Speaker Identification for Whispered Speech Using Modified Temporal Patterns and MFCCs

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

Abstract
Speech production variability due to whisper represents a major challenge for effective speech systems. Whisper is used by talkers intentionally in certain circumstances to protect personal privacy. Due to the absence of periodic excitation in the production of whisper, there are considerable differences between neutral and whispered speech in the spectral structure. Therefore, performance of speaker ID systems trained with high energy voiced phonemes, degrades significantly when tested with whisper. This study considers a combination of modified temporal patterns (m-TRAPs) and MFCCs to improve the performance of a neutral trained system for whispered speech. The m-TRAPs are introduced based on an explanation for the whisper/neutral mismatch degradation of an MFCC based system. A phoneme-by-phoneme score weighting method is used to fuse the score from each subband. Text independent closed set speaker ID was conducted and experimental results show that m-TRAPs are especially efficient for whisper with low SNR. When combining scores from both MFCC and TRAPs based GMMs, an absolute 26.3% improvement in accuracy is obtained compared with a traditional MFCC baseline system. This result confirms a viable approach to improving speaker ID performance between neutral/whisper mismatch conditions.

Index Terms: whispered speech, speaker identification, temporal patterns, MFCCs

1. Introduction
Whisper is a commonly used alternative vocal speech style from neutral speech, employed by talkers for communication in public circumstances to protect privacy. For example, a speaker may prefer employing whispered speech when giving their credit card information, billing address or date of birth when either speaking to a bank verification system, or making a hotel/flights/car reservation over a cell phone. Aphonic individual, as well as those with low vocal capability (i.e. heavy smokers, etc.) also use whisper as the primary form of oral communication. Compared with neutral speech, whisper has no fundamental frequency due to the absence of voice harmonic excitation. The frequency of lower formant is higher than that of the neutral speech, and the spectral slope is more flat in whisper[1, 6, 9, 10, 11]. These differences present unique challenges for effective speaker ID system performance when mixed vocal style is present.

Several studies have been made recently to enhance the performance of speaker ID systems for whispered speech. In [2], a whisperer speaker ID system achieved 8-33% relative improvement by using 5 to 15 seconds of whispered speech as adaptation data. In [3] and [4], several ways of compensation for male’s whisper were proposed for the purpose of text dependent/independent speaker identification using only neutral trained GMMs. All of these studies employed or proposed feature vector extraction method, based on short time analysis in the complete or a large continuous part of the frequency domain. In our study, first, we analyze the reason for the degradation of performance based on MFCCs-GMM. Next, feature vectors extracted from separate subbands based on long-time analysis are considered as complementary to MFCCs. TRAPs (or temporal patterns) are commonly used in phoneme classification by using the temporal patterns of spectral energy in each critical/mel subband [5, 7]. Our feature extraction method proposed here is different from TRAPs as follows: first, instead of using feature vectors extracted from a fixed 1 second temporal context, a dynamic length is employed based on the duration of each particular phoneme and Legendre polynomial fitting is applied to obtain equal length feature vectors. Second, unlike standard TRAPs, we not only use the logarithmic spectral energy contour from each frequency band as feature vectors, the spectral center gravity (SCG) contour of each subband are also concatenated after the energy contour to calculate the feature vectors. Third, because of the unique acoustic properties of whispered speech [4], a linear-band starting after 1000 Hz is used in this study instead of Mel-band or critical-band. Fourth, after feature extraction, no neural network training, PCA or LDA is applied to fuse the feature vectors in all subbands, instead, separate GMMs are trained for each subband. For each test using whispered data, the weighted score based on a phoneme basis from each subband GMMs are combined with scores of the MFCC based GMMs to obtain a final score.

The remainder of this paper is organized as follow. In Sec.2, a general introduction to the UT-Whisper database is first presented. Second, details regarding modified TRAPs are presented, which includes phoneme segmentation, a comparison between standard and modified TRAPs, and procedures for extraction of modified TRAPs. In Sec.3, a description of the training and test procedures for GMMs are provided and experimental results are shown. Conclusions and summary are included in Sec. 4.

This project was funded by AFRL under a subcontract to RAD Inc. under FA8750-05-C-0029 and the University of Texas at Dallas under Project EMMITT. Approved for public release; distribution unlimited.
2. Database and Feature Extraction

2.1. UT-Whisper Corpus Setup

The UT-Whisper corpus developed in [6] is employed here. A small sample of neutral and whispered speech was collected from a total of 10 native American English female subjects. In the spontaneous speech part, those subjects are asked 10 questions, where they are free to answer 3 of under whisper and the other 7 with neutral speech in a total of one or two sentences. In the read speech part, each subject reads 40 different phonetically balanced sentences from the TIMIT database with whispered as well as neutral speech and half of those sentences are required to speak with whispered speech. The whispered and neutral speech for both parts are manually separated to constitute our whisper and neutral corpus. From [6], we also note that all recordings include a pure-tone calibration test sequences to provide ground-truth on true vocal effort for all speakers and sections. Speech data was digitized using sample frequency of 16 kHz, with 16 bits per sample. Speech from all speakers was windowed with a Hamming window of 16 ms, with an 8 ms overlap rate for phoneme segmentation and calculation of spectral temporal patterns.

