Speaker Identification for Whispered Speech based on Frequency Warping and Score Competition

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

In certain situations, talkers will intentionally use whisper instead of neutral speech for the sake of privacy or confidentiality, which severely degrades the performance of speaker identification systems trained with only neutral speech. There are considerable differences in the spectral structure between whisper and neutral speech due to an absence of voice harmonic excitation. This study introduces a new feature based on frequency warping and score competition for the task of speaker identification for whisper. The proposed feature method is evaluated on a corpus of male speakers in both neutral and whisper. Closed set speaker ID results show an absolute 27% improvement in accuracy when compared with a traditional MFCC feature based system. The result confirms a viable approach to improving speaker ID performance between neutral and whisper speech condition.

Index Terms: speaker identification, whisper, frequency warping, score competition

1. Introduction

Whisper is a natural but special mode of speech production which presents special challenges to speech systems. It is commonly used for communication in public circumstances to ensure privacy, as well as for auralic individuals [1], or those with low vocal capability (i.e. heavy smokers, etc.). Although achievements have been made in the field of speaker identification in recent years, few studies have considered the task of speaker ID for whispered speech. In [2], a study considered automatically identifying whisperers, based on the assumption that at least 5 seconds of whispered speech per speaker is available. The resulting system reached a 8-33% relative improvement with 5 to 15 seconds of whispered speech per speaker. However, in real scenarios, speaker dependent whispered training data is generally not available. In this paper, we investigate a method for identification of speakers who whispered for the sake of privacy or confidentiality using GMMs trained only with each speaker’s neutral speech.

There is a significant difference between whisper and neutral speech in the spectral domain, due to a loss of voice excitation structure. In [4] and [5], it was shown that whisper and neutral speech share more similarity in the higher frequency domain and there is some formant shifting in lower frequency domain. In this sense, frequency warping has been employed in the design of filter banks. The first warping stage takes the advantage of their similarity in higher frequencies, while, the second warping stage, which is combined with GMM (Gaussian mixture model) score competition, attempts to compensate the formant shift. The integration of these two processing stages offers an important step forward in addressing whisper/neutral speech for seamless speaker recognition.

The remainder of this paper is organized as follows. In Sec. 2, we develop the Frequency Warping Stage I for speaker ID of whisper. In Sec. 3, we provide detail information for Frequency warping Stage II, which is combined with GMM score competition. In Sec. 4, a general overview of the corpus and experimental setup is presented, then performance of speaker ID based on the proposed method is compared with an MFCC baseline system. Conclusions and summary are drawn in Sec. 5.

2. Frequency Warping Stage I (FWS-I)

In [3], it was shown that there is an upward shift of formant frequencies of vowels in whispered speech data compared to neutral speech, and the shift amount is larger for low formant frequencies. It is also shown in [5] that there is almost no shift in F3 and F4 when comparing whispered and neutral speech. So if more higher frequency information is actually contained in the feature vector, whispered and neutral speech share more traits. A GMM trained only with neutral speech would therefore achieve better performance when tested with whispered speech. The MFCC (Mel frequency cepstral coefficient) feature is obtained from the mel-frequency scale which reflects the frequency resolution of the human ear. However, the mel-scale emphasizes lower frequency and de-emphasizes higher frequency by putting more filters in lower frequency domain [7]. In this case, a new frequency scale which emphasizes the higher frequency information is needed as a substitute to the mel-scale. In [8], new frequency scales for recognition of speech under stress were proposed, which included an exponential mapping function to emphasize mid-frequencies. The general form of the exponential mapping function used in our study is:

\[ y = c \times \left(10^{f/k} - 1 \right), 0 \leq f \leq 8000Hz. \]  

The values of \( c \) and \( k \) are obtained by solving a set of equations specifically for whisper and neutral speech. First, we require that the exponential and mel-scale warping functions are equal at the half sample frequency. Sample frequency is 16k Hz in our study. Thus, we obtain Eq. (2) as follows,
\[ c \times (10^{8000/k} - 1) = 2595 \times \log(1 + \frac{8000}{700}). \] (2)

