ACOUSTIC ANALYSIS AND RECOGNITION OF WHISPERED SPEECH

Taisuke Itoh, Kazuya Takeda and Fumitada Itakura

Center for Integrated Acoustic Information Research,
Nagoya University, Nagoya 464-8603 Japan.
Email: takeda@nuee.nagoya-u.ac.jp

ABSTRACT

In this paper, acoustic properties and the recognition method of whispered speech are discussed. A whispered speech database that consists of whispered speech, normal speech and their corresponding facial video images of more than 6,000 sentences from 100 speakers was prepared. The comparison between whispered and normal utterances show that 1) the cepstrum distance between them is 4 dB for voiced and 2 dB for unvoiced phonemes, respectively, 2) the spectral tilt of the whispered speech is less sloped than the normal speech and 3) the frequency of the lower formants (below 1.5 kHz) is lower than that of the normal speech. Acoustic models (HMM) trained by the whispered speech database attain an accuracy of 60% in syllable recognition experiments. This accuracy can be improved to 63% when MLLR adaptation is applied, while the normal speech HMM adapted with the whispered speech attain only 56 % syllable accuracy.

1. INTRODUCTION

Whispering is a common mode of speaking to communicate quietly or privately. Recently, as the use of the cellular phone spreads widely, the needs for private speech communication in a public place are increasing. In that situation, we usually whisper to the phone so as to reduce the amount of speech being spilled out. Since we do not make vocal cord vibration, the main sound source of the whispered speech is aspiration. Therefore, the standard source-filter model of speech production does not hold in the whispered speech case. Furthermore, because of the lower sound pressure level as compared to normal speech, whispered speech usually has a lower SNR, especially in a public place. Therefore, speech processing, especially recognition, of whispered speech is expected to be more difficult than the normal speech.

This research has been supported by a Grant-in-Aid for COE Research (No. 11CE2005).

Although the importance of processing whispered speech is increasing in the above-mentioned background, there is relatively low amount of research on this topic. The literature provides some research carried out on the prosodic information carried by the whispered speech [1], [2], [3]. Research on gender discrimination [4] and formant-frequency estimation of the whispered vowels [5] are also reported. However, no systematic effort has been reported on the recognition of whispered speech.

In this paper, we report on a study of the acoustic properties and recognition method for whispered speech. For this study, we constructed a database having *parallel* utterances of the normal and the whispered speech with facial video images of both utterances for more than 100 speakers. Using this database, taking frame-by-frame correlates between the whispered and normal speeches, we can directly compare the nature of the two different utterances. Furthermore, speech recognition experiments have been carried out in two different methods. In the first experiments, a whispered speech HMM is trained by the database, and in the second experiment MLLR adaptation is used. Throughout the experiments, we have found that more than 60% syllable accuracy can be obtained for the whispered speech.

2. DATABASE

For the database construction, each speaker whispered 60 sentences among which 50 sentences compose a phonetically balanced set for training and 10 sentences are for testing. The same 60 sentences are also recorded using a normal speaking style in the same session. Speech data are digitized into 16 bits with 16 kHz sampling frequency. To this date, whispered and normal speech of 62 male and 49 female speakers have been recorded.

In addition to the speech recording, facial video images are also recorded by DV camera and stored in an AVI format. The size of the image is 720x480 (pixel). Figure 1 shows the lower half of an example image.

Phoneme boundary information is available in all

Fig. 1. An example of facial video image. (The lower half of the original recording.)

utterances in the database. For whispered speech, the boundary information is given through frame-by-frame correlates to the normal speech. The correlation is calculated by DTW. The phoneme alignment information of the normal speech is given by a standard Japanese HMM [6]. As for the local distance used in DTW for boundary detection between normal and whispered speeches, we have tested an acoustic measure, i.e. cepstral distance, and a visual measure, i.e., square norm of the brightness. From the preliminary comparison between the two measures, we found that the acoustic measure provides more accurate boundaries since some speakers make lip-movement before the utterance starts. Furthermore, from the DTW results, we have found that the whispered speech has more insertions of short pauses because more frequent breathing is needed.

3. ACOUSTIC ANALYSIS OF WHISPERED SPEECH

In order to investigate the acoustical properties of whispered speech, the averaged spectrum of both speaking styles are calculated and shown in Figure 2 (on the next page) for 24 Japanese phonemes.

In Figure 3, the cepstrum distances between whispered and normal speech are illustrated for 24 phonemes. Initially, cepstrum distances are calculated for each phoneme occurrence, then, averaged over all occurrences. The cepstrum distance value is calculated by

$$\frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{29}(c_i^{(n)} - c_i^{(w)})^2},$$

where c^n and c^w are cepstrum coefficients of normal and whispered speeches, respectively. For the spectrum analysis, a 25 ms Hamming window with 10 ms shift was used.