2.2. Modified Temporal Patterns (m-TRAPs)

Fig. 1 shows two whispered sentences from two speakers. We can see that for the first utterance shown in Fig. 1(a), most spectral information in vowels, diphthongs, liquids and glides is lost, with only one or two formants retained. However, for the second whispered utterance in Fig. 1(b), most spectral information is kept relatively well intact and most formant structure is preserved in whisper. The difference in spectral structures between these two utterances is related to the way these subjects produce whispered speech. For some speakers, they pronounce whisper with a much lower volume than others. Hence, even in quiet environments, for low power whisper, part of the spectral structure is hidden or buried in the background noise and results in whisper with low SNR. While in the relative higher power whisper, the corresponding SNR is relatively higher and thus most of the formant structures are preserved. Both ways of producing whisper are quite common among speakers. For example, in our 10 speakers corpus, two speakers’ whispered data have an SNR below 13 dB, two speakers have 73.9% whispered sentences with their SNR below 18 dB, and all other speakers’ whispered data has an average SNR of 20 dB. (note: all speakers’ neutral data has an average SNR of 29 dB). Short time analysis based features such as MFCCs or LFCCs contain spectral information from the complete or a large continuous part of the frequency domain and the coefficients represent the shape of the spectral energy envelop along the frequency axis in a short time domain. For the high SNR whisper, even though there is spectral differences with neutral speech, such as formant shifts in low frequency, lack of excitation, and spectral tilt changes, the overall spectral energy envelope is mostly preserved, so the MFCCs feature vectors are still similar with neutral and thus can still convey speaker information. However, for low SNR whisper, since the spectral structure in some subbands is missing, the resulting spectral energy envelope along the frequency axis is significantly different from that of neutral. Hence, the whisper feature vectors based on short time analysis in the complete frequency domain can not match the corresponding neutral case. For those sentences with an SNR above 20 dB in our corpus, MFCC based GMMs can achieve an average performance of 80%, while for the two speakers with all whisper’s SNR below 13 dB, no test utterance is recognized correctly using MFCC based GMM. We expect that the distribution and contour of subband energy and spectral center gravity belonging to different phonemes will differ among different speakers due to the location, size and movement of articulators, and thus contain speaker information. Noting the fact that for each phoneme in low SNR whisper, especially vowels, at least one or two formant tracks are preserved. In our study, we consider a modified temporal patterns (m-TRAPs) in separate subbands over a long temporal context as a viable method to represent such information.

Phoneme segmentation must be made first for both training (neutral) and testing (whisper) data in order to extract the m-TRAPs of each phoneme for different speakers. There are many studies regarding phoneme segmentation for neutral speech, but few exist for whisper. For whispered speech, after removing all silence parts by energy thresholding, phoneme segmentation for whisper speech is as follows: first, the ratio of the higher frequency energy to lower frequency energy is used to detect the fricatives and stops. Next, the frequency domain is divided into three parts: low frequency band (300-1500 Hz), middle frequency band(1500-3500 Hz) and high frequency band (3500-8000 Hz) for segmenting vowels, diphthongs, liquids and glides. The symmetric relative entropy of each frequency band belonging to neighboring frames are calculated as Eq. (1) to measure the similarity between neighboring frames. When the symmetric relative entropy belonging to one of the frequency domain falls below a certain threshold, we assume that the
speech has entered a new phonemes. Our experiment shows that this method is efficient and easily to realize.

\[
D(p_t, p_{t-1}) = \sum_{f_k \in \mathcal{F}} [p_t(f_k) \log \frac{p_t(f_k)}{p_{t-1}(f_k)} + p_{t-1}(f_k) \log \frac{p_{t-1}(f_k)}{p_t(f_k)}]
\]

where \( p_t(f_k) \) is the probability density for the spectrum and can be estimated as,

\[
p_t(f_k) = \frac{s(f_i)}{\sum_{k=1}^{M} s(f_k)}, i = 1 \cdots M
\]

where \( s(f_i) \) is the spectral energy of the frequency component \( f_i \), and \( M \) is 128 in our study.

TRAPs are commonly used in phoneme classification [5, 7], and only recently considered for text independent speaker identification [8]. In these studies, spectral feature vectors were extracted from a fixed duration of temporal data. In [5, 7], the author used 1 sec and in [8], the duration is 150 ms. However, in reality, the duration for different phonemes varies in different context, and even the same phoneme will differ in duration in the same sentence. Even calculating the mean of all trajectories belonging to all the same phonemes [5] can still not remove all effects from other neighbouring phonemes, and in [8] the proposed trajectory of each phonemes were either segmented into several pieces or buried among other phonemes. Hence, we use a dynamic temporal pattern length here, where the length equals the duration of the particular phoneme for which that trajectory belongs to. The spectral center of gravity trajectories in each subband are also obtained and concatenated after the energy contour to calculate the final feature vectors in order to improve performance. Based on our earlier result [4], we used a linear scale of 13 filter banks that start from 1000 Hz to obtain the spectral energy distribution in each subband. For the purpose of smoothing, decorrelation, and convenience of GMMs training/testing, the weighting coefficients belonging to the linear combination of a set of Legendre polynomials up to order 5 are used to represent the shapes of the spectral energy and gravity center for each subband, and hence we obtain a 10 dimensional feature vector. Fig. 2 illustrates the procedures for feature extraction for our modified TRAPs, and Fig. 3 shows an example of the smoothed energy and spectral center of gravity contours from the 10th filter bank of two speakers, where we can easily tell the difference of temporal patterns between these two speakers.

