A practical method of generating whisper voice: Development of phantom silhouette method and its improvement

Teruhisa Uchida1,* and Masanori Morise2,†

1Research Division, National Center for University Entrance Examinations, 2–19–23 Komaba, Meguro-ku, Tokyo, 153–8501 Japan
2School of Interdisciplinary Mathematical Sciences, Meiji University, 4–21–1 Nakano, Nakano-ku, Tokyo, 164–8525 Japan

(Received 2 December 2020, Accepted for publication 3 March 2021)

Keywords: Noise-vocoded voice, Spectral envelope, Speech synthesis, Voice conversion, WORLD

1. Introduction

The purpose of this research is to expand “expression via voice” more richly and freely. Our immediate aim is to convert standard utterances into whisper voices. If we can do that, we can later convert any part of a dialog or singing voice into a special effect. So far, we have proposed the phantom silhouette method as a practical method for converting standard speech into a whisper voice [1]. To that end, we use a high-quality vocoder, WORLD, which makes it easy to intuitively understand which parts of the speech sound are manipulated [2]. We have now improved that method. This paper describes the improvements and the results of the evaluation experiment.

2. Phantom silhouette method

With the phantom silhouette method, a pseudo-whisper is generated by first extracting the spectral envelope of the standard speech using WORLD. The envelope is then transformed so that it sounds like a whisper. Finally, a devoiced pseudo-whisper is created by driving the manipulated spectral envelope using white noise instead of the vocal cord sound source signal [3].

The spectral envelope is converted by using three operations to manipulate the timbre:

1. Upward shifting of spectrum in \( F_1 - F_2 \) formant frequency bands,
2. Compensation of the breathy sound component in the high-frequency range,

This timbre manipulation is not signal processing performed after voice conversion. Since the spectrum is directly manipulated, the desired timbre can be obtained when synthesizing speech.

The conversion process can be likened to the image of a Halloween ghost. The core of this method is noise-driven devoicing and low-frequency suppression of the spectrum, which is referred to as “phantomization of voice.” Pseudo-whispering generated using only this process is called “truncated noise-vocoded speech” and is used in auditory experiments. The procedure for adding the characteristics of whispering to the spectrum of standard speech is called “spectral silhouette compensation.” We have now improved each of these processes.

3. Improvement of conversion processes

3.1. Upshifting of \( F_1 - F_2 \) frequency bands

Matsuda et al. [4] reported that formants below 1,200 Hz rise when whispering. Therefore, in the conversion method, the standard speech’s spectral frequency axis is partially expanded or contracted on the equivalent-rectangular-bandwidth (ERB) scale. However, in the previous conversion method, because we were particular about fixing the origin \((0,0)\), the shift amount was insufficient, especially in the low-frequency range of male voices.

In contrast, when we analyzed the female whisper voice, the formants did not rise as much as the male ones. Therefore, we changed the shift amount to be controlled in accordance with the median value of \( f_0 \) so that the shift amount is large for low-\( f_0 \) male voices and small for high-\( f_0 \) female voices (Fig. 1).

3.2. Compensation of high-frequency range

Compared with that of actual whispering, the spectrum of standard speech lacks high-frequency components. Therefore, the high-frequency spectrum of standard utterances was compensated using a certain weight in the previous conversion method. This time, we have adjusted this compensation amount. Figure 2 shows the time-averaged spectrum of voices recorded by the same speaker in the narrator’s office, who uttered the same sentence in standard and whisper voices. The spectral gradient difference in the high-frequency range (1.6 kHz–10 kHz) can be considered the amount of compensation required when whispering.

The differences in the spectral gradient in comparison with the \( f_0 \) median for male and female voices are shown in Fig. 3. It is evident that more high-frequency compensation was required for the male whisper voices. Furthermore, the higher the \( f_0 \), the more compensation is required for both men and women. Therefore, after classifying male and female voices on the basis of the \( f_0 \) median, we corrected the weighting of the high-frequency emphasis in accordance with the \( f_0 \) value (Fig. 4).

3.3. Low-frequency spectrum suppression

Matsuda et al. [4] pointed out that, in a whisper, the sound pressure level drops in the range where the frequency is below
1 kHz, while Kishida et al. [5] found that amplitude information from 570 Hz to 1,370 Hz plays an essential role in phonological comprehension in noise-vocoded speech. Therefore, paying attention to phonological comprehension, we set the transition range and suppressed the low-frequency range (Fig. 5). We also reexamined the target to which this process is applied.

We can distinguish standard speech and actual whisper by using the value of the spectral centroid (e.g., 1.6 kHz). Standard speech is converted into pseudo-whisper by combining compensation of the spectral silhouette and phantomization of voice. If a whisper is input, only the phantomization process is applied since the input is already a whispered voice, which means we do not have to compensate for the spectral silhouette. The phantomization process is applied to deal with quality deterioration in the vocoder due to local voiced sounds.