Next, in order to reduce the emphasis of information in the lower frequency domain, which is defined as 0 to 4kHz here, the specific values of c and k that minimize the value of y when f=4000 Hz are obtained through Eq. (3) by minimizing the absolute value of the partial derivative of Eq. (1) with respect to c and k.

\[ \{c, k\} = \min \left\{ \left| 10^{4000/k} - 1 - \frac{4000}{k^2} \times \ln 10 \times c \times 10^{-\frac{4000}{k}} \right| \right\}. \] (3)

By solving Eqs. (2) and (3), we obtain the values of c and k with c=6375 and k=50000. Thus, the exponential warping function used here is as follows,

\[ y = 6375 \times (10^{f/50000} - 1). \] (4)

Fig. 1 shows two mapping functions: the mel-scale and exponential frequency-scale, respectively. For convenience of comparison, we also put the linear frequency-scale together with the exponential one. To compute the new feature vector for whispered speaker ID, a set of 24 triangular bandpass filters on the LP power spectrum was placed according to the exponential scale function in Eq. (4). The cosine transform is then applied to the log energies obtained from the filter banks to obtain a set of cepstral coefficients. Only the first 19 cepstral coefficients are employed for train and test. The LP power spectrum was chosen here instead of the FFT because the LP power spectrum contains more vocal tract information while removing more excitation information [7]. Since one of the main differences between neutral and whispered speech is their excitation [11], this strategy helps to minimize this difference.

\[ f' = \begin{cases} \frac{f}{7} & \text{for } 0 \leq f \leq 1200 \text{ Hz} \\ \frac{f}{4} & \text{for } 1200 < f \leq 8000 \text{ Hz} \end{cases}. \] (5)

Because the value of \( \alpha \) varies according to different speakers, hence instead of using one fixed value of \( \alpha \), we compensate by drawing \( \alpha \) from the following values: 1.2, 1.25, 1.3, 1.35, 1.4, therefore there are 5 sets of feature vectors for each test frame of whisper speech. For a given speaker \( k \), feature vectors obtained from neutral speech through only exponential function warping (FWS-I) were used to train a GMM model \( \Theta^k \). The likelihood of a whispered observation is found using one feature vector \( x_n \) given this GMM model \( \Theta^k \) of speaker \( k \) as \( p(x_n \mid \Theta^k) \). Since we have 5 different warping factors for each whisper test trial, and for each warping factor \( \alpha \) the likelihood can be expressed for \( p(x_n \mid \Theta^k) \), in which \( 1 \leq m \leq 5 \). Thus, we have 5 corresponding test scores respectively for each trial. Considering there are \( K \) speakers, so there will be \( K \) GMM scores. Together for each trial, we have \( K \times 5 \) scores, among which we select the GMM that achieves the highest score in this total \( K \times 5 \) scores. The likelihood of the entire set of feature vectors is therefore estimated as

\[
p(X \mid \Theta^k, m) = \max_{m=1,2,3,4,5} \left\{ \max_{k=1,2,\ldots,K} \left\{ \prod_{n=1}^{N} p(x_n \mid \Theta^k) \right\} \right\}.
\]

(6)

4. Experimental Results

4.1. Corpus and Experimental Setup

The corpus developed in [10] is employed in this study. A sample of neutral and whispered speech are collected from a total of 10 native English-spoken male subjects, with each subject reading 10 sentences from the TIMIT database in two speaker modes: neutral and whisper. Speech data was digitized using a sampling frequency of 16 kHz, with 16 bits per sample. From [10], we note that all recordings include pure-tone calibration test sequences to provide ground truth on true vocal effort for all speakers and sections. Finally, training data per speaker model is approximately 30 sec and each whisper/neutral test is 3-5 sec.

