ACOUSTIC CHARACTERISTICS RELATED TO THE PERCEPTUAL PITCH IN WHISPERED VOWELS

Hideaki Konno[1,2]*, Hideo Kanemitsu*[1]*, Nobuyuki Takahashi*[1]*, Mineichi Kudo*[2]

[1] Department of Humanities and Regional Sciences, Hokkaido University of Education, Hakodate, Japan
[2] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

ABSTRACT

The characteristics of whispered speech are not well known. The most remarkable difference from ordinal speech is the pitch (the height of speech), since whispered speech has no fundamental frequency. In this study, we have tried to reveal the mechanism of producing pitch in whispered speech through an experiment in which a male and a female subjects uttered Japanese whispered vowels in a way so as to tune their pitch to the guidance tone with different five to nine frequencies. We applied multivariate analysis such as the principal component analysis to the data in order to make clear which part of frequency contributes much to the change of pitch. We have succeeded in endorsing the previous observations, i.e. shift of formants is dominant, with more detailed numerical evidence. In addition, we obtained some implications to approach the pitch mechanism of whispered speech. The main result obtained is that two or three formants of less than 5 kHz are shifted upward and the energy is increased in high frequency region over 5 kHz.

Index Terms— whispered vowel, pitch, formant, mel scale filterbank, principal component analysis

1. INTRODUCTION

According to the increase of mobile phones equipped with a speech recognition function, the concern of whispered speech recognition has been also increasing, because of the viewpoint of privacy preservation and avoidance of disturbance to others. However, whispered speech has totally different characteristics from ordinary speech, so that conventional recognition methods cannot be directly applied to it without modification. The most remarkable difference is seen in the pitch which is a perceptual height of speech. The pitch is governed by the fundamental frequency in normally phonated speech, while whispered speech is unvoiced and there is no fundamental frequency. Nevertheless, we can surely feel the pitch even in the whispered speech. This interesting fact derives us to this study.

Previous studies suggested that the pitch of whispered speech is controlled by shifting amount of the formant frequencies [1, 2, 3, 4], or by increasing the intensity of speech [1]. Perceptual experiments using synthesized vowels also made clear that whispered pitch is affected by formant frequencies [5, 6] and spectral tilt [5]. However, the degree and the relationship among them are not fully revealed yet.

In the experiment described in this paper, we present some pieces of evidence to explain the cause of perceptual pitch in whispered speech. We analyze the spectrum of Japanese five vowels uttered in such a way that each uttered vowel is tuned to a guidance tone of a fixed frequency.

2. RECORDINGS AND SPEECH SAMPLES

In this experiment, a male and a female subjects aged around 22 years uttered five Japanese whispered vowels of /i/, /e/, /a/, /o/ and /u/ in different levels of pitch. Their speech was recorded by a condenser microphone (SONY C-48) in a soundproof booth. The other pieces of equipment are a microphone amplifier (AVARON DESIGN M5), a digital audio tape deck (SONY DTC-2000ES) and a digital memory recorder (FOSTEX FR-2). Sampling rate and quantization is 48 kHz and 16 bits, respectively.

At the beginning, the speaker was asked to utter the lowest whispered vowels as he/she could and to search the frequency by tuning a dial of a pure tone generator and by listening to the generated tone through a headphone. Then the frequency of the guidance tone is raised in a half-tone step, e.g., C, C#, D, ..., and they uttered so as to match the guidance tone. In this way, they spoke 5 to 9 whispered vowels with different level of pitch in the range of 90 Hz to 160 Hz by the male speaker and of 400 Hz to 600 Hz by the female speaker.

As a post-processing, we cut out a part of 500 ms from the recorded signals such that the resultant part includes a target vowel only and then applied a high-pass filter with cut-off frequency of 50 Hz. The filter was applied in order to suppress a noise caused by a strong breath. The energy of cut signals was normalized to a constant before analysis. To preserve the original characteristics of the spectrum, no pre-emphasis was applied.

