Intrinsic Variation Robust Speaker Verification based on Sparse Representation

Yi Nie, Mingxing Xu, Haishu Xianyu

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

Intrinsic variation is one of the major factors that aggravate performance of speaker verification system dramatically. In this paper, we focus on alleviating influence caused by intrinsic variation using sparse representation. Because the over-complete dictionary increases the flexibility and the ability to adapt to variable data in signal representation, we expect redundancy of the dictionary could benefit addressing the implicit properties of intrinsic variation within each speaker. Both exemplar dictionary and learned dictionary are evaluated on an intrinsic variation corpus and compared with GMM-UBM, Joint Factor Analysis (JFA) and i-vector systems. In our system, we choose the K-SVD algorithm, generalization of K-means algorithm to learn dictionary with Singular Value Decomposition (SVD). The experiment results show that the two sparse representation systems achieve higher accuracy than GMM-UBM, JFA and i-vector systems consistently, especially outperform GMM-UBM respectively by 37.17% and 41.55%. We also find that the K-SVD based sparse representation system has almost the best performance, which achieve an average Error Equal Rate (EER) of 14.23%.

Index Terms: speaker verification, speaking style, intrinsic variation, sparse representation, K-SVD

1. Introduction

In the field of speaker verification, great progress has been made on addressing the effects of the extrinsic variability while limited research has been done to decrease the performance degradation caused by intrinsic variation including, emotion state, vocal effort, state of health and so on. Shriberg [1] found that vocal effort level had dramatic effects on GMM-UBM system. Ghirucau [2] illustrated that performance of speaker verification system could decrease significantly when the speaker’s emotion state changed.

However, change of speaking style cannot be constrained and predicted in real world multimedia applications or real life scenarios, more general approaches are required. Wei Wu [3] used emotion-dependent score normalization method to decrease the aggravation caused by the mismatch between speaker GMM-UBM models and test utterances. Huanjun Bao [4] proposed the emotion attribute projection method to alleviate the intra speaker emotion variability. Inspired by the channel compensation techniques, Sheng Chen applied Joint Factor Analysis (JFA) [5] and i-vector [6] to speaking style robust speaker verification on an intrinsic variation corpus. The two frameworks both perform better than the GMM-UBM baseline system [7].

Recently, sparse representation provides a new direction to signal processing. Sparse representation means among all the coefficients of the expressed vectors, only a small fraction on the entries are nonzero. It could select the subspace of base vectors (namely the dictionary) which expresses the input signal most concentrated and automatically reject other less concentrated components [8]. Therefore, its discriminative abilities have been exploited to various fields of pattern recognition. In [9] exemplar dictionary based sparse representation was employed to speaker verification and got higher accuracy than i-vector framework. Following this work, Haris [10] proved that K-SVD learned dictionary speaker verification system outperforms the exemplar one.

The over-complete dictionary increases the flexibility and the ability to adapt to variable data in signal representation. Inspired by this, we expect redundancy of the dictionary in sparse representation could benefit addressing the implicit properties of intrinsic variation of each speaker. In this work, the sparse representation is used to model the i-vectors in the total variability space.

The main contribution of our paper is to propose a general solution to speaking style independent speaker verification system using sparse representation. Both exemplar dictionary and learned dictionary are evaluated on an intrinsic variation corpus [11] and compared with well-known methods such as GMM-UBM, JFA and i-vector systems.

This paper is organized as follows: In Section 2, we briefly introduce the i-vector framework. The sparse representation and approaches to build dictionary are described in Section 3. Details of the experimental setup are given in Section 4, followed by results and discussion in Section 5. Finally, we summarize the conclusions in Section 6.

2. Total variability i-vector Speaker Verification system

We use sparse representation to model the i-vectors in the total variability space due to its excellent discriminative capability and small dimensionality. The total variability space contains the speaker and channel information simultaneously. Given an utterance, the GMM mean supervectors $s$ for a speaker can be represented as,

$$s = m + Tw$$

where $m$ is the speaker-independent UBM mean supervector, $T$ is a rectangular total variability matrix of lower rank and $w$ is the so-called i-vector [6]. The channel effects are removed with some compensation techniques, such as Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN) [12]. WCCN uses the inverse of the within-class covariance to normalize the cosine kernel while LDA minimizes the intra-class variance due to channel effects and maximizes

978-616-361-823-8 © 2014 APSIPA

APSIPA 2014
the variance between speakers. After that speaker verification can be accomplished by calculating cosine kernel between training i-vectors and testing i-vectors.

3. Sparse Representation For Speaker Verification

In this section, we provide a brief introduction of approaches to build dictionary. Then the speaker verification system based on sparse representation is demonstrated including sparse coding of utterances and verification methods. The framework is showed in figure 1.

![Diagram](image)

Figure 1: Framework of sparse representation speaker verification systems

3.1. Dictionary Construction

In the sparse representation system, the i-vectors representing the target speakers and a set of non-target background speakers are used to build the dictionary. Two approaches to building the dictionary, exemplar and K-SVD are introduced below.

