WHISPERING SPEAKER IDENTIFICATION

Qin Jin, Szu-Chen Stan Jou, and Tanja Schultz

Interactive Systems Laboratories, LTI SCS, Carnegie Mellon University
{qjin|scjou|tanja}@cs.cmu.edu

ABSTRACT

This paper describes a study of automatically identifying whispering speakers. People usually whisper in order to avoid being identified or overheard by lowering their voices. The study compares performances between normal and whispered speech mode in clean and noisy environment under matched and mismatched training conditions, and describes the impact of feature warping and throat microphone on noise reduction. Score combination strategies are used when only little whisper data is available to improve performance. In sum, we achieved 8% to 33% relative improvements in identification accuracy with only 5 to 10 seconds noisy whispered speech data per speaker.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing the person from his/her voice. Speaker recognition technologies can provide a way to manage and access multimedia databases, which is to retrieve information according to interested speakers. Although speaker recognition has drawn a lot of attention in the research community over the last years, published studies focus on the recognition of "cooperative speakers", i.e. speakers who are willing to be overheard and identified by an automatic speaker recognition system. Uncooperative speakers who would like to avoid being identified may try to intentionally fool a system by changing their speaking behavior or lowering their voices. Kajarekar et al. investigated the effect of intentional voice modifications on the speaker recognition and showed vulnerability in both humans and speaker recognition systems to changed voices [8].

In this paper, we investigate the recognition of speakers who whisper to either intentionally disguise their voice to not being identified or who want to have a confidential conversation in public places and do not want the content of the conversation to be overheard. We investigated how to improve speaker identification accuracy in presence of such uncooperative speakers under both quiet and noisy environment.

For ethical reasons we refused to record data from speakers without their explicit knowledge and permission. Nor did we record these speakers at times when they did not expect to be recorded. We rather instructed our subjects to speak in a whispered mode, pretending that they want to only be heard by a person standing very nearby. In general it is expected that very few such training data are available. Therefore, we are investigating strategies to counteract mismatched conditions. In this study we will (1) compare speaker identification performance between normal and whispered speech modes assuming that sufficient data in both speaking modes are available (up to 90 sec training data per mode), (2) compare speaker identification performance between quiet and noisy environment (3) examine the system performance when speaker models are trained on normal speaking mode but evaluated on whispered speaking mode, and (4) show improvements of score combination strategies that combine the scores of matched and mismatched speaker models.

2. DATABASE AND EXPERIMENTAL SETUP

For the experiments, a small sample of normal and whispered speech data are collected from 22 subjects. In a quiet room, each person reads sentences in two different styles of articulation: normal and whispered. The recordings of both articulation styles were done using both a throat microphone and a close-talking microphone simultaneously. The throat microphone used in our experiments is made of piezoelectric ceramics and can be mounted by wearing it around the neck. It is a commercial product made by Voice Touch. We chose this microphone because it has the best spectral resolution among contact microphones we have experimented with. Similar to [11], we used a USB external sound card to record two channels simultaneously. One channel contains the throat microphone recording, while the other contains the regular close-talking microphone recording. The close-talking microphone is a Sennheiser HMD 410. For each articulation style, there are 50 sentences including 38 phonetically balanced sentences and 12 sentences from news articles. In contrast, a noisy session of recording was also held with the same protocol except that cocktail-party babble noise was played via a pair of loud speakers during the whole session [7]. The training data in different durations (90, 60, 30, and 15 seconds) is selected from the 38 phonetically balanced sentences and the test data in different durations (20, 15, 10, 5, 4, 3, 2, and 1 second) is selected from the 12 news article utterances. The format of the recordings is 16 kHz sampling rate, 2 bytes

1-4244-1017-7/07/$25.00 ©2007 IEEE

per sample, and linear PCM. The total number of test trials is 190, same for all the different test durations.

In our following experiments, there are two evaluation conditions: matched and mismatched. Matched condition refers to the condition in which the training and test data are in the same speaking mode and same acoustic environment. While for mismatched condition, there will be acoustic environmental mismatch and speaking mode mismatch. Our final goal is to improve speaker identification performance when speaker models are trained on clean normal speech but evaluated on noisy whispered speech. Speaker identification accuracy is used for performance measurement, which is the percentage of correctly identified test trials.

