Combination method of Bone-conduction Speech and Air-conduction Speech for Speaker Recognition

Satoru Tsuge\textsuperscript{1}, Takashi Osanai\textsuperscript{2}, Hisanori Makinae\textsuperscript{2}, Toshiaki Kamada\textsuperscript{2}  
Minoru Fukumi\textsuperscript{1}, Shingo Kuroiwa\textsuperscript{3}  

\textsuperscript{1} The University of Tokushima, Japan  
\textsuperscript{2} National Research Institute of Police Science, Japan  
\textsuperscript{3} Chiba University, Japan

Abstract

Recently, some new sensors, such as bone-conductive microphones, throat microphones, and non-audible murmur (NAM) microphones, besides conventional condenser microphones have been developed for collecting speech data. Accordingly, some researchers began to study speaker and speech recognition using speech data collected by these new sensors. We focus on bone-conduction speech data collected by the bone-conductive microphone. This paper proposes a novel speaker identification method which combines “bone-conduction speech” and “air-conduction speech”. The proposed method conducts speaker identification by integrating the similarity calculated by air-conduction speech model and similarity calculated by bone-conduction speech model. For evaluating the proposed method, we conduct the speaker identification experiment using part of a large bone-conduction speech corpus constructed by National Research Institute of Police Science, Japan (NRIPS). Experimental results show that the proposed method can reduce a identification error rate of air-conduction speech and bone-conduction speech. Especially, the proposed method achieves that the average error reduction rate from air-conduction speech to the proposed method is 35.8%.

Index Terms: Speaker identification, Bone-conduction speech, Air-conduction speech

1. Introduction

In recent years, the use of portable terminals, such as cellular phones and PDAs (Personal Digital Assistants), has become increasingly popular. Additionally, it is expected that almost all appliances will connect to the Internet in the future. As a result, it will become increasingly popular to control these appliances using mobile and hand-held devices. Therefore, we believe that a speaker recognition system will be used as a convenient personal identification system in such cases. Actually, a lot of researchers are involved in the study of speaker recognition.

There are many speech corpora\textsuperscript{[1]}, such as YOHO, TIMIT, NIST, KIND for studying speaker recognition. A lot of researches which use these speech corpora are reported\textsuperscript{[2]}\textsuperscript{[3]}\textsuperscript{[4]}. Moreover, there is the Japanese speech corpus\textsuperscript{[5]} for studying Japanese speaker recognition.

Recently, some new sensors, such as bone-conductive microphones, throat microphones, and non-audible murmur (NAM) microphones, besides the conventional condenser microphone were developed for collecting speech data. Using speech data collected by these new sensors, some researches began to study speaker and speech recognition\textsuperscript{[6]}\textsuperscript{[7]}. In this paper, we focus on speech data collected by the bone-conductive microphone. Actually, some researcher use the speech data collected by the bone-conductive microphone\textsuperscript{[6]}. In our previous work, we studied speaker identification using bone-conduction speech. In this paper, to improve the speaker identification performance, we propose a novel speaker identification method which uses bone-conduction speech with air-conduction speech collected by a condenser microphone. The proposed method conducts speaker identification by integrating the similarity calculated by the speaker model of air-conduction speech and similarity calculated by the speaker model of bone-conduction speech. For evaluating the proposed method, we conduct the speaker identification experiments using part of a large speech corpus constructed by National Research Institute of Police Science, Japan (NRIPS)\textsuperscript{[8]}. The number of speakers collected in this corpus is more than 600. In addition, the speech data in this corpus are collected by the condenser microphone and the bone-conductive microphone at the same time. The details of this speech corpus are described in section 2.

This paper will continue as follows: Section 2 introduces the speech corpus collected by NRIPS. In section 3, we propose a novel speaker identification method which combines bone-conduction speech and air-conduction speech. In section 4, we show the speaker identification experimental results using a part of this corpus and compare the results of the proposed method with those of conventional methods. Finally, we summarize this work in section 5.

