SPEAKER IDENTIFICATION WITH WHISPERED SPEECH BASED ON MODIFIED LFCC PARAMETERS AND FEATURE MAPPING

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

Much research recently in speaker recognition has been devoted to robustness due to microphone and channel effects. However, changes in vocal effort, especially whispered speech, present significant challenges in maintaining system performance. Due to the absence of any periodic excitation in whisper, the spectral structure in whisper and neutral speech will differ. Therefore, performance of speaker ID systems, trained mainly with high energy voiced phonemes, degrades when tested with whisper. This study considers a front-end feature compensation method for whispered speech to improve speaker recognition using a neutral trained system. First, an alternative feature vector with linear frequency cepstral coefficients (LFCC) is introduced based on spectral analysis from both speech modes. Next, for the first time a feature mapping is proposed for reducing whisper/neutral mismatch in speaker ID. Feature mapping is applied on a frame-by-frame basis between two speaker independent GMMs (Gaussian Mixture Models) of whispered and neutral speech. Text independent closed set speaker ID results show an absolute 20% improvement in accuracy when compared with a traditional MFCC feature based system. This result confirms a viable approach to improving speaker ID performance between neutral and whispered speech conditions.

Index Terms:whisper, speaker identification, linear scale cepstrum coefficients, feature mapping

1. INTRODUCTION

Whisper is an alternative vocal effort style from neutral speech which can be employed between speakers when conveying information a speaker may consider to be personal. For example, when making a hotel/car reservation over a cell phone in a public area, a speaker may not want to speak at the same vocal effort when giving their credit card information. Individuals with low vocal capability also employ whisper for their oral communication. Compared with neutral speech, whisper has no fundamental frequency because of the absence of voice harmonic excitation, and formant shifting exists in the lower frequency region [1][4]. A significantly different spectral structure exists between whispered and neutral speech and therefore presents unique challenges for effective speaker ID system performance.

Several efforts have been made recently to enhance the performance of speaker ID systems for the whispered speech mode. In [3], a whisper speaker ID system achieved 8-33% relative improvement based on the assumption that a small amount of whispered speech per speaker is available. However, speaker dependent whispered training data is generally not available in real scenarios. A system based on frequency warping and score competition was introduced in [8], which offered an initial step forward in addressing whisper/neutral speech for seamless speaker recognition. However, this approach was text-dependent, which limits its application in real applications. In this study, a 19-dimensional modified linear frequency cepstral coefficients is used as a feature vector that aims to remove frequency components that generally differ between whispered and neutral speech, while maintaining more spectral information that shares similar traits between both speech modes. Also, feature mapping, which has not been considered for mismatch in the speech mode for speaker ID, is applied here to modify the input test data based on two general UBMs (Universal Background Models) trained with whisper and neutral speech respectively. The integration of these two processing stages provides a meaningful step in developing a seamless speaker ID system for whispered and neutral speech.

The remainder of this paper is organized as follows: in Sect.2, a general introduction to the UT-Whisper database is presented. Second, we introduce the details for spectral comparison and feature extraction. Afterward, specific procedures for feature mapping are explained. In Sec.3, performance for closed set speaker ID based on the proposed method is

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

compared with an MFCC baseline system. Conclusions and a summary are drawn in Sec.4.

2. DATABASE AND SYSTEM DESCRIPTION

2.1. UT-whisper Corpus Setup

The UT-Whisper corpus developed in [6] and employed in [8] is also used here. A small sample of neutral and whispered speech was collected from a total of 10 native English male subjects. Each subject read 10 phonetically balanced sentences from the TIMIT database in two different speech modes: whisper and neutral speech. From [6], 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 a 16 ms overlap rate.

2.2. Linear Frequency Cepstrum Coefficient

In [2] and [4], it was shown that the degree of spectral difference between neutral and whispered speech varies in different frequency portions. Therefore, for the whisper/neutral mismatch situation, a feature extraction method which suppresses the mismatch spectral information while preserving similar frequency components, will achieve better performance for speaker ID. In order to find those frequency portions, we analyze both modes across frequency bands.

