Unseen Handset Mismatch Compensation Based on A Priori Knowledge Interpolation for Robust Speaker Recognition

Jyh-Her Yang and Yuan-Fu Liao

Department of Electrical Engineering & Institute of Computer, Communication and Control, National Taipei University of Technology, 1, Sec 3, ChungHsiao E. Rd, Taipei 106, Taiwan
yfliao@ntut.edu.tw

Abstract

Unseen handset mismatch is the major source of performance degradation for speaker recognition in telecommunication environment since handset distortions are tightly coupled with speaker characteristics. In this paper, a soft-decision unseen handset characteristics estimation method based on a priori knowledge interpolation is proposed to decouple the characteristics of the unseen handset and speaker for robust speaker recognition. Experimental results on HTIMIT database showed that the proposed method improved the speaker recognition rate for both seen and unseen handsets.

1. Introduction

A speaker recognition system in telecommunication network environment needs to be robust against distortion of different handsets, where some handsets may not be seen, i.e., unseen handsets, and will cause significant performance degradation. However, knowledge about the distortion of those unseen handsets, i.e., the characteristics of the unseen handsets, is usually not available beforehand. Moreover, the characteristics of the handset and speaker are usually tightly mixed or coupled together. To separate the characteristics of the handset and speaker is essentially a difficult one-to-many mapping problem, unless some knowledge about the unseen handset is available in advance.

Several successful techniques have been proposed to address this handset mismatch problem, including the cepstral mean subtraction (CMS) [1] signal bias removal (SBR) [2], handset detector-based approaches [3,4] and speaker model synthesis (SMS) [5]. The CMS and SBR methods compensate the handset distortion by subtracting the long-term average of handset-distorted cepstral vectors. The handset detector-based approach discriminates the types of handsets, such as carbon button or electret, to remove the distortion or to adapt the underline models using the pre-trained handset-dependent characteristics. The SMS uses pre-trained speaker-independent channel transformations to synthesize a speaker model that corresponds to the channel of the testing session.

However, the CMS and SBR methods may remove not only the characteristics of handset but also the speaker’s as well. The detector- and synthesize-based approaches usually use a hard-decision handset selector front-end and may have the difficulty to deal with speech coming from unseen handsets. They may just select the most likely handset from a set of seen handsets, simply reject them as out-of-handset (OOH) [3] or have to fall back to CMS-based system.

Essentially, the characteristics of the handset and speaker are tightly mixed or coupled together, a priori knowledge about the characteristics of the handsets that speaker are likely to use is crucial to separate each other. In this paper, a priori knowledge interpolation-based handset characteristic estimator (PKIHCE) is proposed to alleviate the problem of the unseen but unavoidable handsets.

Assuming there is a set of seen handsets, the mismatch characteristics of these seen handsets and the enrollment handset are measured using maximum likelihood linear regression (MLLR) [6]. The set of MLLR transformation parameters of the seen handset is collected and used as the a priori knowledge to represent the space of handsets. The characteristics of an unseen handset are then be estimated using a posteriori probabilities-based interpolation algorithm generated from a Gaussian mixture model (GMM)-based handset classifier [7]. Besides, the reliability of applying the a priori knowledge to an input test utterance is measured using a divergence-based measurement to guide the interpolation. Finally, to handle some outliers which are totally uncovered by the a priori knowledge, the recognition scores are further fused with the scores of a GMM/CMS-based speaker recognizer.

This paper is organized as follows: Section 2 briefly reviews the MLLR method for seen handset characteristics estimation. The proposed unseen handset compensation method is described in Section 3 including the a priori knowledge interpolation and the score fusion. In Section 4, experimental results are reported in the well-known HTIMIT database [8]. Some conclusions are given in the last section.

2. MLLR seen handset characteristics estimation

MLLR is a mature method to adapt a well trained speech model to a new environment by transforming the model parameters using a set of regression-based transforms defined by

$$\hat{u}_m = A_m \cdot u_m + b_m,$$

(1)

where $\hat{u}_m$ is the adapted mean, $u_m$ is the original mean, $A_m$ is the transformation matrix and $b_m$ is the bias of the $m$-th mixture component.
In this paper, MLLR is used to measure the transformation relationship between a set of \( N \) seen handsets that speakers are likely to use and the enrollment handset, i.e., \( A_{n,n} \) and \( b_{n,n} \), where \( n \) is the index of the seen handset. The set of the transformation matrix and bias, \( \{A_{n,n}, b_{n,n}, n = 1 \sim N\} \), is then collected as the \( \textit{a priori} \) knowledge to represent the space of the handsets.

It is worthy to note that MLLR can only be used to measure the mismatch between the enrollment and seen handsets in advance. It can not be used to directly adapt the enrollment speaker models to an unseen handset online, since the difference between speakers will be eliminated and all speaker models will collapse together.

3. PKIHCE unseen handset characteristics estimation

To compensate an unseen handset, the characteristics of the unseen handset are estimated using the linear combination of the characteristics of the set of seen handsets. Specifically, before performing recognition, we compute one set of transformation parameters for each type of handsets using MLLR method mentioned in the previous section. During an evaluation session, the unseen handset used by the speaker is compensated by an interpolation of the set of transformation parameters of the seen handsets.