3.1.1. Exemplar Dictionary

Given \( N_{tar} \) target training samples represented as \( D_{tar} \) and \( N_{bg} \) non-target background training samples represented as \( D_{tar} \), the exemplar dictionary \( D \) is constructed as follows,

\[
D = [D_{tar}, D_{bg}] = [q_{tar,1}, \ldots, q_{tar,N_{tar}}, q_{bg,1}, \ldots, q_{bg,N_{bg}}]
\]

where \( q_{tar,i} \) is an i-vector of the \( i \)-th target speaker from the test data set and \( q_{bg,i} \) is the i-vector of \( i \)-th background speaker from development data set. Then test i-vectors can be represented by sparse vectors as the linear combination of atoms in the dictionary \( D \).

3.1.2. K-SVD Learned Dictionary

The K-SVD algorithm is a generalization of the K-means clustering algorithm seeking best possible representation for each member of training examples with a constraint on sparsity. The dictionary learning problem is represented as,

\[
\min_{D,X} \{ \| Y - DX \|^2_F \}, \text{ s.t. } \| x_i \|_0 \leq T_0
\]

where \( Y \) is the set of training i-vectors of both target and background speakers, \( D \) is the dictionary to be learned, \( X \) is the set of sparse vectors corresponding to \( Y \) and \( T_0 \) is the constraint on sparsity.

The dictionary learning is an iterative process that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data with Singular Value Decomposition (SVD).

3.2. Sparse Representation Speaker Verification System

Once the dictionary is constructed, the test i-vector \( y \) can be represented by a sparse vector \( x \) as a linear combination of dictionary atoms:

\[
y = DX
\]

where \( D \) is an \( F \times N \) matrix, \( F \) is the dimension of the input signal, \( N \) is the number of the training utterances. Usually \( N \gg F \), the solution to equation (4) is undetermined. Hence, this problem turns into solving the optimization problem as follows,

\[
x = \arg\min \| x \|_0, \text{ s.t. } y = DX
\]

However, equation (5) is an NP-hard problem. Recent research on sparse representation demonstrates that if \( x \) is sparse enough, \( l_0 \) norm minimization can be converted to \( l_1 \) norm minimization equivalently as:

\[
x = \arg\min \| x \|_1, \text{ s.t. } y = DX
\]

As for the speaker verification with exemplar dictionary, take the first target speaker for example, if \( y \) is from the target, its sparse representation \( \hat{x}_1 \) has larger component on \( q_{tar,1} \). So, score on the i-th target speaker is found using \( l_1 \) norm ratio:

\[
\text{score} = \frac{\| \hat{x}_1(i) \|_1}{\| \hat{x}_1 \|_1}
\]

While in the case of learned dictionary, there is no direct class label associated with the atoms. Cosine kernel metric is used to measure the similarity between the sparse representation of the claimed speaker and that of the test utterance as follows:

\[
\text{score} = \frac{\langle \hat{x}_{clm}, \hat{x}_{tst} \rangle}{\| \hat{x}_{clm} \| \| \hat{x}_{tst} \|}
\]

where \( \hat{x}_{clm} \) and \( \hat{x}_{tst} \) are the sparse representation of the claimed and test speakers.

4. Experiment

4.1. Intrinsic Variation Corpus

The speaker verification system is evaluated on an intrinsic variation corpus containing 110 (46 males and 64 females) native Chinese university students aged from 18 to 20 years old [11].

Neutral speech at normal rate and volume in Chinese is defined as the base case and the eleven different variation forms are noted as reading, fast, slow, loud, soft, whispered, angry, happy, denasalized, mumbled, and English. The derivation process is presented in Figure 2. These utterances are recorded with a sample rate of 8k. The duration of each utterance is 180 seconds. The channel and noise variability in the recording environment are minimal.
4.2. Experiment Setup

4.2.1. Feature Extraction

We use an energy based voice activity detector to remove the non-speech parts from input data. Mel Frequency Cepstral Coefficients (MFCC) are extracted with 32ms window length and 16ms frame rate from utterances. The 39-dimension MFCC feature vectors are composed of 12 cepstral coefficients and energy, adding their first and second order derivatives.

4.2.2. Database and Parameters

A gender-independent UBM model with 512 Gaussian mixtures is used throughout the experiment. Total variability matrix with 200 columns for i-vector based system, LDA matrix and the background speaker part of the exemplar based dictionary are created using 3069 speech utterances of 30 speakers in development database. The 50-dimension WCCN matrix are also trained using 20 speakers’ utterances in development database.

There are 12 intrinsic variation forms, so the experiment should be repeated 12 times. Each original utterance of a specific style is split into 10 pieces by 18 seconds, which are used for enrollment and test. Hereafter, the “utterance” only refers to the split 18-seconds piece. For each subject, only one utterance in the particular form is used for enrollment while utterances with all the twelve variation forms are used for testing. There are 20 utterances for enrollment and 1719 for testing. Data partitions are showed in Table 1.