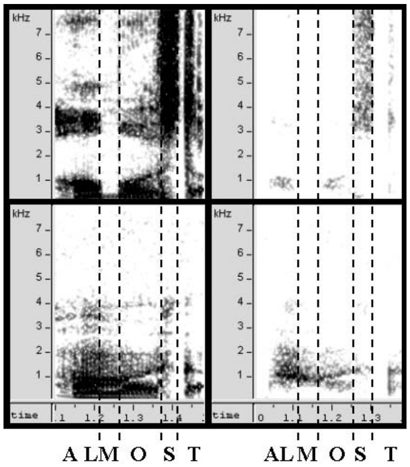

Fig. 1. Spectrogram of the word '*ALMOST*'. Upper row: close-talking mic. Lower row: throat mic. Left column: normal speech. Right column: whispered speech.

Table 1. SNR for different mic and environment

	Close-talking Mic		Throat Mic	
	Clean	Noisy	Clean	Noisy
Normal	29.0	11.5	36.2	37.7
Whispered	9.3	2.5	4.6	12.2

Whispered speech is produced by holding vocal cords open and without vibration. While the speech motor control plans are similar in normal and whispered speech, whispered speech has flatter spectral shape and its spectral variation is smaller [9] [10]. Figure 1 shows the spectrogram of the word 'ALMOST' in normal speaking mode on the left and whispered speaking mode on the right. We can tell from the figure that whispered speech has significant lower signal-to-noise ratio (SNR) and lower energy than normal speech. Table 1 shows the SNRs for different microphones and environments. In the noisy data we observed Lombard effect which means "characteristic changes in articulation due to environmental influence" [12]. Therefore, SNR on throat microphone is higher under the noisy condition than its clean counterpart. Figure 1 also shows that the throat microphone recording is low-passed at about 4 kHz. We believe the reason is due to the bandwidth of skin vibration on the throat.

3. SYSTEM DESCRIPTION

3.1. Feature Processing and Speaker Modeling

Our system uses 13 Mel Frequency Cepstral Coefficients (MFCC) as speaker features and Gaussian Mixture Model (GMM) with 128 Gaussian components as speaker models. We also compared different features such as MFCC, LPCCEP, and wavelet. MFCC achieves better performance than the other two in both speaking modes. We applied the Feature Warping technique [2] [3] [4] over MFCC, which warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval. It appears to be a good way to make the features more robust to different channel and noise effects. The warping is implemented via CDF matching. In our experiments, the window size is 300 frames and the window shifts one frame. Zeros are padded at the beginning and at the end of the raw feature stream, which is MFCC in this paper.

3.2. Score Combination

Suppose we have sufficient normal speech data but only small amount of whispered speech data for each speaker, which is realistic: it is easier to collect normal speech data than whispered speech data. How can we utilize the data to do better job for recognizing speakers when they whisper? We can train two GMM models for each speaker with each one in one of the speaking mode. We propose this "frame based score competition (FSC)" approach to compute the likelihood of an observation given the two GMM models.

The likelihood of an observation (for example one feature vector x_n) given a GMM model Θ^k of speaker k is estimated as

$$p(x_n|\Theta^k) = \sum_{i=1}^{M} \frac{w_i}{\sqrt{2\pi|\Sigma_i|}} exp\{\frac{-(x_n-\mu_i)^T\Sigma_i^{-1}(x_n-\mu_i)}{2}\}$$