Figure 3: Comparison between two speakers’ smoothed energy and spectral center of gravity contour from the 10th filter bank, (a) speaker I’s energy contour, (b) speaker I’s SCG contour, (c) speaker II’s energy contour, (d) speaker II’s SCG contour
use neutral data for training because whispered adaptation data is usually not available in reality. If we apply the method in [5], [7] or [8], those subbands for which whisper has no speaker information, while neutral contains some individual messages, will be included in the final GMM for computing the score and degrade performance. With 13 separate GMMs for each subband, we can disregard or weight the score from those subbands mentioned above. Hence, in this study, we provide a simple phoneme-by-phoneme score weighting method. For each testing utterance, we first divide it into separate phonemes and for each phoneme only those subbands that contain formant structure will be considered for scoring, and the score from the remaining subbands will not be included in the final score for this phoneme. Next, scores from all phonemes will be summed together to obtain the score for this utterance. For the extraction of MFCCs, we only considered information above 1000 Hz for both whisper and neutral mode based on the result in [4]. The traditional method for training and testing of GMMs based on MFCCs is employed here. The combined score is represented as follows,

\[ S = \alpha S_{MFCC} + (1 - \alpha)S_{m-TRAPs} \]  

(3)

where \( \alpha \) is dependent on the SNR of the test utterance. If the SNR is below 18 dB, we assume this test data belongs to lower SNR whisper and \( \alpha \) is set to 0.2. While if SNR is above 18 dB, \( \alpha \) is set to 0.8.

3.2. Experimental Results

As noted in Sec. 2.1, we have 20 read TIMIT and 7 spontaneous sentences for neutral speech and 20 read TIMIT and 3 spontaneous sentences for whisper. First, half of the neutral data is used for training the GMMs based on m-TRAPs and the other half of the neutral data is used as test. For this neutral/neutral matched case, the final score is the sum of the scores from all subbands GMMs. A performance of 87.2% is achieved, which demonstrates the capability of modified TRAPs to carry speaker information. The loss of performance compared with traditional MFCCs is caused by the redundant information contained in some subbands.

Next, a traditional MFCC based GMM is used here for comparison. By using all the neutral data for training and all whisper data for testing, the baseline system obtains a performance of 44.1% and when only using test whisper sentences from the corpus with an SNR below 18 dB, the performance of MFCCs based GMMs is 6.0%. Next, subband GMMs based on m-TRAPs for each speaker are trained using all feature vectors from neutral speech. We obtain a performance of 57.5%, which is far beyond the MFCC baseline, by testing those low SNR whisper data using m-TRAPs and score weighting. When combining the MFCC based GMMs and m-TRAPs based GMMs together using the method in Sec 2.3 for all whisper test data, an overall 70.4% performance is achieved. The result is listed in Table 1.

Table 1: Experimental results from a close speaker set ID system.

<table>
<thead>
<tr>
<th>System (trained with neutral)</th>
<th>Speaker Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low SNR whisper</td>
</tr>
<tr>
<td>MFCCs</td>
<td>6.0%</td>
</tr>
<tr>
<td>M_TRAPs+MFCCs</td>
<td>57.5</td>
</tr>
</tbody>
</table>

4. Discussion and Conclusion

Whisper is an alternative speech production mode, which is used by talkers commonly in natural conversational scenarios to avoid being overheard and protect personal privacy. Due to the difference in production of whispered and neutral speech, speaker identification systems trained with only neutral speech degrade significantly when tested with whisper.

In this study, we first analyze reasons why traditional MFCCs-GMM speaker identification systems fail to provide effective overall performance for whispered test data. After introducing a simple but efficient entropy phoneme segmentation method for whisper, a modified TRAPs (m-TRAPs) and phoneme-by-phoneme score weighting strategies were proposed and combined together as a complement for the MFCC feature. Compared with traditional TRAPs, our feature vectors are extracted from dynamic length trajectories based on the duration of the particular phoneme, as well as the spectral gravity center in each subband being concatenated after the spectral energy. The modified-TRAPs and phoneme-by-phoneme score weighting methods demonstrate significant improvement for whisper data with low SNR and when combines with MFCCs, an overall improvement of 26.1% is achieved.

The results represent one of the first advancements in developing a seamless speaker ID system for whisper/neutral mismatch condition. Our future work will focus on demonstrating the effectiveness of the proposed algorithms in a larger and more comprehensive corpus for actual voice communication/voice dialog systems.

5. References