The block diagram of the proposed PKIHCE approach is illustrated in Figure 1. It includes a PKIHCE front-end/GMM-based speaker recognizer, a CMS front-end/GMM-based speaker recognizer and a nonlinear score fusion module. The detailed block diagram of the PKIHCE front-end is shown in Figure 2 where the characteristics of the seen handsets are interpolated to estimate the characteristics of an unseen handset. The function of the PKIHCE front-end and the score fusion module will be described in detail in the following subsections.

![Figure 1: A block diagram of the proposed PKIHCE scheme for robust speaker recognition.](image1)

![Figure 2: The proposed scheme of the \( \textit{a priori} \) knowledge interpolation-based handset characteristics estimator.](image2)

3.1. Handset space and \( \textit{a posteriori} \) probabilities

Assume that there are \( N \) seen handsets that speakers are likely to use. In the training phase, the handset characteristics \( H = \{h_1, h_2, \ldots, h_N\} \) of these \( N \) seen handsets and their corresponding GMMs \( \Lambda = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_N\} \) are computed from the speech observations of each handset \( O_n = \{o_{n,1}, o_{n,2}, \ldots, o_{n,T}\} \), using the expectation-maximum (EM) [9] algorithm and MLLR, respectively. The set of the handset characteristics \( H \) and GMMs \( \Lambda \) of those seen handsets are then used as the kernel functions to represent the space of handsets.

Furthermore, the input speech observations \( O' = \{o'_{1}, o'_{2}, \ldots, o'_{T}\} \) of a speaker from an unseen handset are fed into the set of seen handset GMMs \( \Lambda \) to compute \( N \) likelihoods \( L(O' \mid \Lambda_n) \) function. These \( N \) likelihoods \( L(O' \mid \Lambda_n) \) function are transformed into the \( \textit{a posteriori} \) probabilities \( p(\Lambda_n \mid O') \) using the following equation:

\[
p(\Lambda_n \mid O') = \frac{\exp\left(L(O' \mid \Lambda_n)\right)}{\sum_{i=1}^{N} \exp\left(L(O' \mid \Lambda_i)\right)} \quad \ldots \quad (2)
\]

3.2. Reliability of the \( \textit{a priori} \) handset knowledge

The reliability of applying the \( \textit{a priori} \) knowledge to the input speech observation \( O' \) is then be estimated using a divergence measure to compare the distribution of the \( \textit{a posteriori} \) vector \( P \):

\[
P = \left[ p(\Lambda_1 \mid O'), p(\Lambda_2 \mid O'), \ldots, p(\Lambda_N \mid O') \right] \quad \ldots \quad (3)
\]

with a uniform distribution reference vector \( U \):
\[ U = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ N & N & \ldots & N \end{bmatrix}, \quad \ldots \quad (4) \]

The equation of the divergence measure (or called Jensen difference) \( J(P, U) \) is defined by

\[ J(P, U) = S\left( \frac{P + U}{2} \right) - \frac{1}{2} \left[ S(P) + S(U) \right], \quad \ldots \quad (5) \]

where \( S(\cdot) \), called the Shannon entropy, is given by

\[ S(Z) = - \sum_{n=1}^{N} z_n \log z_n, \quad \ldots \quad (6) \]

and \( z_n \) is the \( n-th \) component of \( Z \). If \( P \) approximates \( U \), \( J(P, U) \) is small, then the a priori knowledge is not reliable. It means the observation speech \( O' \) may come from a handset never seen before. On the other hand, if \( P \) is far away from \( U \), \( J(P, U) \) is large, the a priori knowledge is reliable. It means the observation speech \( O' \) may come from a handset which is very similar to some of those seen handsets.

The divergence measure \( J(P, U) \) is further converted into the reliability measure \( R \) using a zero-one sigmoid \( \cdot \) function defined by

\[ R = \frac{1}{1 + \exp(-\lambda(-J(P, U) + \beta))}, \quad \ldots \quad (7) \]

where \( \lambda \) and \( \beta \) are the parameters of the sigmoid \( \cdot \) function.

### 3.3. Interpolation and score fusion

The reliability measure \( R \) is then utilized to generate the weights \( \alpha_n \) for the linear interpolation algorithm using the following equation:

\[ \alpha_n = \frac{\exp \left( R \cdot L(O | \Lambda_n) \right)}{\sum_{n=1}^{N} \exp \left( R \cdot L(O | \Lambda_n) \right)}, \quad \ldots \quad (8) \]

Then, the estimate \( \hat{h} \) of the characteristics of the unseen handset is defined as:

\[ \hat{h} = \sum_{n=1}^{N} \alpha_n h_n, \quad \ldots \quad (9) \]

And the distortion of the unseen handset is compensated by adapting all speaker GMMs using the MLLR equation defined in Equation (1).