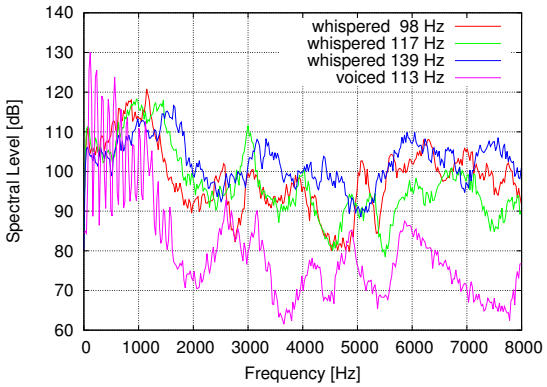

Fig. 1. Voiced and whispered vowels of /a/ by a male speaker.

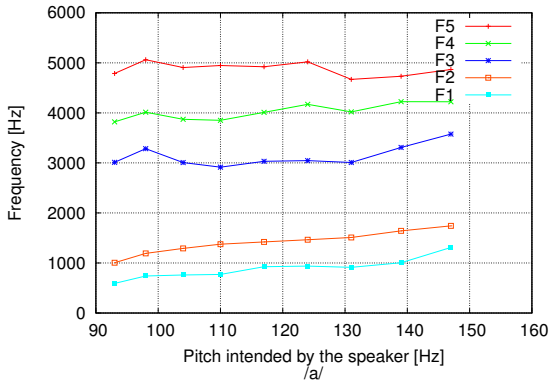

Fig. 2. Formant frequencies of /a/. The x-axis shows the pitch level intended by a male speaker.

3. ACOUSTIC ANALYSIS

3.1. Difference in spectrum

The speech samples were transformed to spectra in the frequency domain by FFT at 48 kHz sampling rate with shifting windows of length 42.7 ms (2048 points). The window is the Hamming window and the shift length is a half of the window length. The spectra collected by moving windows over a sample were averaged with respect to the same speaker, same vowel, and same level of pitch. An example of the averaged spectrum of whispered /a/ uttered by a male speaker is shown in Fig. 1 with that of normally phonated (voiced) /a/ by the same speaker as a reference. We see a clear difference between whispered and voiced vowels. There exists a clear harmonic structure less than 1 kHz in the voiced vowel but no such structure is seen in the whispered vowel. On the contrary, the energy of high-frequency part of whispered vowel is larger than that of voiced vowel. In addition, we can observe some characteristics in the whispered vowel according to the increase of level of pitch: 1) formants (spectral peaks) with frequencies less than 2 kHz shifted upward and 2) the tilt of the spectrum becomes more flat. The shift of formants is the observations already known in previous studies [1, 2]. We investigate these observations quantitatively in the following sub-sections.

3.2. Difference in formants

To make clear the amount of shift of formants, we applied the linear predictive coding (LPC) analysis to the speech samples. The samples were down-sampled to 12 kHz and LPC analysis with the order of 14 through 17 was applied. This time, the frame length is 256 points and the frame-shift length is 128 points. The five formants were calculated from the roots of LPC equation, taking into consideration the pole bandwidth and the continuity between them. The results are shown in Fig. 2. This figure shows the case of /a/, but the same tendency is observed in the other vowels. It becomes clear that the first two to three formants less than 5 kHz moves upward as the pitch level increases. For other vowels, see the appendix.

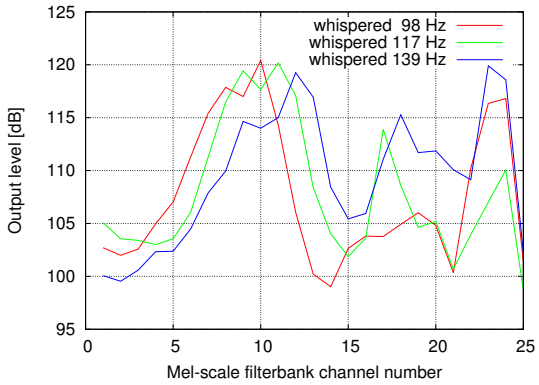

Fig. 3. Output of filter bank of /a/ uttered by a male speaker.

3.3. Mel scale filterbank output

To analyze the difference in spectra in a global manner, we examined the outputs of a filter bank equally arranged along the mel scale. The samples were down-sampled to 16 kHz. The number of channels of the filter bank was 25. The software package of HTK was used for the analysis. We took an average for the output of each filterbank channel. An example of the results is shown in Fig. 3. From this figure, we see the increase of intensity of region #15–#20 (2.3 kHz–4.2 kHz) as the pitch level increases. To evaluate this tendency analytically we applied the principal component analysis (PCA) to the averaged outputs of the mel-scale filterbank. PCA was carried out on all the outputs of all vowels and all levels of pitch in each speaker.