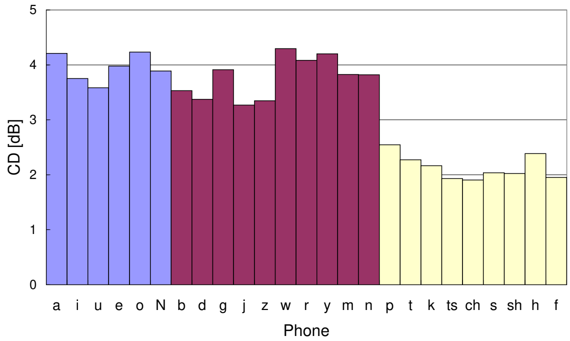

Fig. 3. Cepstrum distance between whipered and normal speeches.

The results obtained are as follows. For the voiced phonemes:

- The power of the lower frequency band, i.e., less than 1.5 kHz, is smaller and the spectral tilt of the whispered speech is less sloped than that of normal speech.
- The frequency of the lower frequency formants shifts to a lower frequency.
- The cepstrum distance between the normal and whispered speech is about 4 dB.

For the voiceless phonemes:

- The cepstrum distance between the normal and the whispered speech is less than 2 dB.
- There is no significant difference in the spectral shape of the phonemes that has the same place of articulation but the voiced-unvoiced distinction exists.

4. RECOGNIZING WHISPERED SPEECH

Recognition experiments of whispered speech have been carried out using the devised database. As for the training data, 4,000 sentences uttered by 40 male and 40 female speakers are used whereas 200 sentences uttered by 2 male and 2 female speakers are used for the evaluation task. Feature parameters used for the experiments are listed in Table 4. The trained whispered speech model consists of 43 monophone HMMs. Each monophone model has 3 states with 32 mixture densities.

For comparison purpose, the same structured monophone HMMs of the normal speech are also trained using a standard Japanese speech database [7] which consists of 14,000 sentences. The recognition accuracy is measured by the syllable recognition accuracy.

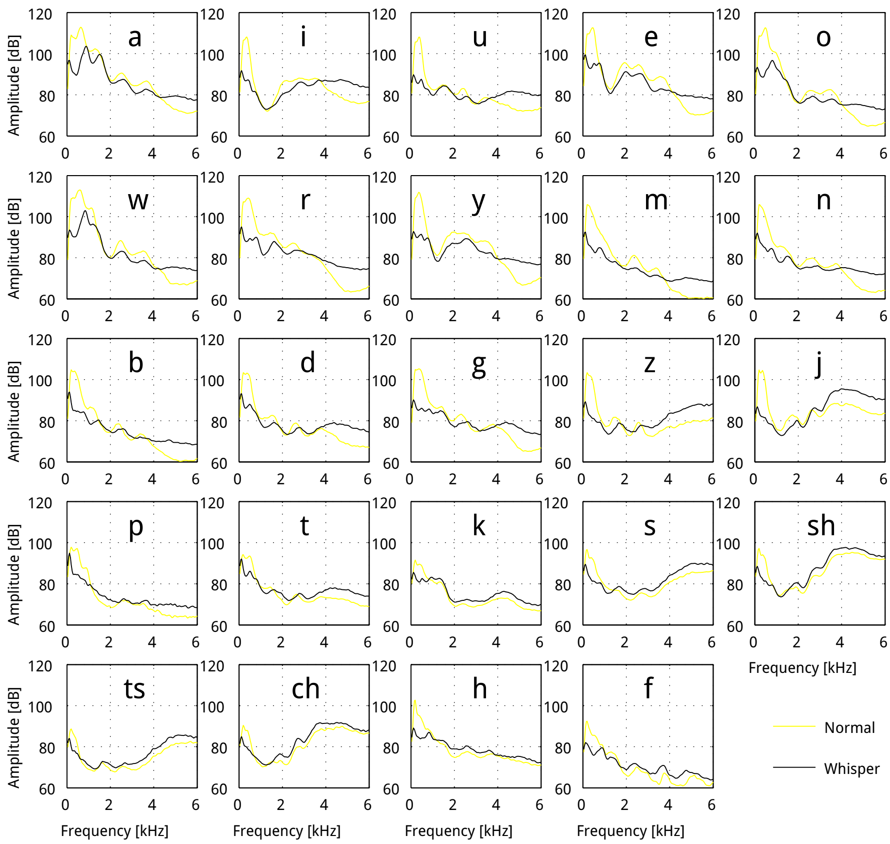

Fig. 2. Comparison between the averaged spectrums of whispered and normal speeches.

Table 1. Feature parameters for recognition experiments.

analysis window	Hamming
frame width	25 ms
frame shift	10 ms
feature parameters	mfcc(12) $+\Delta$ mfcc(12) $+\Delta$power

Table 2. Conditions for the adaptation experiments.

	original model	adaptation data no. of sents / no. of spkrs
(1)	whisper	10/target spkr
(2)	normal	10/target spkr
(3)	whisper	80/ non-target 80 spkrs

4.1. Recognition Results

The recognition results are shown in Figure 4. The baseline performance of this experiment, i.e., the accuracy of the HMM trained by the normal speech (normal model), was 80% in syllable accuracy. When the normal model is applied directly to the whispered speech, the recognition accuracy falls to less than 40%. Using the HMMs trained by the whispered speech (whisper model) attains 60 % accuracy for the whispered speech, whereas the whisper model can recognize the normal speech with approximately 50 % accuracy.