Speech from all speakers was windowed with a Hamming window of 20ms, with a 10ms overlap rate. Because the spectral slope of whispered speech is more flat compared with that of neutral speech [6], so instead of using a traditional pre-emphasis filter given by

\[ H(z) = 1 - 0.97z^{-1}. \] (7)

3. Frequency Warping Stage II (FWS-II) and Score Competition

In [5], it was observed that the value of F1 for whispered speech vowels is about 1.3-1.6 times the value to neutral speech. For male speakers: the range of frequencies for F1 is approximately 150-850 Hz. In order to compensate for this difference between whispered and neutral speech, we move the position of the first three filter banks in the exponential-scale frequency domain upward to, say, \( \alpha \) times the value of the original during feature extraction for whispered speech. The remaining filter banks are kept at the same position for neutral speech. In this case, it is equivalent to a linear warp of the frequency components under 1200 Hz of whispered speech by the factor \( \alpha \). The new frequency \( f' \) in the linear frequency domain after this warping is:

\[ f' = \begin{cases} \frac{f}{7} & \text{for } 0 \leq f \leq 1200 \text{ Hz} \\ \frac{f}{4} & \text{for } 1200 < f \leq 8000 \text{ Hz} \end{cases}. \] (5)

Because the value of \( \alpha \) varies according to different speakers, hence instead of using one fixed value of \( \alpha \), we compensate by drawing \( \alpha \) from the following values: 1.2, 1.25, 1.3, 1.35, 1.4, therefore there are 5 sets of feature vectors for each test frame of whisper speech. For a given speaker \( k \), feature vectors obtained from neutral speech through only exponential function warping (FWS-I) were used to train a GMM model \( \Theta^k \). The likelihood of a whispered observation is found using one feature vector \( x_n \) given this GMM model \( \Theta^k \) of speaker \( k \) as \( p(x_n \mid \Theta^k) \). Since we have 5 different warping factors for each whisper test trial, and for each warping factor \( \alpha \) the likelihood can be expressed for \( p(x_n \mid \Theta^k) \), in which \( 1 \leq m \leq 5 \). Thus, we have 5 corresponding test scores respectively for each trial. Considering there are \( K \) speakers, so there will be \( K \) GMM scores. Together for each trial, we have \( K \times 5 \) scores, among which we select the GMM that achieves the highest score in this total \( K \times 5 \) scores. The likelihood of the entire set of feature vectors is therefore estimated as

\[
p(X \mid \Theta^k, m) = \max_{m=1,2,3,4,5} \left\{ \max_{k=1,2,\ldots,K} \left\{ \prod_{n=1}^{N} p(x_n \mid \Theta^k) \right\} \right\}.
\]
we use an adaptive pre-emphasizer for both whispered and neutral speech, which is proposed in [9] as
\[ H(z) = 1 - \tilde{a}_n z^{-1}. \]  
where \( \tilde{a}_n = r_n(1)/r_n(0) \). The variable filter coefficients are represented as the ratio of the first to the zeroth order lag autocorrelation parameters.

Also, as noticed in [3], unvoiced consonants are generally similar for neutral and whisper speech. In this case, the frequency warping Stage II should not be applied to whisper speech, where the corresponding neutral speech is unvoiced speech. In most cases, this is generally unvoiced fricatives, such as the phonemes [f], [s], [b], etc. By spectral energy analysis, we find that unvoiced fricatives are distributed across the frequency domain, while for other phonemes, because of the effect of formants movement and lack of periodic excitation, the spectral energy varies significantly across the frequency range with a higher concentration in relatively lower frequency domain. In order to detect this part of whisper, we simply compute the ratio of the spectral energy above and below 4000 Hz and then set a threshold. We assume those frames that have a ratio above the threshold are unvoiced fricative frames and so no linear frequency warping (FWS-II) is applied to those frames.

For neutral speech, after applying front-end processing based on exponential frequency warping, a 19-dimensional feature vector is extracted from all speech data. Using these training vectors, a GMM with 32 Gaussian mixtures is trained for each speaker. For testing, the windowed whispered speech was first processed for unvoiced fricative detection, then for the part whose corresponding neutral speech is unvoiced fricative, the feature extraction method is the same as that for neutral, while the remaining speech data for whisper are processed through the filter banks based on both Frequency Warping Stage I and II. For both cases, we obtain 19-dimensional feature vectors for testing.