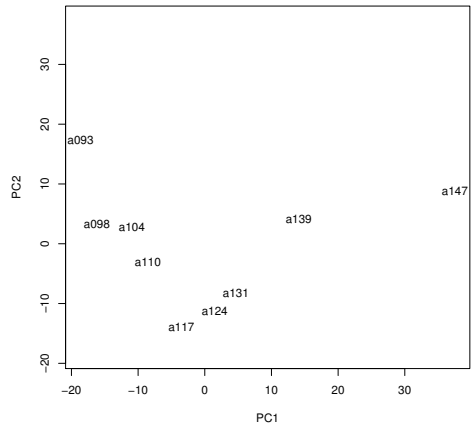

Fig. 4. Configuration of /a/ uttered by a male speaker in the plane of the first two principal components. A label along each point shows the vowel and its pitch frequency.

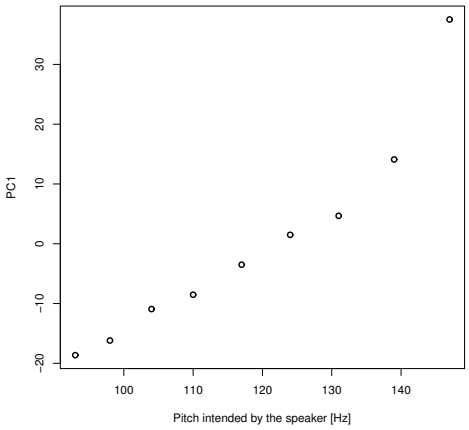

Fig. 5. The first principal component values of each sample of /a/ uttered by a male speaker.

A few principal components were sufficient in the sense of squared error to explain the data expressed in vectors of the outputs of the filterbank channels. In /a/ of the male speaker, for example, the first five of them contributed for reconstruction by 64%, 85%, 91%, 95%, 97%, respectively, when they are combined in turn. As an example, all /a/'s uttered by the male speaker are plotted in the plane spanned by the first two principal components in Fig. 4, where nine samples of /a/ with different nine levels of pitch are displayed, where the number attached to a point indicates the intended pitch frequency. It can be observed that the first component corresponds to the height of pitch faithfully. Indeed, an almost linear relationship can be seen between the pitch frequency and the value of first principal component in Fig. 5. The corresponding first eigenvector is shown in Fig. 6. This eigenvector shows that the components lower than #11 (1.3 kHz) work negatively to the increase of pitch frequency, while those components between #11 and #22 work positively. This eigenvector suggests that if we want to increase the pitch level, decrease the energy in low frequencies under 1 kHz and increase the formant frequencies between 1 kHz to 5 kHz upward. Indeed, through the eigenvectors, the decrease of energy below 550 Hz is observed in common to all whispered vowels when the speaker intended to increase the level of pitch.

To examine how well the pitch level can be predicted from a small number of principal components, we carried out a multivariate linear regression of the pitch frequency by the three principal components. An example of the results is shown in Fig. 7. The average p-value over all vowels was 4.9%. This means by three major factors we can explain the reasons why the pitch was increased or decreased in whispered speech. The most influential factor is a combination of

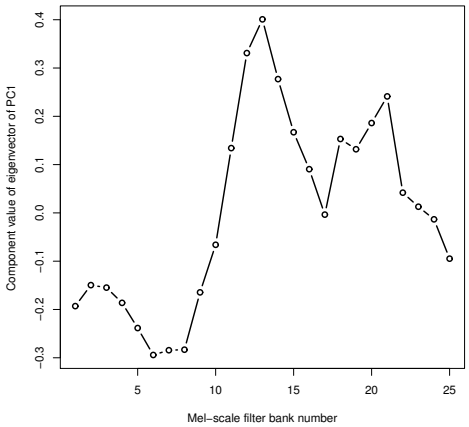

Fig. 6. Eigenvector of the first principal component for /a/ uttered by a male speaker.

energy in a low frequency part and formant shift in a middle frequency part as described in the case of the first eigenvector.