Finally, to handle some outliers which are totally uncovered by the a priori handset knowledge, the recognition scores of the PKIHCE \( S_1 \) and the scores of the conventional GMM/CMS-based speaker recognizer \( S_2 \) are nonlinearly fused to get the final scores \( S_f \) using the following equation:

\[ S_f = \frac{1}{\gamma} \log \left[ \frac{\exp \gamma S_1 + \exp \gamma S_2}{2} \right], \quad \ldots \quad (10) \]

where \( \gamma \) controls the degree of nonlinear fusion. It is worthy to note that by using this fusion function, the final score will be the dominant score if one score is much larger that the other, or be the average score, if their values are close.

### 4. Experiments

To evaluate the performance of the proposed approach, the HTIMIT database [8], which was recorded for studying the handset mismatch problem, was chosen. There were in total 384 speakers, each gave ten utterances using a Sennheizer head-mounted microphone (called senh). The set of 384*10 utterances was then 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, and one portable cordless phone (called pt1). Therefore, there are in total 384*10*10 utterances.

HTIMIT uses 8 kHz sampling rate. However, only the speech signals between 300~3200 Hz were used to avoid handset/channel distortions. For training the speaker GMMs, 38 mel-frequency cepstral coefficients (MFCCs) including 12 MFCCs, 12 \( \Delta \)-MFCCs, 12 \( \Delta^2 \)-MFCCs, \( \Delta \)-log energy and \( \Delta^2 \)-log energy were computed with window size of 30 ms and frame shift of 10ms.

In this paper, all experiments were performed on 356 speakers (178 females and 178 males) in this database using first 16 seconds speech of each speaker from the senh (enrollment handset) subset to build a 32-order speaker GMM for each speaker and testing on other ten four-second speech sessions of each speaker from 10 evaluation handsets, respectively. Furthermore, the leave-one-out style cross-validation strategy was used to evaluate the proposed method in the unseen handset mismatch situation which in turn used nine handsets as the seen handsets to build nine 64-order handset GMMs and estimate nine set of handset characteristics, \( \{A_{n \alpha}, b_{n \alpha}, n = 1 \sim 9\} \), using MLLR algorithm, respectively, as the a priori knowledge and left one handset out as the unseen test handset.

First, the speaker recognizer using the conventional CMS method to remove the handset bias was evaluated as the baseline. The result was shown in Table 1; an average
recognition rate of 55.8% was achieved. Compared with the one reported in [8], the baseline results were not bad.

The proposed PKIHCE approach without score fusion was then tested using the cross-validation style test. Different number of MLLR classes, including one for speech or two for consonant and vowel classes were evaluated, respectively. The performance was significantly improved to 64.8% (see Table 1). Lastly, the PKIHCE approach with fusion was applied and showed that it could further improve the average recognition rate to 65.2% (see Table 1).

Moreover, the recognition results for the nine unseen handset turns in the cross-validation test were separated and shown in Table 2. It showed that PKIHCE+Fusion approach could increase the performance from 54.6% to 60.2%. Therefore, the results in Table 1 and 2 showed that the proposed method could improve the recognition accuracy for both seen and unseen handsets.

Table 1: The average speaker recognition rates (%) in the cross-validation evaluation achieved by the CMS front-end, PKIHCE front-end without and with fusion methods, respectively, on the HTIMIT database.

<table>
<thead>
<tr>
<th># of MLLR classes</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM/CMS</td>
<td>55.8</td>
</tr>
<tr>
<td>PKIHCE</td>
<td>64.8</td>
</tr>
<tr>
<td>PKIHCE + GMM/CMS</td>
<td>65.2</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, the PKIHCE approach is proposed and the experimental results have shown that it can improve the recognition accuracy for both seen and unseen handsets. Unlike conventional hard-decision handset classifier-based approaches, which use only coarse carbon/electret classes and have to reject out-of-handsets or fall back to CMS-based system, the proposed soft-decision method can seamlessly deal with unseen handsets. It is therefore a promising method to alleviate the problem of the unseen but unavoidable handsets for robust speaker recognition.

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 2: The speaker recognition rates (%) of the nine unseen handset turns in the cross-validation evaluation achieved by the CMS and PKIHCE+fusion approaches, respectively, on the HTIMIT database.

<table>
<thead>
<tr>
<th># of MLLR classes</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>GMM/CMS</td>
<td>-</td>
<td>62.9</td>
<td>63.2</td>
<td>25.3</td>
<td>35.4</td>
<td>68.5</td>
<td>61.0</td>
<td>55.6</td>
<td>63.5</td>
<td>55.9</td>
</tr>
<tr>
<td>PKIHCE</td>
<td>1</td>
<td>73.9</td>
<td>67.7</td>
<td>28.4</td>
<td>43.5</td>
<td>74.2</td>
<td>60.1</td>
<td>61.2</td>
<td>66.0</td>
<td>60.4</td>
</tr>
<tr>
<td>PKIHCE + GMM/CMS</td>
<td>2</td>
<td>72.8</td>
<td>71.1</td>
<td>27.2</td>
<td>41.3</td>
<td>75.3</td>
<td>63.5</td>
<td>62.9</td>
<td>68.0</td>
<td>60.1</td>
</tr>
</tbody>
</table>