Eigen-Prosody Analysis for Robust Speaker Recognition under Mismatch Handset Environment

Zi-He Chen¹, Yuan-Fu Liao² and Yau-Tarng Juang¹

¹Department of Electrical Engineering National Central University, Chung-Li, Taoyuan, 32054, Taiwan
²Department of Electrical Engineering & Institute of Computer, Communication and Control, National Taipei University of Technology, 1, Sec 3, ChungHsiao E. Rd, Taipei 106, Taiwan

Abstract

Most speaker recognition systems utilize only low-level short-term spectral features and ignore high-level long-term information, such as prosody and speaking style. This paper presents a novel eigen-prosody analysis (EPA) approach to capture long-term prosodic information of a speaker for robust speaker recognition under mismatch environment. It converts the prosodic feature contours of a speaker’s speech into sequences of prosody symbols, and then transforms the speaker recognition problem into a full text document retrieval-similar task. Experimental results on the well-known HTIMIT database have shown that, even only few training/test data is available, a remarkable improvement, about 28.7% relative error rate reduction comparing with the GMM/cepstral mean subtraction (CMS) baseline, could be achieved.

1. Introduction

A speaker recognition system in telecommunication network environment needs to be robust against distortion of different handsets or channels, since the recognition accuracy usually drops dramatically while mismatch handsets or channels are encountered. To address this problem, prosodic features, which are known to be less sensitive to handsets and channels mismatch, are attractive recently. Several successful techniques to utilize the prosodic information have been proposed including the distribution [1], the n-gram [2] and the discrete hidden Markov chain (DHMM) [3] approaches.

In the distribution approach, the per-frame pitch and energy values are extracted and modeled using traditional distribution models, such as the conventional Gaussian mixture models (GMMs) [4]. In the n-gram approach, the dynamics of the pitch and energy trajectories are described by sequences of symbols and modeled by the n-gram statistics. In the HMM method, the sequences of prosody symbols are further modeled by the state observation and transition probabilities. However, the distribution approach may not adequately capture the temporal dynamic information of the prosodic feature sequences and the n-gram and HMM methods usually require large amount of training/test data to get a reasonable performance.

In this paper, a novel eigen-prosody analysis (EPA) approach is proposed to add robustness to conventional cepstral features-based GMMs speaker recognition system under the situation of mismatch handsets/channels while only few training/test data is available. The basic idea of EPA is to extract and convert the prosodic feature contours of a speaker’s enrollment speech into sequences of prosody symbols, and to treat the sequences of prosody symbols as a text document which records the detail prosody/speaking style of the specific speaker. The speaker’s test speech is also converted into sequences of prosody symbols and uses as the query keywords to recall the most related document (speaker). By this way, the speaker recognition problem is transformed into a full-text document retrieval-similar task and some well-established information retrieval algorithms, such as latent semantic indexing (LSI) [5], could be directly applied to build an eigen-prosody space to represent the space of speakers. Moreover, it is possible to get benefit form the prosodic information even when only few speech data is available.

This paper is organized as follows. Section 2 gives the information of the HTIMIT database. Section 3 describes the proposed eigen-prosody analysis in detail. Section 4 reports the experimental results evaluated on the HTIMIT database [6]. Some conclusions are given in the last section.

2. HTIMIT database

To evaluate the proposed approach, the well-known HTIMIT database [6], which was recorded for studying the handset mismatch problem, was chosen. HTIMIT was a reading style database and was recorded using 8 kHz sampling rate. There were in total 384 speakers, each uttered ten utterances using a Sennheizer head-mounted microphone (called senh). The set of 384*10 utterances was playback and recorded through nine other different handsets include four carbon button (called cb1, cb2, cb3 and cb4), four electret (called el1, el2, el3 and el4) handsets, one portable cordless phone (called pt1). There were in total 384*10*10 utterances.

In this study, all experiments were performed on 346 speakers (173 females and 173 males). The first seven utterances of each speaker from the senh subset were used as the enrollment speech. The last three utterances of each speaker from each handset (ten handsets in total) were used as the evaluation data, respectively. Moreover, the pitch contours of all utterances were extracted using the popular Wavesurfer/Whack sound toolkit [7].

3. The proposed eigen-prosody analysis approach

The basic idea of the proposed EPA method is to transform the speaker recognition issue into a full text document retrieval-similar problem in order to capture the long-term
prosodic behaviors of a speaker and to avoid the requirement of large amount of training/test speech data.