Also, the entire set of feature vectors X are assumed to be independent and identically distributed (i.i.d.). Accordingly, the likelihood of observation sequence X given Θ^k is estimated as

$$p(X|\Theta^k) = \prod_{n=1}^{n=N} p(x_n|\Theta^k)$$

We call the likelihood value as "score" in the following sections. As mentioned earlier in this section, a realistic situation is that we have enough amount of normal speech samples and small amount of whispered speech samples, we can then train two models for each speaker with one in each speaking mode. In FSC, we compare a feature vector of each frame to both GMMs $N\Theta^k$ and $W\Theta^k$. $N\Theta^k$ and $W\Theta^k$ refer to the GMM model of speaker k trained on normal speech and whispered speech respectively. we pick the highest score for the given frame. So the likelihood of the entire set of feature vectors X is estimated as

$$p(X|\Theta^k) = \prod_{n=1}^{n=N} \max\{p(x_n|N\Theta^k), p(x_n|W\Theta^k)\}$$

4. EXPERIMENTAL RESULTS

4.1. SID under Matched Condition

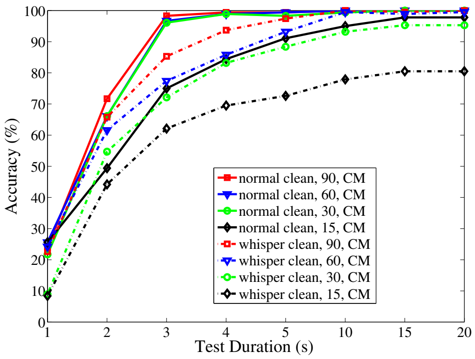

Fig. 2. Accuracy on normal vs. whispered Speech

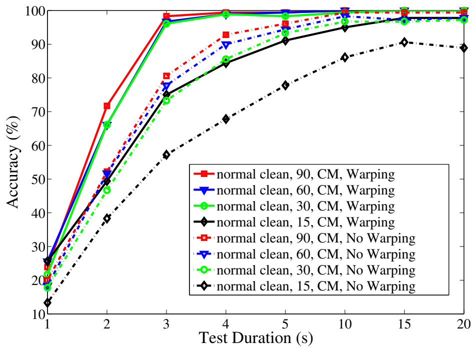

Fig. 3. Impact of feature warping

Our first set of experiments aims to compare the speaker identification performance between normal and whispered speaking mode under matched conditions. Figure 2 compares identification accuracy with different training durations and test durations. The first column in the legend refers to the training/test condition, and the number in the second column represents the training duration, and last column means the microphone type: CM refers to close-talking mic and TM refers to throat mic. It is obvious that performance increases with more training data and longer test duration. This is true for both normal and whispered speech. We can also see that better performance is achieved on normal speech than on whispered speech, which indicates that whispered speaker identification is more difficult.

Figure 3 shows performance improvement by feature warping on normal clean speech. Significant improvement is achieved by applying feature warping, which matches our results as shown in [4]. Therefore, feature warping is applied in the feature extraction step in all the following experiments.

Figure 4 compares the speaker identification accuracy on clean whispered speech vs. on noise whisper speech with close-talking microphone. We see performance degradation due to the presence of noise. Figure 5 compares system performance under noisy environment with close-talking microphone vs. throat microphone. We can see significant improvement is achieved by using throat microphone.