Since the spectral slope of whispered speech is known to be more flat [5], the same adaptive pre-emphasizer used in [8] is also applied here. Also, an LP-based power spectrum was chosen here instead of an FFT to compute the spectral energy of each frame. Since one primary difference between neutral and whisper is the voiced excitation, an LP-based power spectrum has the advantage of characterizing information of the vocal tract function while suppressing most excitation information. A 24-band linear-scaled triangular bandpass filter shown in Fig. 1(a) with LP-based spectral energy as input is used to obtain a 24 dimensional vector for each frame to represent the spectral energy distribution. For each speech mode, a UBM was trained after applying a log function to the vectors from all speech in our corpus. Next, Eq. (1) is used to fuse the mean of the UBM mixture together to obtain a general difference between whisper and neutral speech among various phonemes,

$$\mu' = \sum_{m=1}^{M} \omega_m \mu_m. \qquad (1)$$

where M is the total mixture size for the UBM, which is 32 in this study; $\omega_m$is the mixture weight of the m^{th} component of UBM; μ_m is the mean of UBM's m^{th} component; and μ' is the fused mean of the UBM. Two fused means for whisper and neutral speech respectively can be obtained with Eq. (1), which are both plotted in Fig. 1(b) for comparison. As can be seen here, the whisper and neutral speech's spectral energy share more similarity above 1000 Hz, while there are significant differences in the lower frequency portion. It is known that the Mel-scale emphasizes the low frequency portion and de-emphasizes higher frequencies by placing more filters in the lower frequency domain, and therefore, fails to take advantage of this spectral characteristic of whisper and neutral speech. Hence, a linear scale filterbank is applied here for feature extraction to retain more higher frequency structure and we therefore remove the spectral information from 0 Hz to 1000 Hz, which corresponds to the range covered by the first 3 filter banks in Fig. 1(a). The cosine transform is then applied to the **log** energy obtained from the remaining 21 linear filters. Only the first 19 coefficients are kept as the feature vector in the following experiment. It is noted that while LFCC was discussed here, it has been employed extensively in other research studies for speaker recognition. The particular implementation here is modified based on the knowledge learned between whispered/neutral speech.

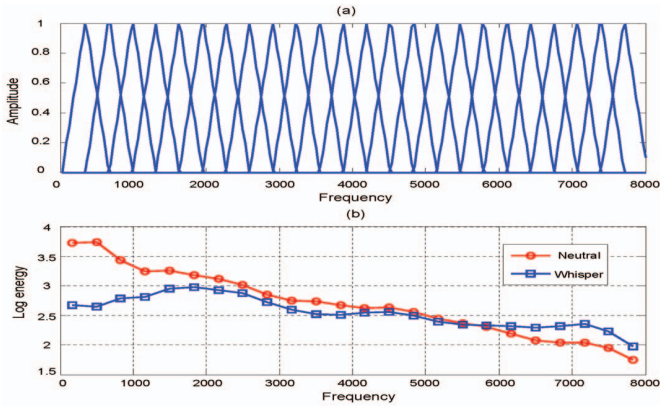

Fig. 1. Comparison between fused mean for whisper and neutral UBM, (a) linear frequency filter bank, (b) resulting long-term spectral energy present versus frequency

2.3. Feature mapping for whispered speech

A number of feature compensation and mapping techniques have been proposed for channel compensation [7,9,10]. However no methods have considered feature mapping for whisper/neutral speaker ID. To compensate for the difference between neutral and whispered speech and further improve the performance, feature mapping was applied on a frame-by-frame basis. Compared with other attempts to compensate for whispered speech based on fixed statistics[1], feature mapping has the advantage of addressing the variability of the difference between whisper and neutral speech among different phonemes.

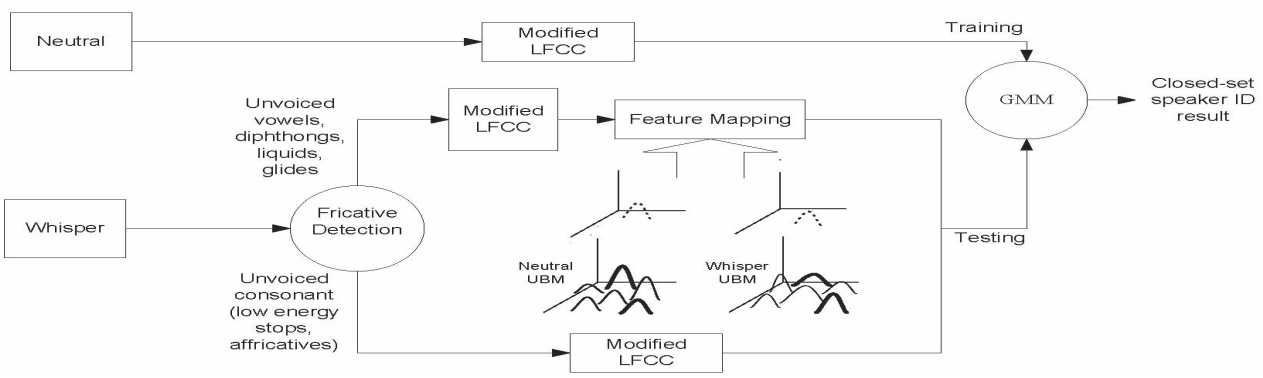

Fig. 2. System flow diagram for close set speaker ID system for whispered speech.