The procedures (see Fig.1) includes: (1) stylizing the prosodic feature contours of speakers into sequences of prosody symbols, (2) segmenting the sequences to extract important prosody keywords, (3) calculating the occurrences statistics of these prosody keywords for each speaker to form a prosody keyword-speaker occurrence matrix, (4) applying the singular values decomposition (SVD) [5] to decompose the prosody keyword-speaker occurrence matrix to build an eigen-prosody space to represent the constellation of speakers, and (5) measuring the speaker distance using the cosine of the angle between two speaker vectors in the eigen-prosody space. The detail procedures (see Fig.1) are described in the following subsections.

![Figure 1: The proposed scheme of the eigen-prosody analysis for robust speaker recognition: (a) construction of the prosody keyword-speaker occurrence matrix and (b) eigen-prosody space analysis using SVD.](image)

### 3.1. Prosodic feature contours stylization

First of all, the prosodic feature contours are needed to be converted into sequences of prosody symbols. In this study, syllables are chosen as the basic processing units. Four types of prosodic features including pitch contour, log-energy contour, pause duration and pitch jump between two vowels are used here.

The slope of the pitch contour of a vowel segment and the difference between the average log-energy of two vowel segments are computed and quantized to three levels, i.e., pitch rising (“/”) and flat pitch (“\”) and pitch falling (“\”) and volume increasing (“<”), equal volume (“=”), and volume decreasing (“>”), respectively. The pause durations are also classified into three categories, i.e., short (“S”), middle (“M”) and long (“L”) pauses. Finally, the value of the pitch jump between two vowels is labeled as jump up (“U”), flat jump (“F”) and jump down (“D”).

A typical example of the stylization is shown in Figure 2, where a segment of a HTIMIT utterance is labeled as “<SU />LU \>MD”. By this way, four-character text symbols are used to describe the prosody phenomenon of a syllable. The sequences of the prosody symbols generated from the enrollment utterances of a speaker are treated as a prosody text document to book the prosodic behaviors of the speaker.

![Figure 2: Illustration of the stylization of a segment of a HTIMIT utterance into a sequence of prosody symbols, i.e., “<SU />LU \>MD”](image)

### 3.2. Prosody keyword extraction

After the stylization, the prosody text documents of all speakers are searched to find important prosody keywords in order to form a prosody keywords dictionary. Essentially, the compilation of the dictionary can be treated as an unknown-word extraction problem [8] and an n-gram approach for finding high frequency collocations is adopted.

First, all possible combinations of the prosody words, including single words and word pairs (uni-gram and bi-gram), are listed and their frequency statistics are computed. After calculating the histogram of all prosody words, frequency thresholds are set to leave only high frequency ones. One typical histogram example on analyzing all enrollment utterances is shown in Figure 3. By examining the figure, one can find that some highest frequency single prosody words, such as “<SF” and “<LF”, are really very common phenomena in reading speech. However, some lowest frequency prosody word pairs, such as “/>SU” and “/>MF/>MD”, are less common, on the other hand. These rare keywords will be the informational cues to discriminate between different speakers.

![Figure 3: The histogram of the found prosody keywords from the enrollment set of the HTIMIT database.](image)
3.3. Prosody keyword-speaker occurrence matrix statistics

The prosody text document of each speaker is then parsed using the generated prosody keywords dictionary by simply giving higher priority to longer words. The occurrence counts of keywords of a speaker are booked in a prosody keyword vector to represent the long-term prosodic behaviors of the specific speaker. Therefore, the prosody keyword-speaker occurrence matrix \( A \) is made up of the collection of all speaker prosody keyword lists vectors.

It is worthy to note that the prosody keyword-speaker occurrence matrix \( A \) is a very high dimensional sparse matrix with each column vector represents the relationship of a speaker to these prosody keywords and each row vector represents the relationship of a prosody keyword to all speakers.

Moreover, to emphasize the uncommon keywords and to deemphasize the very common ones, the inverse document frequency (IDF)\(^5\) weighting method is applied. Specifically, \( A \) is normalized by the log of the global counts that the keywords occur in all speakers.

3.4. Eigne-prosody analysis

In order to reduce the dimension of the prosody space, the sparse prosody keyword-speaker occurrence matrix \( A \) is further analyzed using SVD to find a more compact eigen-prosody space.