2. Japanese Large Speech Corpus

In this section, we introduce Japanese large speech corpus collected by NRIPS. Figure 1 shows the overview of the recording system. As shown in this figure, this system collects four-channel speech data. These channels are as follows:

- The speech data of channel 1 is collected by the electric condenser microphone.
- The speech data of channel 2 is collected by the bone-conductive microphone.
- The speech data of channel 3 is collected by the electric condenser microphone over the mobile-phone network.
- The speech data of channel 4 is collected by the bone-conductive microphone over the mobile-phone network.

In this corpus, “SONY ECM-23FS” was used as the electric condenser microphone, “TEMCO EMLIA SB 9040” was used as the bone-conductive microphone, and “EDIROL R-4” was used as the digital recorder. After collection, the speech data of the four channels are synchronized. The time difference in the four channel speech data is adjusted to less than 20 msec.
2.1. Speakers

The speech data of each speaker for two days is collected in this corpus. The number of male and female speakers at the first session are 336 and 328 respectively. At the second session, the number is 313 and 319, respectively. The same speakers participated in both sessions. In this corpus, each speaker uttered the recording-set (described in 2.2), twice on each recording session. Hence, there is the speech data of the recording-set on four times per speaker for 632 speakers, but there is speech data of the recording-set for only two times per speaker for 32 speakers.

2.2. Recording set

The following recording-sets are used in this corpus:

- Japanese syllable (The number of words is 100.),
- Japanese isolated words (The number of words is 66.),
- Japanese sentences (The number of sentences is 14.),
- ATR Japanese balance sentences (The number of sentences is 50.).

The contents of the Japanese words are digit strings, pronouns, and so on.

Each speaker uttered this recording-set twice per recording session. Hence, there are 920 utterances (230 (number of utterances) × 2 (times per day) × 2 (recording days)) per speaker in this corpus.

3. Combination method of bone-conduction speech and air-conduction speech for speaker identification

In this section, we propose the combination method of bone-conduction speech and air-conduction speech for speaker identification. Figure 2 shows the flow of the proposed method. The proposed method does not combine feature parameters of air- and bone-conduction speech but combines the distortions of the two. The details of the proposed method are described as follows.

In speaker model training process, the speaker model of air-conduction speech, $\lambda_{s_a}^i$, and the speaker model of bone-conduction, $\lambda_{s_b}^i$, are individually trained. In this paper, we used the vector quantization, VQ, centroids for speaker model.

In the speaker identification process shown in figure 2, using these speaker models, we calculate VQ distortion between the feature parameters of input speech, $X = \{(x_{a}^1, x_{b}^1), \ldots, (x_{a}^N, x_{b}^N)\}$; $x_{a}^i$ and $x_{b}^i$ are the feature parameter of air-conduction speech and bone-conduction speech of frame $i$, respectively, in following equation.

$$D_s = \sum_{i=0}^{N-1} \left( \alpha_i \cdot d(x_{a}^i, \lambda_{s_a}^i) + (1 - \alpha_i) \cdot d(x_{b}^i, \lambda_{s_b}^i) \right),$$  \hspace{1cm} (1)

where $D_s$ indicates the VQ distortion of speaker, $s$. $\alpha_i$ indicates a weight value. $x_{a}^i$ and $x_{b}^i$ are the feature parameter of air-conduction speech and the feature parameter of bone-conduction speech of frame $i$. $d()$ is distortion calculation function described as follows:

$$d(x_{a}^i, \lambda_{s_a}^i) = \min_m \left( |x_{a}^i - C_{s_a}^m|^2 \right),$$  \hspace{1cm} (2)

$$d(x_{b}^i, \lambda_{s_b}^i) = \min_m \left( |x_{b}^i - C_{s_b}^m|^2 \right),$$  \hspace{1cm} (3)

where $C_{s_a}^m$ and $C_{s_b}^m$ are VQ centroid in the air-conduction speaker model, $\lambda_{s_a}^i$ and VQ centroid in the bone-conduction speaker model, $\lambda_{s_b}^i$. The proposed method calculates the VQ distortion between speaker model and the feature parameter of input speech as weighted summation of the VQ distortion of air-conduction speech and bone-conduction speech. Because of using the information of air- and bone-conduction speech, we believe that the proposed method can improve the speaker identification performance of the individual speech.
4. Speaker Identification Experiment

Using part of a Japanese speech corpus described in section 2, we conducted speaker identification experiments. In this experiment, we used 99 female’s speech data which were down-sampled from 44.1 kHz to 8 kHz. Hence, in this paper, we conducted the 99 female speaker identification experiment.