<table>
<thead>
<tr>
<th>Function Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM Model</td>
<td>30 speakers 18 hours 12 variation forms</td>
</tr>
<tr>
<td>Total variability LDA matrix</td>
<td>30 speakers 18 hours 12 variation forms</td>
</tr>
<tr>
<td>$D_{bg}$ part of SRC dict.</td>
<td>20 speakers 12 hours 12 variation forms</td>
</tr>
<tr>
<td>WCCN matrix</td>
<td>20 speakers 12 hours 12 variation forms</td>
</tr>
<tr>
<td>Testing Data</td>
<td>20 speakers 12 hours 12 variation forms</td>
</tr>
</tbody>
</table>

4.2.3. Sparse Representation Framework

The sparse representation systems are established on the i-vector framework followed by LDA and WCCN. In order to solve the optimize problem of equation (6), we use Gradient Projection for Sparse Reconstruction (GPSR) [14], which converts the problem to:

$$\text{minimize} \ ||Dx - y||^2_2 + \lambda \ ||x||_1$$  \hspace{1cm} (9)

where $y$ is the i-vector of test or enrollment speakers, $D$ is the dictionary constructed using i-vectors of both target and background speakers, $x$ is the sparse representation of $y$, $\lambda \geq 0$ is the regularization parameter. Finally, Error Equal Rate (EER) and Detection Error Tradeoff (DET) plot are used to measure the performance of the speaker verification system.

5. Results and Discussion

5.1. Robustness of SRC System on Intrinsic Variation

We investigate the performance of the sparse representation systems when the styles of enrollment utterances and test utterances are matched and mismatched. Figure 3 shows the sparse solutions of true speaker and the imposter speaker under the two conditions discussed above. It can be seen that sparse representation is useful to model the intrinsic variation.

Table 2 shows the performances of four variation forms when the enrollment and test utterances are matched and mismatched. The three variation forms, spontaneous, reading and whisper are the most common scenarios in real life application. The average of 12 variation forms are listed because what we expect in the practical condition is that no matter what variation form of utterances are enrolled and tested, the speaker verification system could have robust performance. On average, the sparse representation systems outperform GMM-UBM by about 38%, especially the K-SVD learned dictionary based system achieves an EER of 14.82% when the enrollment utterances and the test utterances are not matched.

5.2. Overall Performance of SRC Systems

Table 3 shows the performance of the GMM-UBM, i-vector and sparse representation (exemplar and learned dictionary) systems for all twelve variation forms measured with EERs. 

$^1$ Results in Table 3 of JFA system are cited from [7]
The average EERs of all enrollment conditions are showed at the bottom of the table. The best results are formatted in bold across each row. It is obvious that K-SVD based sparse representation system obtains lower EERs than the exemplar dictionary based system. For Sparse Representation Classification with learned dictionary system (SRC with learned dict.), the best performance comes from the spontaneous form with an EER of 11.18% while the worst performance comes from the whisper which gives an EER of 25.42%.

5.3. Performance Comparison

The relative reduction on EERs of speaker verification system for all the enrollment conditions are presented in Table 4 compared with GMM-UBM, JFA and i-vector systems. It is obviously that sparse representation systems outperform the other three approaches. Still, the best results are obtained from SRC (learned dictionary) system with the significant relative reduction of EER around 41.55% compared with the GMM-UBM system and 10.47% compared with the i-vector speaker verification system. The corresponding DET curve showed in Figure 4 also demonstrates the robustness of the SRC systems for addressing the intrinsic variation.

<table>
<thead>
<tr>
<th>Systems</th>
<th>GMM-UBM</th>
<th>JFA</th>
<th>i-vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRC (xmplr dict.)</td>
<td>37.17</td>
<td>6.93</td>
<td>3.71</td>
</tr>
<tr>
<td>SRC (learned dict.)</td>
<td>41.55</td>
<td>13.26</td>
<td>10.47</td>
</tr>
</tbody>
</table>

Table 2: EER Relative Reduction of SRC systems (%)

Figure 4: DET plot of the speaker verification systems

6. Conclusion

This paper investigates the robustness of the sparse representation based speaker verification systems on an intrinsic variation corpus. The exemplar dictionary and K-SVD learned dictionary based SRC systems both outperform the GMM-UBM, JFA and i-vector systems, especially the K-SVD learned dictionary based speaker verification system achieves the lowest average EER of 14.23%. Through the experiment on mismatched utterances of enrollment and test, we find sparse representation based verification systems also get 37.17% and 41.44% improvement compared with GMM-UBM, respectively. Our hypothesis has been proved that the redundancy of the dictionary in sparse representation is useful to alleviate the intrinsic variation.

However, performance of SRC systems is less than satisfactory compared with progress made in the field of extrinsic variation. Besides, the database we use now is small and we will enlarge it. In the future, we will try to apply manifold learning and Deep Neural Network (DNN) to model intrinsic variation. In addition, prosodic and paralinguistic information will also be used to present the information related to speaking style.

7. Acknowledgements

The work in this paper is partially supported by the National Basic Research Program of China (973 Program, No.2012CB316401) and National Nature Science Foundation of China (No.61171116).

8. References