4. DISCUSSION

There are two cues known to explain the perception of pitch. They are the place cue and the temporal cue [7]. The place cue corresponds to the characteristic frequency of auditory nerves and is applicable for all sounds in the audible frequency range. The temporal cue is obtained by temporal intervals of impulses in neural activities and is said to be effective for the frequency range less than 5 kHz where the phase-locking can

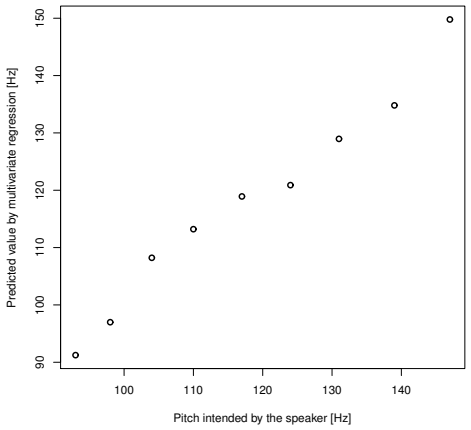

Fig. 7. Results of a multivariate linear regression.

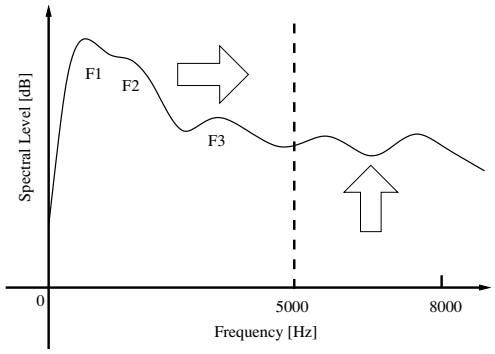

Fig. 8. Illustrative explanation.

occur. Our results might be related to this phase-locking. One possible explanation on the basis of phase-locking is as follows. Even in a whispered waveform, there exists a periodicity corresponding to formant frequencies, although the periodicity is not clear. If we assume that both of the place and temporal cues could influence on perceiving pitch of whispers, the region of frequency less than 5 kHz should be impact more than the region of frequency more than 5 kHz on the perception of pitch in whispered speech. Hence, increasing the pitch of whispered speech may cause upward shift of the formants less than 5 kHz firstly and/or enhancement of the power of high frequency range more than 5 kHz.

5. CONCLUSION

We have tried to reveal the mechanism of whispered speech to bring the perceptual pitch in spite of lack of the fundamental frequency. We designed an experiment so as to obtain Japanese whispered vowels with different levels of pitch and analyzed the difference from ordinary speech in spectrum. The results showed that producing higher pitch is made in a combinatorial way of 1) shifting upward the first two or three formants of less than 5 kHz and 2) increasing the energy in high-frequency region beyond 5 kHz (Fig. 8). These are confirmed by principal component analysis and multivariate linear regression. This might be useful to develop whispered speech interfaces. For the time being, we have only a few pieces of possibility to explain the reason why 5 kHz is the border, but it would be revealed by our future studies.

6. ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 23500340, and a presidential grant of Hokkaido University of Education.

7. REFERENCES

[1] W. Meyer-Eppler, "Realization of prosodic features in whispered vowels," *J. Acoust. Soc. Am.*, vol. 29, no. 1, pp. 104–106, 1957.

[2] T. Hirahara, "Acoustic analysis of whispered vowels in different notes," Tech. Rep. TR-A-0120, ATR, 1991, pp. 1–28 (in japanese).

[3] X. Li and B. Xu, "Formant comparison between chinese whispered and voiced vowels," in *Proc. ICA2004*, 2004, pp. 3291–3294.

[4] X. Chen and H. Zhao, "Relationship between fundamental and formant frequency in whispered mandarin," in *Proc. ICALIP2008*, 2008, pp. 303–306.

[5] H. Konno, J. Toyama, M. Shimbo, and K. Murata, "The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality," Tech. Rep. SP95-140, IEICE, 1996, pp. 39–45 (in Japanese).

[6] M. Higashikawa and F.D. Minifie, "Acoustical-perceptual correlates of whispered pitch in synthetically generated vowels," *J. Speech, Lang. and Hear. Res.*, vol. 42, no. 3, pp. 583–591, 1999.

[7] J.O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, 1988.

A. APPENDIX

All the other graphs of whispered vowels corresponding to Fig. 2 are shown in Fig. 9 and Fig. 10. In Fig. 9, formant frequencies are not plotted if formants disappeared.