4.1. Experimental Conditions

4.1.1. Registered data

For the registered data, i.e., the speaker model training data, we used five text sentences, which are ATR Japanese balance sentences, per speaker. We used speech data that was collected first time in the first session for these registered data. The average length of the registered data which included the silent section is about five seconds.

4.1.2. Testing data

35 utterances per speaker were used. For testing, we used 3,465 utterances and these testing utterance were divided to following four test sets:

1. TD-SC
   The text of the utterances in this test set is the same as the registered data, i.e., Text-Dependent speaker identification (TD). The recording session of this test set is the same as registered data, i.e., this test set was collected on first session (SC: Session Closed). The number of utterances per speaker in this test set is 5. Hence, the total number of testing utterances is 495 (99 speakers × 5 sentences).

2. TI-SC
   The text of the utterances in this test set is not contained in the registered data, i.e., Text-Independent speaker identification (TI). The recording session of this test set is the same as registered data. The number of utterances of each speaker in this test set is 10. Hence, the total number of testing utterances is 990 (99 speakers × 5 sentences).

3. TD-SO
   The text of the utterances in this test set is the same as the registered data, i.e., Text-Dependent speaker identification. The recording session of this test set is different from that of registered data (SO: Session Opened). The number of utterances of each speaker in this test set is 10. Hence, the total number of testing utterances is 990.

4. TI-SO
   The text of the utterances in this test set is not contained in the registered data, i.e., Text-Independent speaker identification. The recording session of this test set is different from that of registered data. The number of utterances of each speaker in this test set is 10. Hence, the total number of testing utterances is 990.

4.1.3. Feature vector and acoustic model

All data, sampled at 8kHz, were segmented into overlapping frames of 25ms, producing a frame every 10ms. A Hamming window was applied to each frame. Mel-filtering was performed to extract 12-dimensional static MFCC, as well as a logarithmic energy (log-energy) measure. The 12-dimensional delta MFCC and delta log-energy were extracted from the static MFCC and the log-energy, respectively. After that, by omitting the log-energy, we constituted a 25-dimensional feature vector (12 static MFCCs + 12 delta MFCC + delta log-energy). Cepstral Mean Subtraction (CMS) was applied on the static MFCC vectors. In this experiment, we used only the speech section which are detected by voice active detection using power information. We used HTK version 3.4[9] for the feature extraction.

For speaker model, we used the VQ centroids. The number of centroids is set to 64.

For comparing to the proposed method, we conducted the speaker identification experiments by combining the feature parameter of air-conduction speech to the feature parameter of bone-conduction speech. We set the weighted value, \( \alpha_{t} \), to 0.7 in all frames.

4.2. Experimental results

Table 1 shows the identification error rate (IER). In this table, AIR and BONE mean the results of air-conduction speech and bone-conduction speech, respectively. FEAT and DIST mean the results of feature parameter combination method and distance combination method, which is the proposed method, respectively.

Comparing AIR to BONE, we can see from this table that the IERs of AIR are lower than those of BONE under all test sets. Especially, the performances of AIR are better than those of BONE under the condition that the text of registered data is the same as that of the identification data, which are TD-SC and TD-SO. In this experiment, we used the conventional methods, which are feature extraction, speaker models, and so on, for the speaker recognition. Hence, these results imply that the novel methods are needed for the speaker recognition using bone-conduction speech.

We can see from this table that both combination methods, which are FEAT and DIST, can reduce the IERs of individual speech, which are AIR and BONE. Therefore, we conclude that the combination of the information of two speech is available for speaker identification. In addition, we can see that the IERs of DIST is lower than those of FEAT. These results show that it is possible to improve the speaker identification performance by combining these speech on the distance calculation. In addition, we investigate the details of these results in the following section.