Specifically, given an \( m \) by \( n \) (\( m >> n \)) matrix \( A \) of rank \( R \), \( A \) is decomposed into the product of three matrices defined as:

\[
A = U \sum V^T
\]

where \( U = (u_1, \ldots, u_n) \) and \( V = (v_1, \ldots, v_n) \) and

\[
\sum = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),
\]

\[
\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n = ... = \sigma_n = 0 \ldots \ldots (2)
\]

The \( \sigma_i \)’s are called the singular values of \( A \), \( u_i \) and \( v_i \) are the left and right singular vectors associated with \( \sigma_i \), \( i = 1 \sim R \), and the \( i \)-th singular triplet of \( A \) is defined as \( (u_i, \sigma_i, v_i) \). The matrix \( A \) can also be expressed as a sum of \( R \) rank-one matrices:

\[
A = \sum_{i=1}^{R} \sigma_i u_i v_i^T \quad \ldots \ldots (3)
\]

and further approximated using only the largest \( K \) singular values as:

\[
A = U \sum V^T = A_K = U_K \sum_K V_K^T \quad \ldots \ldots (4)
\]

where \( A_K, U_K, V_K \), and \( \sum_K \) matrices are the rank reduced matrices of the respective matrices.

All speaker-to-speaker prosody keywords list vectors inner products based on this approximation are given by

\[
A_K^T A_K = V_K \sum_K V_K^T \quad \ldots \ldots (5)
\]

This implies that it is possible to compute meaningful association values among speakers, even if the speakers do not have any prosodic keyword in common.

By this way, EPA is capable to give a compact eigen-prosody space to model the long-term prosodic behaviors of the speakers. A typical example of the eigen-prosody space by the analysis of the senh enrollment set is shown in Figure 4. It is worthy to note that speakers with flat speaking style are located on the lower right-hand corner. On the other hand, speakers with more dynamic behaviors are distributed on the upper right-hand corner. Therefore, it is possible to separate speakers with different speaking style in the eigen-prosody space.

3.5. Score measurement and system fusion

The problem of the speaker recognition is now formulated as a pseudo-document testing as in the LSI approach. The test utterances of a speaker are first stylized and parsed to form a pseudo query document \( y_v \), i.e., the prosody keywords list vector of the test speaker, and then transformed into a query vector \( v_o \) in the compact eigen-prosody speaker space by

\[
v_o = y_v U \sum^{-1} \quad \ldots \ldots (6)
\]

The distance between the test speaker and the \( i \)-th registered speaker is defined as the cosine of the angle between input query vector \( y_v \) and the \( i \)-th speaker vector \( v_{K,i} \) in the eigen-prosody space, i.e.,

\[
S_i^P (v_o, v_{K,i}) = \frac{v_o^T v_{K,i}}{\|v_o\| \|v_{K,i}\|} \quad \ldots \ldots (7)
\]

Finally, the score of the \( i \)-th speaker generated from the prosody models \( S_i^P \) is nonlinearly fused with the scores of the conventional GMM/CMS-based recognizer \( S_i^{G} \) to get the final score \( S_i^F \) using a nonlinear fusion function defined as:

\[
S_i^F = \alpha \text{sigmoid}(S_i^P) + (1 - \alpha) \text{sigmoid}(S_i^G) \quad \ldots \ldots (8)
\]

where \( \alpha \) is a weighting constant.

Figure 4: The distribution of the prosody keywords and speakers castellation on the two dimensional eigne-prosody space.
4. Experiments

In this paper, all experiments were performed on 346 speakers in HTIMIT. The first seven utterances of each speaker from the senh subset were used to build a 32-order speaker GMMs. HTIMIT was recorded using 8 kHz sampling rate, but only the speech signal in 300-3200 Hz was used to reduce the handset/channel distortions. 38 mel-frequency cepstral coefficients (MFCCs), including 12 MFCCs, 12 ΔMFCCs, 12 Δ^2-MFCCs, Δ-log energy and Δ^2-log energy, were computed with window size of 30 ms and frame shift of 10ms. Besides, for training the eigen-prosody model, slope of the pitch contour of a vowel segment, change of the mean log-energy and pitch jump between two vowel segments and pause duration between two syllables were measured. The evaluation was performed on the last three utterances of each speakers from each handset (ten handsets in total). Furthermore, CMS was used to alleviate the handset mismatch while building a 32-order GMM for each speaker.

First, the speaker recognizer using the GMM/CMS method to remove the handset bias was evaluated as the baseline. The result is shown in Table 1. An average recognition rate of 62.7% was achieved. A comparison between the baseline and the one reported in [6] showed that the result was not bad.

Then, the proposed EPA approach was evaluated. From the first seven utterances of each speaker from the senh subset, about 917 prosody keywords were extracted. The occurrence counts of the prosody keywords lists of all 346 speakers were collected to form a sparse 917*346 dimensional prosody keywords-speakers occurrence matrix A. The matrix A was further analyzed by SVD to find a more compact eigen-prosody space.

