepstral Coefficients (MFCC) are extracted with 32ms window length and 16ms frame rate from utterances. The MFCC features are composed of 12 cepstral coefficients and energy, adding derivatives of first and second order to produce 39 dimensional feature vectors. A gender independent UBM composed of 512 gaussian components is used throughout our experiments. The JFA system is made of 50 speaker factors and 20 session factors while 200 total factors are used for the total variability space in the i-vector based systems. In the LDA compensation, the 25 best eigenvectors are used to form the LDA projection matrix. The NAP transformation matrix is composed of top 100 eigenvectors having the best eigenvalues of the within-class covariance matrix.

4.3. Experiments

Performances of speaker verification systems are investigated for each enrollment condition when utterances with all the twelve variation forms are used for testing. 2400 utterances from 20 speakers are used for enrollment and testing. There are 12 conditions of intrinsic variations and every subject has 10 utterances for each condition. For each subject, 1 utterance of the enrollment condition is used for training and the other 119 utterances with the twelve variation forms are used to create target trials. All the 2280 utterances of the other subjects are used to create imposter trials. The training and testing procedure is repeated 10 times for each subject by choosing another enrollment utterance. Error Equal Rate (EER) and Detection Error Tradeoff (DET) plot are used to measure the performances of speaker verification systems.

Table 2: EERs(%) for each enrollment condition when utterances with all the twelve variation forms are used for testing.

<table>
<thead>
<tr>
<th>Speech Variation</th>
<th>Variation Form</th>
<th>GMM-UBM</th>
<th>LDA</th>
<th>WCCN</th>
<th>NAP</th>
<th>LDA+WCCN</th>
<th>JFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Form</td>
<td>Spontaneous</td>
<td>21.39</td>
<td>18.66</td>
<td>14.41</td>
<td>16.89</td>
<td>13.57</td>
<td>12.79</td>
</tr>
<tr>
<td>Speaking Style</td>
<td>Reading</td>
<td>25.13</td>
<td>20.25</td>
<td>17.39</td>
<td>20.46</td>
<td>14.92</td>
<td>17.50</td>
</tr>
<tr>
<td>Speaking Volume</td>
<td>Loud</td>
<td>27.10</td>
<td>20.08</td>
<td>21.22</td>
<td>19.96</td>
<td>17.06</td>
<td>15.50</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>31.51</td>
<td>22.69</td>
<td>17.82</td>
<td>20.63</td>
<td>16.68</td>
<td>13.13</td>
</tr>
<tr>
<td></td>
<td>Whispered</td>
<td>46.93</td>
<td>32.69</td>
<td>33.24</td>
<td>32.02</td>
<td>25.88</td>
<td>29.92</td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>Fast</td>
<td>27.49</td>
<td>23.19</td>
<td>20.80</td>
<td>22.52</td>
<td>19.92</td>
<td>15.92</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>23.03</td>
<td>19.58</td>
<td>19.12</td>
<td>18.15</td>
<td>16.93</td>
<td>15.21</td>
</tr>
<tr>
<td>Emotional State</td>
<td>Angry</td>
<td>26.60</td>
<td>23.28</td>
<td>21.43</td>
<td>23.36</td>
<td>19.41</td>
<td>16.29</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>23.49</td>
<td>18.49</td>
<td>16.43</td>
<td>18.24</td>
<td>15.42</td>
<td>12.33</td>
</tr>
<tr>
<td>Physical Status</td>
<td>Denasalized</td>
<td>20.71</td>
<td>17.94</td>
<td>16.39</td>
<td>19.75</td>
<td>14.41</td>
<td>13.17</td>
</tr>
<tr>
<td></td>
<td>Mumbled</td>
<td>22.52</td>
<td>18.49</td>
<td>16.34</td>
<td>18.07</td>
<td>15.25</td>
<td>12.75</td>
</tr>
</tbody>
</table>
4.4. Results and Discussion

Table 2 presents the EERs obtained in the GMM-UBM baseline system, the JFA system and i-vector based speaker verification systems. The best results are formatted in bold across each row. It is obvious that factor analysis based systems outperform the GMM-UBM based system for each enrollment condition and the best results are obtained in the JFA system. The combination of LDA and WCCN perform best to compensate for the intrinsic variability in the intersession compensation techniques.

Table 3: Performances of Speaker Verification Systems

<table>
<thead>
<tr>
<th>System</th>
<th>EER(%)</th>
<th>Relative Reduction(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-UBM(baseline)</td>
<td>27.33</td>
<td></td>
</tr>
<tr>
<td>i-Vector+LDA</td>
<td>21.35</td>
<td>21.89</td>
</tr>
<tr>
<td>i-Vector+WCCN</td>
<td>19.78</td>
<td>27.65</td>
</tr>
<tr>
<td>i-Vector+NAP</td>
<td>20.96</td>
<td>23.32</td>
</tr>
<tr>
<td>i-Vector+LDA+WCCN</td>
<td>17.29</td>
<td>36.76</td>
</tr>
<tr>
<td>JFA</td>
<td>16.44</td>
<td>39.85</td>
</tr>
</tbody>
</table>

The overall EERs of speaker verification systems for all the enrollment conditions are presented in Table 3. The best results are obtained in the JFA system with the significant relative reduction of EER around 39.85%, compared to the GMM-UBM baseline system. The corresponding DET curve presented in Figure 2 demonstrates that the JFA approach and the combination of LDA and WCCN in the i-vector modeling are the most useful to compensate for intrinsic variations. These results indicate the effectiveness of factor analysis techniques in addressing the effects of intrinsic variability.

5. Conclusions

This paper investigates how the recent factor analysis techniques perform in addressing the effects of intrinsic variability from six aspects, namely speaking style, speaking rate, speaking volume, emotional state, physical status, and speaking language. Factor analysis approaches of JFA and i-vector framework are used to model the speaker and intrinsic variability for robust speaker verification. LDA, WCCN and NAP are tested for the compensation of intrinsic variability in the i-vector framework and the best results are obtained with the combination of LDA and WCCN. Experiments in the intrinsic variation corpus show that JFA and i-vector framework perform better than the GMM-UBM paradigm in modeling the intrinsic variability. Compared to the GMM-UBM baseline system, significant relative reductions in EER of around 39.85% and 36.76% are obtained respectively in the JFA and i-Vector+LDA+WCCN speaker verification systems.

6. Acknowledgment

The work in this paper is partially supported by the National Natural Science Foundation of China (No. 61171116) and National Basic Research Program of China (973 Program, No. 2012CB316400).

7. References