4.3. Discussion

4.3.1. Discussion of weight value

In the proposed method, we calculated the VQ distortion of each speaker by weighted summation of the VQ distortion of air- and bone-conduction speech in equation (1). Hence, the weight value influences to the identification performances. In this section, we investigate the relationships between the weight value and the IER. Figure 3 shows the identification error rate as a...
function of the weight value, $\alpha_i$.

From these results, we can see the best identification performance occur under the condition that the weight value, $\alpha_i$, sets to 0.7. This implies that the weight of the distance of AIR is higher than that of BONE because the IER of AIR is lower than the IER of BONE. In the future, we investigate the detail of these results and propose the weight decision function.

4.3.2. Investigation of the number of correct/incorrect utterances

For analyzing the detail of the identification results, we investigated the number of correct/incorrect utterances. Table 2 shows the number of utterances for each result. In this table, ‘C’ and ‘I’ mean correct utterance and incorrect utterance, respectively.

From this table, we can see that the proposed method can identify the utterances to be correct speaker even if it is not possible to correctly identify the utterances on both speech (AIR: I, BONE: I, DIST: C). The number of these utterances is 77. Hence, we investigate the ranking of the correct speaker and the ranking of the incorrect speaker of these utterances.

This investigation shows that the average ranking of the correct speaker of air-conduction speech and the average ranking of the correct speaker of bone-conduction speech are 2.3 and 7.9, respectively. From this investigation, we can confirm that the ranking of the correct speaker is high even if the utterance is not identified correctly. On the other hand, the average ranking of the speaker with minimum distance on air-conduction speech under bone-conduction speech is 44.8 and the average ranking of the speaker with minimum distance on bone-conduction speech under air-conduction speech is 33.1. This investigation implies that the ranking of one speech may not be high even if the distance of other speech is small. Therefore, we believe that the proposed method which uses the information of both speeches can improve the speaker identification performance of each speech.

5. Summary

In this paper, we proposed the novel speaker identification method which combines “bone-conduction speech” and “air-conduction speech”. The proposed method conducts speaker identification by integrating the similarity calculated by air-conduction speech model and similarity calculated by bone-conduction speech model. For evaluating the proposed method, we conducted the speaker identification experiment using female speech data.

Experimental results showed that the identification error rates of the proposed method were lower than those of air-conduction speech and those of bone-conduction speech under all conditions. From this experiment, we concluded that the combination of air-conduction speech and bone-conduction speech was useful for the speaker identification.

In future, we will investigate the detail of these results and the speaker model. In addition, we will investigate the feature of bone-conduction speech and study a new method for the speaker recognition using the bone-conductive microphone.

6. Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 19700172, Scientific Research (B), 17300065, and the Okawa Foundation for Information and Telecommunications.

7. References


Table 2: Details of identification results (Num. of utterances)

<table>
<thead>
<tr>
<th>AIR</th>
<th>BONE</th>
<th>DIST</th>
<th>TD-SC</th>
<th>TD-SI</th>
<th>TD-SO</th>
<th>TI-SC</th>
<th>TI-SO</th>
<th>test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>457</td>
<td>913</td>
<td>425</td>
<td>315</td>
<td>315</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>I</td>
<td>C</td>
<td>33</td>
<td>47</td>
<td>403</td>
<td>320</td>
<td>320</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>I</td>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>I</td>
<td>C</td>
<td>C</td>
<td>4</td>
<td>17</td>
<td>52</td>
<td>73</td>
<td>73</td>
<td>C</td>
</tr>
<tr>
<td>I</td>
<td>C</td>
<td>I</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>31</td>
<td>31</td>
<td>C</td>
</tr>
<tr>
<td>I</td>
<td>I</td>
<td>C</td>
<td>0</td>
<td>5</td>
<td>27</td>
<td>45</td>
<td>45</td>
<td>C</td>
</tr>
<tr>
<td>I</td>
<td>I</td>
<td>I</td>
<td>1</td>
<td>3</td>
<td>55</td>
<td>184</td>
<td>184</td>
<td>C</td>
</tr>
</tbody>
</table>

Figure 3: IERs as a function of weight value, $\alpha_i$. 