4. Evaluation of pseudo-whisper

4.1. Experimental design

The improved version of the phantom silhouette method is called “Phantom Silhouette 2” (“P.S.2”), Figure 6 shows the waveform and spectrogram of the speech data used in the
sound quality evaluation experiment. The leftmost panel shows standard speech. The next panel shows a pseudo-whisper converted from the standard speech using P.S.2. The next panel shows a phantomized voice converted by P.S.2 from an actual whisper. The next panel is a voice resynthesized simply by WORLD from the same actual whisper. The last panel is the actual natural whisper.

The original speech data were eight sentences (S1–S8) from the ATR phoneme balanced sentences and standard speech and whisper voice samples from four male (M1–M4) and four female (W1–W4) professional narrators.

Experimental Condition 1 was a pair of a voice resynthesized by WORLD from an actual whisper and a voice phantomized by P.S.2 from the same whisper voice. Condition 2 was a pair of a voice resynthesized by WORLD from an actual whisper and a pseudo-whisper converted by P.S.2 from “standard speech.” Condition 3 was a pair of an actual whisper and a pseudo-whisper generated by P.S.2. Condition 4 was a pair of two pseudo-whisper generated by P.S.2, with white noise and modified velvet noise as the drive sound source. Since the adoption of modified velvet noise is currently being considered for WORLD [6], we also investigated its effect on the phantom silhouette method.

4.2. Procedure

The participants in the experiment were 123 first-year students from 10 national universities in the Tokyo metropolitan area. They were divided into eight groups, and voice comparison pairs were also divided into eight test sets.

For the evaluation items, we used 20 items related to voice quality expressions, including the evaluation of whisper and the naturalness of the voice [1,7]. The voice’s naturalness was evaluated using four words: natural, understandable, clear, and audible.

We presented the speech stimuli from a portable CD player via earphones. Two speech stimuli of the comparison pair were presented once. The participants were asked to evaluate which voice stimulus was more relevant and appropriate to the item words in ten stages. The rating time for 20 items per pair was 60 s.

4.3. Result

The ten-stage rating was converted into a number (from −4.5 to 4.5) for each evaluation item. A positive number means that the voice generated by P.S.2 was more conformable for that item. The participants’ average score for each item for each comparison pair was used as the evaluation index. Then, each comparison pair was analyzed based on the results of a two-sided t-test with $\mu = 0$ as the null hypothesis $H_0$.

Figure 7 shows the results by the experimental condition of a paired comparison in which the evaluation item was “more like a whisper.” The voice phantomized by P.S.2 from
an actual whisper was evaluated as more like a whisper than the voice resynthesized by WORLD from an actual whisper ($t_{15} = 4.38, p < 0.001$). The pseudo-whisper converted by P.S.2 from the standard speech was evaluated as more like a whisper than the voice resynthesized by WORLD from the actual whisper ($t_{15} = 3.76, p < 0.01$). Furthermore, for a male voice, the pseudo-whisper generated by P.S.2 was evaluated as more like a whisper than the actual male whisper ($t_{7} = 2.09, p < 0.10$). However, there was no significant difference for the female voice ($t_{7} = -1.31, n.s.$). In terms of “whispering,” the quality was reasonable. No significant effect was observed depending on the type of drive noise ($t_{15} = -0.71, n.s.$).

Figure 8 shows the results by the experimental condition in which the evaluation score was “naturalness of voice.” The voice phonemized by P.S.2 from the actual whisper was evaluated as “more natural” than the voice resynthesized by WORLD ($t_{15} = 1.90, p < 0.10$). Then, there was no significant difference between the pseudo-whisper by P.S.2 from standard speech and the voice resynthesized by WORLD from the actual whisper ($t_{15} = 1.44, n.s.$). However, when comparing the actual whisper with the pseudo-whisper generated by P.S.2, the real whisper was evaluated as more natural ($t_{15} = -7.66, p < 0.001$). No significant effect of the drive noise type was found ($t_{15} = -0.28, n.s.$).

5. Conclusions

Although WORLD is a predecessor of the phantom silhouette method, pseudo-whisper generated by the improved version of the phantom silhouette method (Phantom Silhouette 2) was evaluated as “more whispering-like” than the voice resynthesized by WORLD from actual whispers. In addition, they were more than comparable in quality to the actual whispers, evidence that “whispering” was well realized. Nevertheless, improvement is still needed in terms of naturalness. Since the noise type of the driving sound source did not affect the quality, the phantom silhouette method can continue to be used even if modified velvet noise is adopted for WORLD. However, further improvements are needed regarding the effects of the speech organs.

The advantage of the parametric conversion used in the phantom silhouette method is that it is clear which acoustic attributes are being manipulated. So, this conversion method is applicable to auditory research, and the creation of teaching materials for acoustic education [8–10].

Acknowledgement

This work was supported by JST PRESTO Grant Number JPMJPR18J8, JSPS KAKENHI Grant Number JP19K12736.

References