The results also show that the performance difference between baseline performance and the whispered speech model on whispered speech is larger in accuracy than %correct score. The reason for this difference is the insertion errors due to the misclassification among voiced and unvoiced phonemes that have the same place of articulation. Therefore, incorporating the lexical information is expected to improve the performance of the whisper model.

4.2. Adaptation

Finally, the effectiveness of MLLR[8] adaptation to the whispered speech is tested in order to examine if the normal model can be used for whispered speech through adaptation. The experiments are performed in three conditions as listed in Table 4. The number of adaptation matrices is set to 40.

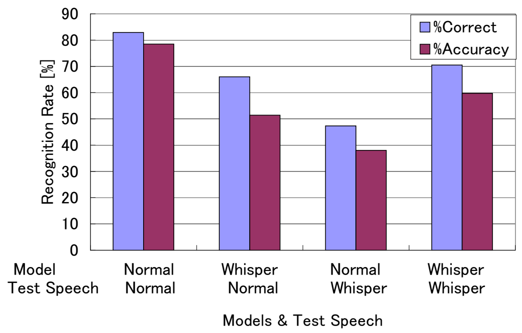

Fig. 4. Recognition results in syllable recognition rate. From left to right, 1) recognizing the normal speech by the normal speech models, 2) recognizing the normal speech by the whispered speech models, 3) recognizing the whispered speech by the normal speech models and 4) recognizing the whispered speech by the whispered speech models.

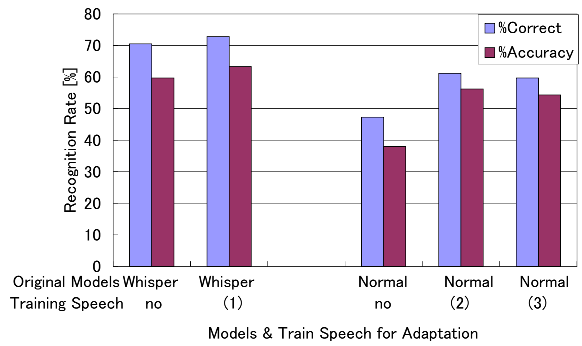

Fig. 5. Recognition results of the adapted models in syllable recognition rate. From left to right, 1) original performance of the whispered speech model, 2) its adaptation result, 3) original performance of the normal speech model, 4) its adaptation result by the 10 sentences of target speaker's speech and 5) its adaptation result by the 80 sentences of non-target speaker's speech.

The recognition results using MLLR adaptation is summarized in Figure 4. When the adaptation is done using 80 sentences of whispered speech, the performance of the normal model is improved by 17% (from 38% to 55%) even though the adaptation data does not include the target speaker's speech. There is no significant difference between case (2), i.e., adaptation using the 10 sentences of the target speaker's speech, and case (3). Therefore, it can be concluded that the adaptation from normal speech model to the general whispered speech is feasible. However, the performance of the adapted model is still much lower than that of the whispered speech model. The error of the adapted models is similar to that of the fully trained models.

5. SUMMARY

In this paper, we reported on the spectral characteristics and the recognition method of the whispered speech.

The comparison between whispered and normal utterances show that 1) the cepstrum distance between them is 4dB for voiced and 2dB for unvoiced phonemes, respectively, 2) the spectral tilt of the whispered speech is less sloped than the normal speech and 3) the frequency of the lower formants (less than 1.5kHz) lowered. Acoustic models (HMM) of whispered speech, trained by the database attain 60% accuracy in syllable recognition experiments. The accuracy is improved by 3% when MLLR adaptation is applied, while the normal speech HMM trained by the whispered speech attain only 56 % of the syllable accuracy score.

Further evaluation of the performance under real environment conditions, where the whispered speech captured at lower SNR, is indispensable for real applications.

6. REFERENCES

[1] W. Meyer-Eppler, "Realisation of prosodic features in whispered speech," J. Acoust. Soc. Am., 29, pp.104-106 (1957)

[2] I.B.Thomas, "Perceived pitch of whispered vowels," J. Acoust. Soc. Am., 46, pp. 468-470 (1969)

[3] Holmes J.N., and A.P. Stephens, "Acoustic correlates of intonation in whispered speech," J. Acoust. Soc. Am., 73, S87. (1983)

[4] I. Eklund and H. Traunmuller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," . Phonetica 54, pp.1-21 (1996)

[5] Ken J.Kallail, Floyd W.Emanuel "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects," J. Speech and Hearing Research 27, pp. 245-251 (1984)

[6] T.Kawahara, T.Kobayashi et al., "Japanese Dictation Toolkit: Plug-and-play Framework For Speech Recognition R&D," Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99), pp.393–396 (1999)

[7] K.Itou, K.Takeda et al., "Design and development of Japanese speech corpus for large vocabulary continuous speech recognition," Proc. of Oriental COCOSDA (May, 1998, Tsukuba)

[8] C.J.Leggetter and P.C.Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," Proc. of the ARPA Spoken Language Technology Workshop, 1995, Barton Creek