ACOUSTIC ANALYSIS FOR SPEAKER IDENTIFICATION OF WHISPERED SPEECH

Xing Fan and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA

{xxf064000, john.hansen}@utdallas.edu http://crss.utdallas.edu

ABSTRACT

Whisper is an alternative speech production mode from neutral speech, which is used by talkers intentionally in natural conversational scenarios to protect personal privacy and avoid being overheard. Due to differences between whispered and neutral speech in vocal excitation and vocal tract function, the performance of speaker ID systems trained with neutral speech degrades significantly. In this study, a neutral trained closed-set speaker ID task based on MFCC-GMM is considered. It is observed that for whisper speaker recognition, the degradation is concentrated for a certain number of speakers. Next, an acoustic analysis is conducted in order to determine the reason affecting the degradation for those speakers. Finally, a confidence space is proposed to measure the quality of whispered speech for the task of speaker ID. Experimental evaluations demonstrate the effectiveness of this method in searching whispered utterances with poor speaker information for a neutral/whisper mismatch speaker ID system. The proposed method makes it possible to compensate for those poor utterances, meanwhile avoiding any harm to other utterances that remain the performance of neutral speaker ID task.

Index Terms— Whispered speech, speaker information, speaker identification, MFCC

1. INTRODUCTION

Whisper is a natural alternative speech production mode. It is commonly used in public circumstances to avoid being overheard or to protect personal information. For example, a speaker may prefer using whisper when providing their date of birth, credit card number and billing address when making a hotel/flight/car reservation over the phone. Aphonic individuals, as well as those with low vocal capability, including heavy smokers, also employ whisper as a primary form of oral communication. There is a significant difference between whisper and neutral speech in the spectral domain, caused by a loss in voiced excitation structure and shifting of formant locations with lower frequency [1,2,3,4,5]. Considering the fact that speaker dependent whisper adapting data is generally not available in real scenarios, these differences present a major challenge for effective speaker ID system performance trained with neutral speech.

Several studies have recently considered compensation of whisper for a whisper/neutral mismatch speaker ID system [6,7,8] and it was found in [9] that features based on short time analysis fail to capture most of the speaker ID information for whisper with low SNR. However, no study has considered a detailed study concerning the varieties of difference between whispered and neutral speech among different speakers and its effect on neutral trained MFCC-GMM systems. In this study, a neutral trained closed-set speaker ID system based on MFCC-GMM is first formulated. It is observed in whisper speaker recognition that the degradation is concentrated for a certain number of speakers, while for other speakers, their whisper achieves comparable performance with the neutral testing task. It would therefore be useful to determine for which speaker will whisper speech severely impact speaker ID. Next, acoustic analysis is conducted to compare the misrecognized and correctly recognized whisper utterances based on the log energy output from linear scale filterbanks. It will be observed that SNR is not the only reason for the degradation for whispered speech. The spectral slope and distribution of spectral energy along the frequency axis also affects the quality of the whispered speech. Based on the analysis above, we propose a two dimensional space for measuring the quality of whispered speech for a neutral trained MFCC-GMM speaker ID system. Experiments will demonstrate the effectiveness of this method in searching for those whisper speakers who perform poorly for the MFCC-GMM system.

The remainder of this paper is organized as follows. In Sec.2, a general introduction to the database, MFCC-GMM speaker ID system is presented. Next, recognition results are presented and acoustic analysis is provided based on two
2. DATABASE AND SYSTEM

2.1. UT-Whisper Corpus Setup

The UT-Whisper corpus developed in [10] is employed here. The corpus consists of a total of 102 speakers, 37 male and 75 female. Whispered and neutral speech from a sub-set of 28 native American English female subjects was used. In the first part, each subject reads 40 sentences alternatively in neutral and whisper mode. Those sentences come from the TIMIT database and are phonetically balanced. In the second part, two paragraphs cited from a local newspaper was produced by each subject. For each paragraph, four whisper-islands were produced with each island containing 1-2 sentences. In the third part, the same paragraphs in the second part were used. Instead of sentences, five phrases were read in whispered mode, with each phrase 2-3 words in duration. The whispered and neutral speech for all three parts are manually separated to constitute whisper and neutral corpora. From [10], we also note that all recordings include pure-tone calibration test sequences to provide ground-truth on true vocal effort for all speakers and sections. Speech data was digitized using a sample frequency of 16 kHz, with 16 bits per sample. Speech from all speakers was windowed with a Hamming window of 32 ms, with an overlap of 16 ms for the following parts.