To understand the impact of the proposed FWS-I, FWS-II processing, a baseline system was built based on the 19-dimensional LP-MFCC, in which no frequency warping was introduced for comparison. Next, FWS-I was applied to both neutral and whispered speech front-end processing to demonstrate the effectiveness of our proposed algorithm. Next, linear frequency warping below 1200 Hz (FWS-II) is introduced in our system. In order to demonstrate the effect of unvoiced fricative detection, we perform recognition experiments with and without this detection and compare the results. Fig. 2 depicts the overall flow diagram for the final improved system based on frequency warping and score competition.

4.2. Experiment Results

To understand the impact of the proposed feature on speaker ID system when the test data is neutral speech instead of whisper, speaker ID experiments were first performed with neutral test data. When a total of 15 sec neutral training data available for each speaker, a closed set speaker recognition rate of 92% is obtained. Compared with the system based on traditional MFCC features [12], there is a slight degradation. However, this problem can be solved with whisper detection pre-processing, which can separate whisper from neutral speech. After the whisper detection procedure, different feature extraction methods can be applied based on different speech modes of the test data. Next, we concentrate on whispered speech.

Results for closed set speaker ID experiments are shown in Fig. 3, where all test data is open whisper speech. As we observe in Fig. 3, GMMs from a ten speaker set are trained using 19-dimensional static MFCC vectors and tested with whispered speech using the same feature extraction method. We obtain a closed set speaker recognition rate of 53%. This significant loss in speaker ID performance shows how whisper speech causes speaker ID system to fail. Next, FWS-I exponential-frequency warping is applied for GMM training and testing with 19-dimensional feature vectors extracted based on the method in Sec 2. The resulting accuracy with FWS-I feature process-
ing engaged is 64%, which represents an 11% improvement over our baseline. If both Frequency Warping Stage I and II are applied without unvoiced-fricative detection, the speaker ID system’s performance increases to 71% with an 18% increase compared to the baseline. Finally, when the unvoiced-fricative detection processing task is applied together with exponential frequency warping and linear frequency warping in lower frequency domain (FWS-I, FWS-II), the system achieved an overall performance of 80%, which represents a 27% improvement over the original baseline system performance.

5. Conclusion

In many natural conversational scenarios, subjects will employ whisper speech instead of neutral speech in order to avoid being heard or identified. In speech communication or voice dialog systems, personal information is sometimes required and speakers would prefer not to broadcast credit card/personal numbers/names, etc. for others to hear. Speech produced in whisper changes significantly in spectral structure since there is no periodic/harmonic excitation for voiced phonemes. Therefore, the performance of GMM based speaker ID systems degrades because high energy voiced phonemes are the mainstay for GMM training. It has been shown that speaker ID systems degrade significantly when vocal effort is mis-match between train and test [10], which is particularly the case for whisper-neutral mismatch.

In this study, a new speech feature processing approach has been proposed based on a two stage frequency warping strategy in combination with GMM score competition. Exponential-Frequency warping is based on the fact that whisper and neutral speech share more common high frequency structure when overall gain adjustment is achieved. In the proposed scheme, linear frequency warping is also applied primarily to correct for formant shifts, which occur in whisper speech at lower frequencies. Since the spectral structure for unvoiced neutral and whisper speech are similar (i.e., no difference due to excitation structure), an unvoiced-fricative detection scheme was developed to determine if a particular whisper speech frame requires linear frequency warping at low frequencies or not. Using a previously collected corpus, where pure-tone calibration test signals provide ground truth on true vocal effort, a closed-set speaker recognition evaluation was performed. Employing the two new frequency warping strategies (FWS-I+FWS-II), along with unvoiced-fricative detection and GMM score competition, we achieved an overall 80% recognition accuracy using only neutral speech data for training. When compared with an MFCC baseline system (53%), there is a 27% absolute improvement in performance.

While this is encouraging, it is clear that a much larger and more comprehensive corpus collection is needed to demonstrate the effectiveness of the proposed algorithms in actual voice communications/voice dialog systems. The results, though, do represent one of the first advancements in addressing whisper for speaker ID, and confirms the viability of the overall algorithm to improve performance on whisper-neutral speech conditions.

6. References

[4] X. Li, “Reconstruction of Speech from Chinese Whispers,” (PhD), Nanjing University, China, 2004