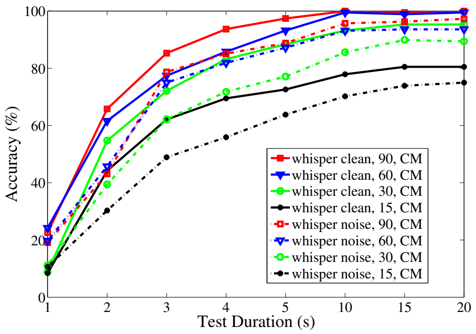

Fig. 4. Performance degradation by noise

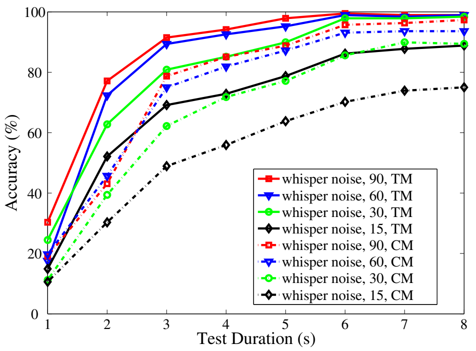

Fig. 5. Performance improvement by throat microphone

4.2. SID under Mismatched Speaking Mode

Our next set of experiments is to examine system performance on noisy whispered speech when sufficient clean normal speech data is available (for example 90 seconds) while no or only small amount of noisy whispered speech samples is available during training. As we have shown above that throat microphone provides better noise robustness, in this set of experiments we focus on data recorded using throat microphone. In figure 6, the bottom curve shows speaker identification baseline under mismatched condition, which means speaker models are trained on 90-second clean normal speech and evaluated on noisy whispered speech. In this case, we don't have any noisy whispered speech samples for training at all, so we label it as 'Mismatched + 0'. There is no doubt that seeing some noisy whispered speech samples in training will help improve the performance. Then the question is: if we only have small amount of noisy whispered speech samples for each speaker, how should we use them? Intuitively, we can use the small amount of noisy whispered speech samples to train a matched speaker model and then we can do test under matched condition. The dotted middle curves shows the identification accuracy under the matched condition with 5 to 15 seconds noisy whispered speech training data. The solid lines refer to the identification accuracy by applying FSC to combine the clean normal speech model with noisy normal speech model. We see clear improvement when test duration is longer than 3 seconds. Especially when only 5 seconds of noisy whispered speech samples are available during training, 33% relative improvement is achieved compared to the performance under corresponding matched condition.

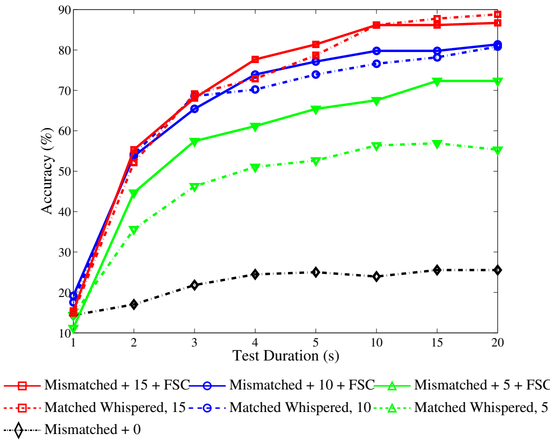

Fig. 6. FSC impact on system performance

5. CONCLUSION

People intend to lower their voices, i.e. to whisper, when they want to avoid being identified or overheard in public places. We investigated how to improve speaker identification accuracy in presence of such uncooperative speakers. We presented a series of experiments of speaker identification on whispered speech. By comparing the performance of our SID system on normal vs. whispered speech, we showed that SID on whispered speech is a more difficult task. The experimental results showed that feature warping brought significant improvement. The presence of noise hurts the SID system performance. Throat microphone provided more noise robustness. In the case of small amount of noisy whispered speech samples available during training, we presented a new score combination approach: "frame based score competition (FSC)". This approach utilized two models per speaker, one trained on normal speech and the other trained on whispered speech, in a competing way. This approach achieved significant improvements over the matched baseline. In sum, we reached 8% to 33% relative improvement with 5 to 15 seconds of noisy whispered speech samples per speaker.

6. REFERENCES

[1] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Speech Communication, Vol. 17, No. 1-2, p. 91-108, August 1995.

[2] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," Proc. Speaker Odyssey 2001 conference, June 2001.

[3] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy and R. Gopinath, "Short-time Gaussianization for Robust Speaker Verification," in Proc. ICASSP, 2002.

[4] Q. Jin, Y. Pan, and T. Schultz, "Far-field Speaker Recognition," in Proc. ICASSP, 2006.

[5] H. Valbret, E. Moulines, and J. P. Tubach, "Voice Transformation Using PSOLA Technique," Speech Communication, vol. 11, pp. 175-187, 1992.

[6] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," Computer Speech and Language, vol. 12, pp. 75-98, 1998.

[7] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone," in Proc. ICSLP, 2004.

[8] S. Kajarekar, H. Bratt, E. Shriberg, and R. Leon, "A Study of Intentional Voice Modifications for Evading Autmatic Speaker Recognition," in Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop.

[9] J. Coleman, E. Grabe, and B. Braun, "Larynx Movements and Intonation in Whispered Speech," in Summary of research supported by British Academy grant SG-36269, 2002.

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

[11] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement," in Proc. ASRU, 2003.

[12] L. Rabiner, and B.-H. Juang, "Fundamentals of Speech Recognition," Prentice Hall, New Jersey, 1993.