To begin with, all speech from the UT-Whisper corpus was used to obtain a speaker and speech mode independent 32 mixtures UBM. Next, a speaker independent whisper and neutral UBM was obtained by adapting that UBM with available whisper and neutral data, respectively. For observation vector $\mathbf{x}_t$ at time t, both the whisper UBM Λ_w and neutral UBM Λ_n are tested. For each UBM, we first compute the output probabilities of each mixture $N(\mathbf{x}_t|\mu_m, \mathbf{\Sigma}_m, \omega_m)$. Next, the probability of the m^{th} mixture given vector $\mathbf{x_t}$ will be given by Eq. (2) as follows:

$$Pr(m|\mathbf{x_t}) = \frac{\omega_m N(\mathbf{x}_t|\mu_m, \mathbf{\Sigma}_m, \omega_m)}{\sum_{m=1}^{M} \omega_m N(\mathbf{x}_t|\mu_m, \mathbf{\Sigma}_m, \omega_m)}, \quad (2)$$

where $\mu_\mathbf{m}$, $\mathbf{\Sigma_m}$ and ω_m are the mean vector, covariance matrix, and mixture weight of the m^{th} component of the UBM. In this study, only diagonal covariance GMM is considered.

Pr in Eq. (2) represents the probability that observation $\mathbf{x_t}$ can be classified into the cluster of phonemes represented by the m^{th} mixture. The higher the probability, the higher the chance that $\mathbf{x_t}$ belongs to the corresponding cluster of phonemes. However, most of the time, an observation will fall between two or even more clusters of phonemes. Therefore, selecting the one with the highest Pr is not sufficient. In this study, based on the fact the total mixture size is 32, we choose the first two mixtures with the highest probabilities and combine them to obtain a mean and variance to represent the possible cluster of phonemes for each observation. The mathematical expectation of the means belonging to the two highest possible mixtures are calculated according to Eq. (3),

$$\mu_{si} = \frac{\sum_{k=1}^{2} Pr(k|\mathbf{x}_t)\mu_k}{\sum_{k=1}^{2} Pr(k|\mathbf{x}_t)}. \quad (3)$$

In the same way, it is possible to compute a mathematical expectation of the variance of these two mixtures. According to Eqs. (2) and (3), a speaker independent mathematical expectation of mean and covariance of observation $\mathbf{x}_t$ can be obtained for both the whisper and neutral UBM and are represented as: μ^{siw}, $\mathbf{\Sigma}^{siw}$, μ^{sin}, and $\mathbf{\Sigma}^{sin}$. In this study, all test data, represented as $\mathbf{x}_t$ in the above equations are whispered speech. Hence, when using μ^{sin}, and $\mathbf{\Sigma}^{sin}$ obtained from $\mathbf{x}_t$ and the neutral UBM, we assume that even whispered and neutral speech differs, for the whisper vowel and vowel-like part, the neutral UBM can still classify the data to the corresponding neutral phoneme cluster. This assumption is reasonable considering the fact that each vowel has its unique vocal tract function and the corresponding whispered speech keeps the basic structure even with some shift in F1, F2, which is already partly suppressed by our feature extraction method (since data from 0-1 kHz is removed). As for the whispered unvoiced consonant frames, because they share more similarity with the neutral ones [2], no feature mapping was applied. Fig. 2 illustrates both unvoiced consonant and vowel-like/vowel based feature processing. In order to map $\mathbf{x_t}$ to a new feature vector that compensates for the difference between neutral and whispered speech, Eq. (4) is used to modify each dimension of $\mathbf{x_t}$,

$$x_{tn}^{'} = x_{tn} + \delta, \quad (4)$$

$$\delta = (\mu_n^{sin} - \mu_n^{siw})\sqrt{\frac{\mathbf{\Sigma}_n^{sin}}{\mathbf{\Sigma}_w^{sin}}}. \quad (5)$$

where n is the dimension index, which ranges from 1 to 19, Σ_n is the n^{th} component of the diagonal matrix Σ. Based on the above analysis, δ can be seen as the speaker independent difference between neutral and whispered speech particularly to the cluster of phonemes corresponding with $\mathbf{x}_t$. Hence, a level of compensation was made to map the test whisper data to neutral speech.