2.2. MFCC-GMM Speaker ID system

First, all silence parts in the whisper and neutral speech is removed by an energy threshold that depends on the SNR of each particular sentence. GMMs for the total 28 speakers are trained using 19-dimensional static MFCC vectors only from the neutral corpus. Close set speaker recognition is conducted using whispered data as the test utterances. For the total 961 test utterances, an average accuracy of 79.29% is obtained. A confusion matrix of the system is provided in Fig1 to show the distribution of accuracy. The darker shade denotes a higher probability. The numbers along the x and y axis is the speaker index. We can see that the accuracy is not evenly distributed among the 28 speakers. For example, for speaker 4 and 12, an accuracy of 100.00% is achieved for whispered speech. However, for speaker 18 and 25, the average accuracy is below 50%. There are also a number of speakers that have an accuracy between 70.0% to 80.0%. The recognition result suggests that the similarity of speaker information in neutral speech and in whispered speech depends on the way the speaker produces whisper: some speakers continue producing good quality whisper, some provide low quality speaker ID content whisper, while the rest are distributed from good to poor depending on the context.

Fig1. Confusion matrix of the recognition result

For the whisper/neutral mismatch speaker ID system, compensation is usually applied to the whispered speech to enhance performance[7,8,9]. However, compensation is not necessary for those high quality speaker where whisper does not impact performance. If a method can be found to separate the poor quality utterances from those good quality ones, the performance based on MFCC-GMM for those good quality whisper will remain and, any damage the compensation strategy may introduced them will be avoided. For those poor quality utterances, a compensation strategy that is more focused can be employed.

2.3. Acoustic Analysis

In [9], it was determined that whisper with low SNR is difficult for correct speaker ID with an MFCC-GMM system based on a 10 subjects database. In this study, we conduct an analysis of the SNR of correctly recognized and misrecognized utterances based on the MFCC-GMM system with a 28 female subjects database. The result is shown in Fig2. The utterances are classified based on their SNR. The x axis represents the percentage of utterances falling into the SNR range on the y-axis. As we can see, the percentage of correctly recognized speaker utterance, with high SNR is higher than that of misrecognized utterances, which confirms that whisper with high SNR retains more speaker dependent information compared with neutral speech. However, a portion of the utterances with low SNR can also be correctly recognized. Therefore, SNR cannot be used as the only property to indicate the quality of the whisper.

In order to find out other acoustic properties that cause loss of neutral speaker information in whispered speech, two UBM's are trained using the log energy output from a 26 linear scale filterbanks for misrecognized and correctly recognized utterances. It should be noted that unvoiced consonant of whisper maintain the same acoustic features as neutral speech[1]. Hence, unvoiced consonant parts are removed from all whispered speech before feature extraction. An ef-
icient way to remove fricatives is to calculate the ratio of high frequency to low frequency energy and the entropy of the frequency domain. Next, a threshold is found for decision making.

Fig 2. Analysis of SNR for test utterances

The fused mean of two UBM s is plotted in Fig 3. We can see in Fig 3 that for those utterances that have been correctly recognized, there is more spectral energy in the high frequency domain. Since whispered and neutral speech shares more similarity in higher frequency domain[5], more information in this area increases the chance of correct speaker recognition. Also, speaker utterances being recognized correctly have less percentage of spectral energy from 1000-2000 Hz. It is known that more speaker information is contained in F3, F4[11], and formant shifts of whispered speech happen mostly for F1 and F2. Hence, information in this area contains less speaker information that is shared between whispered and neutral speech, which explains why those misrecognized utterances usually have a higher percentage of energy in this area than those that have been correctly recognized.

Fig 3. Comparison of UBMs for two kinds of whisper

One way to measure the percentage of information in high frequency is the spectral slope. In Table 1, we list the distribution of the spectral slope of misrecognized utterances and correctly recognized utterances. In order to measure the distribution of information along with frequency, we simply calculate the ratio of the energy in 1000-2000 Hz to the energy in 1000-8000 Hz, which will be referred to as ratio₁ for simplicity. The result of ratio₁ is listed in Table 2. We can observe that a higher percentage of misrecognized whispered speech has lower spectral slope. For ratio₁, it also confirms our observation of means of UBMs. For example, the ratio₁ of 45.5% of misrecognized whispered speech is between 0.4 – 0.5, while only 19.3% of correctly recognized whisper is in this area.