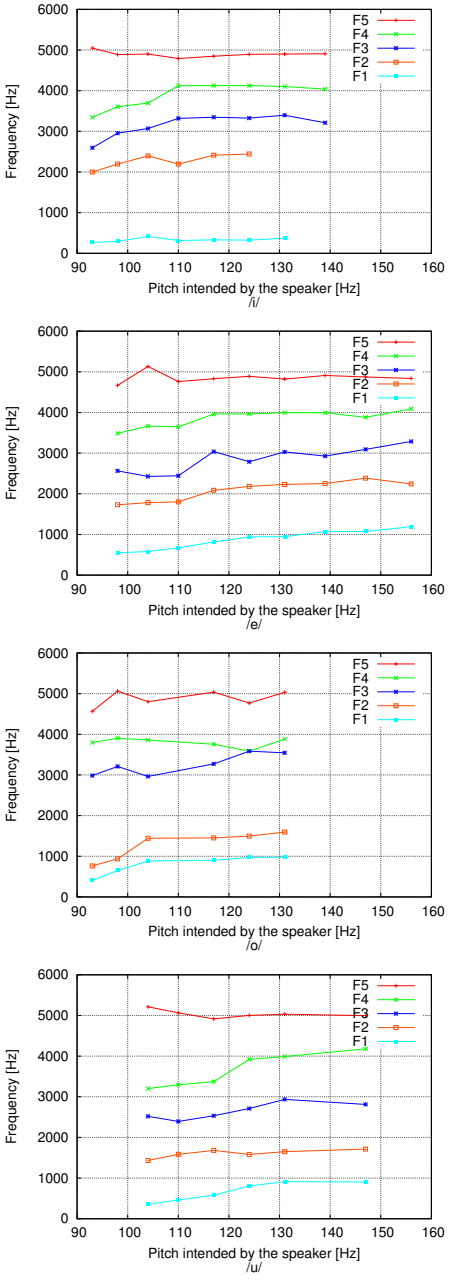

Fig. 9. Formant frequencies of the utterances of a male speaker.

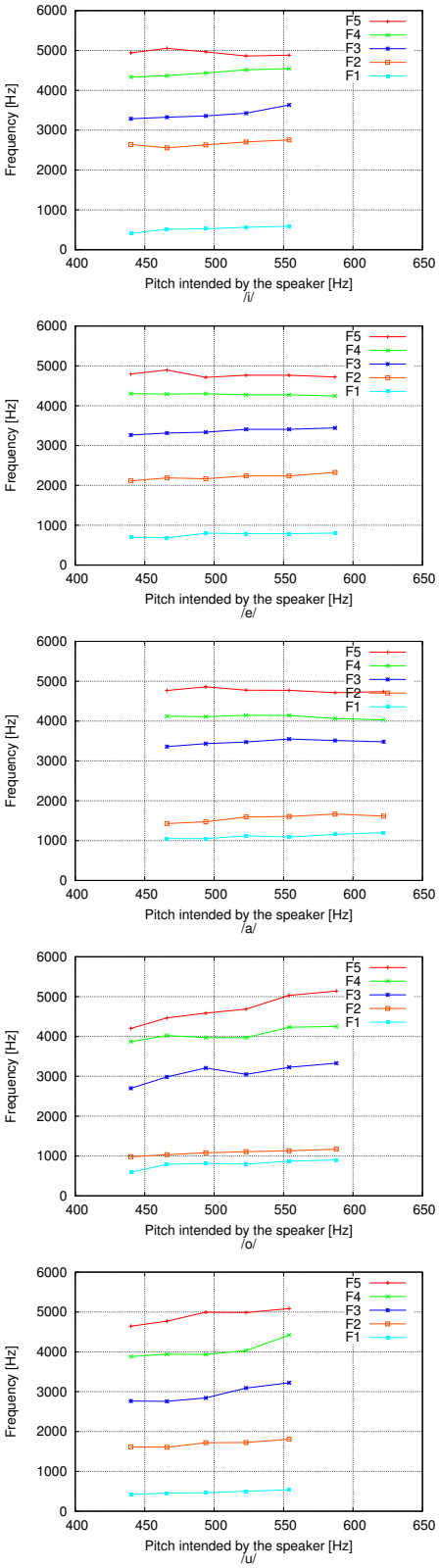

Fig. 10. Formant frequencies of the utterances of a female speaker.