3. EXPERIMENTAL RESULT

As noted in Sec. 2.3, feature mapping was applied only to the vowel/vowel-like frames of whisper, so the consonant detection method used in [8] was also employed here.

Also, in order to compare improvement achieved by our proposed method, a baseline system was developed based on 19-dimensional static LP-MFCC. Next, 19-dimensional static modified Linear Frequency Cepstral Coefficients as presented in Sec.2.2 were used as the second feature vector. Finally, feature mapping and LFCC were combined together to demonstrate the potential impact of our method. Again, Fig. 2 depicts the overall flow diagram for the final system. Five neutral sentences from the TIMIT database for each speaker were used to build ten corresponding GMMs. Also, five whisper utterances from each speaker, different from the ones corresponding to the neutral training data, were used for testing. In order to demonstrate reliability of the models trained with neutral speech, another five neutral utterances that were not used in training were employed first for testing, resulting in a 94% speaker ID accuracy.

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

System(trained with neutral)	Speaker Recognition rate	
	Neutral	Whisper
MFCC	94%	48%
LFCC	x	58%
LFCC+Feature mapping	x	68%

Table 1 summarizes the experimental results. The baseline system using GMMs trained with neutral 19-dimension static MFCC vectors provides a closed speaker set recognition rate of 48% using whispered speech. When 19-dim MFCC was substitute with 19-dim LFCC, a +10% improvement is observed. If both feature mapping and LFCC are applied, the resulting system achieves a 68% speaker recognition rate, which represents +20% improvement over the original baseline system performance.

4. DISCUSSION AND CONCLUSION

Whisper is an alternative speech production mode that is commonly used for communication in public circumstances to protect personal privacy. However, the performance of traditional speech systems degrades due to this whisper/neutral mismatch situation, because of the significant difference between neutral and whispered speech production.

In this study, a new closed set speaker ID system was established based on a modified set of LFCC feature and feature mapping. Through spectral distribution analysis for both speech modes with UBMs combined with linear scaled filter banks, it was observed that whisper and neutral speech share similarities in the higher frequency range, while they differ from each other in the frequency range 0 to 1000 Hz. LFCC was applied as a feature vector in order to preserve similar information shared between them, while eliminating those parts that are different (e.g., 0-1 kHz). Feature mapping, which has been used for channel compensation, was introduced in this study to help address challenges brought to speaker ID systems by vocal effort mismatch since it can capture the difference in variability between whisper and neutral speech among various phonemes. Feature mapping was applied only to frames identified as vowel/vowel-like (includes vowels, diphthongs, liquids, glides), since a difference between neutral and whispered speech is seen in vowel/vowel-like part. Using a previously collected corpus, the advantage of both techniques were shown. With modified LFCC, +10% improvement was observed and when modified LFCC was combined with feature mapping, the system achieved an absolute +20% enhancement compared with an MFCC baseline system.

For future work, it is clear that a larger and more comprehensive corpus is needed to demonstrate the repeatability of the proposed methods in actual speech communication systems. Yet, the results do represent one of the first advancements in developing a seamless whisper/neutral speaker ID system, and also confirms the viability of the proposed methods for improving performance on whisper/neutral speech conditions.

5. REFERENCES

[1] X. Li, "Reconstruction of Speech from Chinese Whispers," [PhD], Nanjing University, China, 2004

[2] T. Ito, K. Takeda and F. Itakura, "Analysis and Recognition of Whispered Speech," Speech Communication 45, pp.139-152, 2005

[3] Q. Jin, S. S. Jou and T. Schultz, "Whispering Speaker Identification," IEEE International Conference on Multimedia and Expo, 2007.

[4] M. Matsuda and H. Kasuya, "Acoustic Nature of the Whisper," EUROSPEECH, pp.137-140, 1999

[5] S. T. Jovicic, "Formant Feature Differences between Whispered and Voiced Sustained Vowels," Acustica-acta, 84(4), pp.739-743, 1998

[6] C. Zhang and J. H. L. Hansen, "Analysis and Classification of Speech Mode: Whisper through Shouted," INTERSPEECH 2007

[7] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," vol.2, 2003, pp.II-53-6,in IEEE International Conference on acoustic, speech and signal processing

[8] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," INTERSPEECH 2008

[9] P.J. Moreno, B. Raj, R.M.Stern, "Data-driven environmental compensation for speech recognition: a unified approach", Speech Communication 24 pp.267-285, 1998.

[10] M. Mason, R. Vogt, B. Baker and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification", pp, 3109-3112, In INTERSPEECH-2005.