**Table 1. Comparison of spectral slope.**

<table>
<thead>
<tr>
<th>Spectral Slope</th>
<th>Whisper</th>
<th>Misrecognized</th>
<th>Recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; −2</td>
<td>75.6%</td>
<td>52.6%</td>
<td></td>
</tr>
<tr>
<td>−2 − −1.5</td>
<td>23.58%</td>
<td>40.6%</td>
<td></td>
</tr>
<tr>
<td>−1.5 − −1.0</td>
<td>0.0%</td>
<td>5.0%</td>
<td></td>
</tr>
<tr>
<td>&gt; −1.0</td>
<td>0.8%</td>
<td>1.8%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Comparison of ratio₁.**

<table>
<thead>
<tr>
<th>ratio₁</th>
<th>Whisper</th>
<th>Misrecognized</th>
<th>Recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.3</td>
<td>0.8%</td>
<td>28.6%</td>
<td></td>
</tr>
<tr>
<td>0.3 − 0.4</td>
<td>41.5%</td>
<td>47.6%</td>
<td></td>
</tr>
<tr>
<td>0.4 − 0.5</td>
<td>45.5%</td>
<td>19.3%</td>
<td></td>
</tr>
<tr>
<td>&gt; 0.5</td>
<td>12.2%</td>
<td>4.5%</td>
<td></td>
</tr>
</tbody>
</table>

3. CONFIDENCE SPACE AND EVALUATION

Based on the above analysis, SNR, ratio₁, and spectral slope are direct indicators of the quality of whispered speech that impact speaker ID. In order to suppress the potential side effect of compensation made to good quality whisper as much as possible, a criterion needs to be proposed that provides confidence for speaker recognition based on MFCC-GMM system for each test whispered speech. For simplicity, we use Eq. (1) to calculate $SR$ in order to combine the spectral slope and ratio₁ together,

$$SR = \frac{1}{ratio₁ (1 - e^{-0.2\text{SNR}})}.$$  

Next, the $SR$ and the SNR can be employed to constitute a two-dimensional-space to indicate the quality of the whispered speech. For simplicity, we call this space the confidence space. Fig 4 shows the distribution of misrecognized and correctly recognized whispered speech in this confidence space. We can see that the misrecognized utterances converged toward to the left down corner of the confidence space, which is indicated with a square in Fig 4. While for those that have been recognized correctly, most have relative higher SNR,
and even those with low SNR, their values of SR tend to have high values. More sophisticated methods, such as SVM or neural networks could also be applied to search for a data fusion boundary.

By moving the threshold along the SNR and SR axis, we can obtain a set of error rates as shown in Fig5. We can see that there is a tradeoff between detecting good and poor quality of whisper speech for speaker ID. For example, when 30% of misrecognized whisper is not detected by setting a hard decision threshold, about 60% of utterances that can be correctly recognized by the MFCC-GMM speaker ID system are detected as good quality. While when 90% of whisper that does not achieve effective speaker recognition with MFCC-GMM system is detected, about 30% of utterances that can be correctly recognized is classified as good quality. However, as we can observe in Fig4, even though some misrecognized utterances sit outside of the hard threshold, they are near the boundary, thus the position in the confidence space can still provide valuable information about the quality of the whisper, and further determines the degree of compensation needed. On the other hand, although some part of the correctly recognized speech for speaker ID can be identified clearly as good quality in the confidence space, other parts of correctly recognized speech fall into the boundary of misrecognized whisper. Since they have similar acoustic properties with those misrecognized whisper, further compensation may still be reasonable.

4. CONCLUSION

Whisper is an alternative speech production mode that is commonly employed for communication in public circumstances to protect personal privacy. However, the performance of traditional MFCC-GMM speaker ID systems degrades due to significant differences between whispered and neutral speech production. Compensation strategies have been proposed in the past [7,8,9]. However, no study has considered analyzing which kind of whispered actually needs compensation versus no compensation.

In this study, a MFCC-GMM systems was established and based on the analysis of the speaker recognition, it is found that the performance varies among different speakers. The differences in recognition was caused by varieties in whisper production among speakers. After that, Acoustic properties of misrecognized and correctly recognized whispered were addressed by calculating the SNR and two UBMMs based on a linear scale filterbanks. Comparison was also made in spectral slope and ratio. A two dimensional confidence space was proposed to measure the quality of whispered speech in preserving speaker information in neutral speech. Evaluation results showed that although part of the correctly recognized whispered can be detected as poor quality, a major part of them is indicated as good quality with high confidence in the confidence space. Most of the misrecognized whisper based on MFCC-GMM can be successfully detected and therefore either set aside or processed with whisper compensation[7,8,9]. Future work will be focused on looking at other acoustic properties and classifying methods.

5. REFERENCES