The scores of the EPA method were fused together with the scores of the GMM/CMS to form the final recognition scores. Table 1 reports the recognition results of the fused EPA+GMM/CMS system. Can be found from the table, a significant improvement was achieved. While using 50, 100 and 200 dimensional eigen-prosody space, 22.3%, 26.3% and 28.7% relative error rate reductions were achieved, respectively. This indicates that EPA is a promising approach for robust speaker recognition under the mismatch handset condition.

5. Conclusions

This paper presents a novel eigen-prosodic analysis approach to capture the long-term prosodic information for robust speaker recognition under mismatch environment. Unlike the n-gram or DHMM-based methods, which need a lot of speech data to reach a reasonable performance, EPA requires only few training/test utterances (in this case, seven/three utterances). Experimental results on the HTIMIT database have shown that, even only few training/test data is available, a remarkable improvement, about 28.7% relative error rate reduction comparing with the conventional GMM/CMS baseline, could be achieved. It is therefore a promising method to alleviate the problem of the mismatch handsets for robust speaker recognition. Furthermore, EPA can be applied to against mismatch channel as well.

6. Acknowledgement

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 92-2213-E-027-037 and Ministry of Education under the project with contract A-93-E-FA06-4-4.

7. References


<table>
<thead>
<tr>
<th># of eigen-vectors</th>
<th>senh</th>
<th>cb1</th>
<th>cb2</th>
<th>cb3</th>
<th>cb4</th>
<th>el1</th>
<th>el2</th>
<th>el3</th>
<th>el4</th>
<th>pt1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GMM/CMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>85.8</td>
<td>75.7</td>
<td>77.7</td>
<td>32.4</td>
<td>45.7</td>
<td>76.0</td>
<td>69.1</td>
<td>69.4</td>
<td>73.4</td>
<td>66.5</td>
<td>67.2</td>
</tr>
<tr>
<td>30</td>
<td>92.5</td>
<td>81.5</td>
<td>79.2</td>
<td>33.8</td>
<td>46.0</td>
<td>80.3</td>
<td>73.1</td>
<td>71.7</td>
<td>73.1</td>
<td>67.1</td>
<td>69.8</td>
</tr>
<tr>
<td>50</td>
<td>94.2</td>
<td>81.5</td>
<td>80.3</td>
<td>35.3</td>
<td>47.7</td>
<td>81.8</td>
<td>73.7</td>
<td>71.4</td>
<td>75.8</td>
<td>68.2</td>
<td>71.0</td>
</tr>
<tr>
<td>100</td>
<td>96.8</td>
<td>82.9</td>
<td>79.8</td>
<td>39.6</td>
<td>50.3</td>
<td>80.6</td>
<td>75.1</td>
<td>76.3</td>
<td>74.0</td>
<td>69.7</td>
<td>72.5</td>
</tr>
<tr>
<td>200</td>
<td>98.3</td>
<td>83.8</td>
<td>82.9</td>
<td>39.9</td>
<td>51.4</td>
<td>81.2</td>
<td>75.8</td>
<td>76.0</td>
<td>74.3</td>
<td>70.8</td>
<td>73.4</td>
</tr>
<tr>
<td><strong>EPA+ GMM/CMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>77.7</td>
<td>76.0</td>
<td>73.7</td>
<td>26.3</td>
<td>39.6</td>
<td>72.8</td>
<td>64.2</td>
<td>65.3</td>
<td>70.2</td>
<td>61.6</td>
<td>62.7</td>
</tr>
<tr>
<td>30</td>
<td>85.8</td>
<td>75.7</td>
<td>77.7</td>
<td>32.4</td>
<td>45.7</td>
<td>76.0</td>
<td>69.1</td>
<td>69.4</td>
<td>73.4</td>
<td>66.5</td>
<td>67.2</td>
</tr>
<tr>
<td>50</td>
<td>92.5</td>
<td>81.5</td>
<td>79.2</td>
<td>33.8</td>
<td>46.0</td>
<td>80.3</td>
<td>73.1</td>
<td>71.7</td>
<td>73.1</td>
<td>67.1</td>
<td>69.8</td>
</tr>
<tr>
<td>100</td>
<td>94.2</td>
<td>81.5</td>
<td>80.3</td>
<td>35.3</td>
<td>47.7</td>
<td>81.8</td>
<td>73.7</td>
<td>71.4</td>
<td>75.8</td>
<td>68.2</td>
<td>71.0</td>
</tr>
<tr>
<td>200</td>
<td>96.8</td>
<td>82.9</td>
<td>79.8</td>
<td>39.6</td>
<td>50.3</td>
<td>80.6</td>
<td>75.1</td>
<td>76.3</td>
<td>74.0</td>
<td>69.7</td>
<td>72.5</td>
</tr>
</tbody>
</table>