Acoust. Sci. & Tech. **44**, 3 (2023)

PAPER

A practical method for generating whispers from singing voices: Application of improved phantom silhouette method

Teruhisa Uchida*

Research Division, National Center for University Entrance Examinations, 2–19–23 Komaba, Meguro-ku, Tokyo, 153–8501 Japan

(*Received 30 May 2022, Accepted for publication 4 January 2023*)

Abstract: The previously proposed phantom silhouette method is promising for converting ordinary speech into whispered speech. It is a simple parametric method that uses high-quality vocoder-type speech analysis and synthesis. An ordinary speech sample is first analyzed using the WORLD vocoder. Then, based on the extracted spectral envelope, spectral features are manipulated so that the voice sounds like a whisper. The target speech is synthesized by driving it with white noise instead of the vocal source signal to make the whole speech sound voiceless. In this study, this method was applied to singing voices to generate whisper voices. In addition to actual singing voices, virtual singers' voices were generated using a Vocaloid voice synthesizer, and AI singers' voices synthesized using a NEUTRINO neural singing synthesizer were also tested to generate whisper voices from singing voices.

Keywords: Noise-vocoded speech, Spectral envelopment, Voice conversion, Speech synthesis, WORLD vocoder

1. INTRODUCTION

1.1. Whispered Speech Generation from Ordinary Speech

The previously proposed phantom silhouette method is a promising approach to converting ordinary speech into whispered speech [1]. It is a simple parametric method using a high-quality traditional WORLD vocoder [2,3].

The WORLD speech analysis/synthesis system has been used in various applications, such as voice conversion and statistical parametric speech synthesis. It first accurately decomposes the speech waveform into the fundamental frequency (f_o), spectral envelope, and aperiodicity. These elements can then be transformed to meet the user's needs. Next, it synthesizes new voices by integrating the transformed f_o, spectral envelope, and aperiodicity. To that end, the phantom silhouette method makes it easy to understand which parts of the speech sound can be manipulated intuitively.

This "whisper conversion" method can be used for already recorded speech, so it can also provide special voice effects. For example, it has been used to create a demonstration movie for auditory illusions as acoustical teaching material [4,5].

*e-mail: uchida@rd.dnc.ac.jp
[doi:10.1250/ast.44.239]

1.2. Generating Whispers from Singing Voices

One possible application of this "whisper conversion" method is to convert singing voices into whispers. As in the case of monologues, if recorded singing voices can be subsequently changed into whispers, the freedom of singing voice expression can be further expanded. This paper reports on the conversion of not only ordinary speech but also "singing voices" into "whisper voices" using an improved version of the phantom silhouette method.

1.3. Converting Virtual/AI Singers' Voices into Whispered Voices

In terms of applicability, the ability to convert virtual singers' voices generated using such tools as Vocaloid, Sinsy, and UTAU and AI singers' voices generated using such tools as CeVIO AI and NEUTRINO into whisper voices would further expand the expressiveness of singing.

When using singing voice synthesis software, information corresponding to the lyric text and musical notation is input, and singing voices are synthesized and output. In this process, the singing voice usually cannot be satisfactorily produced with the system output as is, so fine-tuning the pitch and timing of the singing voice, known as "singing voice adjustment," is currently required.

Given that enhancing the expressiveness of the singing voice is already recognized as an essential process, there should be little resistance to using new modules that can

add additional expressiveness. If users find value in creating their own singing voice through trial and error, the whispering module will be helpful in the context of singing voice synthesis.

1.4. Singing Voice and Phantom Silhouette Method

The phantom silhouette method was initially developed to convert ordinary speech into a whisper voice. The transition range of the fundamental frequency (f_o) of the singing voice differs greatly from ordinary speech. In addition, features such as "singer's formant" may appear in the spectral envelope, which is different from that in ordinary speech. Therefore, actual human singing voices must first be converted into whispers, which are then listened to.

In addition to the actual singing voices, this study tested the generation of synthetic singing voices using singing voice synthesizing software for conversion into whisper voices. More specifically, attempts were made to generate whisper voices using an improved phantom silhouette method from both virtual singers' voices synthesized using Vocaloid voice synthesis software [6] and AI singers' voices synthesized using a NEUTRINO neural singing synthesizer [7].

2. ORIGINAL PHANTOM SILHOUETTE METHOD

In the phantom silhouette method [1], the spectral envelope of an ordinary speech sample is first extracted using the WORLD vocoder. The envelope is then transformed so that the speech sounds like a whisper. Finally, a wholly devoiced pseudowhisper is created by driving the manipulated spectral envelope using white noise instead of the vocal cord sound source signal [8]. The spectral envelope is converted using the following three operations to manipulate the timbre

(1) Upward shifting of spectrum in F_1 and F_2 formant frequency bands (F_1–F_2 frequency band)
(2) Compensation of breathy sound component in the high-frequency range
(3) Suppression of low-frequency band in the spectral envelope

The core of this method is noise-driven devoicing transformation and low-frequency suppression of the spectrum, which are referred to as "phantomization of ordinary voice."

The procedure for adding the characteristics of a whisper voice to the spectrum of ordinary speech is called "spectral silhouette compensation." More specifically, it is upward shifting of the F_1–F_2 frequency bands and compensating for the high-frequency components.

2.1. Upshifting of F_1–F_2 Frequency Bands

In the case of whisper speech, the formants below 1,200 Hz are elevated compared with those of ordinary speech [9]. Therefore, in "whisper conversion," the ordinary speech spectral frequency axis is partially expanded or contracted on the equivalent-rectangular-bandwidth (ERB_{RATE}) scale, which corresponds to the critical bandwidth of hearing.

A certain amount of shift is required for male voices. Nevertheless, the increase in F_1–F_2 band is not as large for female whisper voices as for male whisper voices. Therefore, the amount of shift is controlled by the median value of f_o so that it is large for low-f_o male voices and small for high-f_o female voices.

2.2. Compensation for High-frequency Components

Ordinary speech may lack high-frequency components compared with actual whisper speech. The phantom silhouette method compensates for high-frequency components when generating whisper voices by manipulating the spectral gradients in the high-frequency range (1.6 kHz to 10 kHz).

In general, more high-frequency compensation is required for male whisper voices. Furthermore, the higher the f_o, the more compensation is required for both male and female voices. Therefore, after male and female voices are classified based on the median f_o, the weighting of high-frequency emphasis is corrected following the f_o value [1].

2.3. Low-frequency Spectrum Suppression

In whispered voices, the sound pressure level drops in the range of frequency below 1 kHz [9]. Since amplitude information from 570 Hz to 1,370 Hz plays an essential role in phonological comprehension in noise-vocoded speech [10], the transition range was set, and the low-frequency range was suppressed while paying attention to phonological comprehension.

2.4. Generation of Pseudowhisper Voice

After the three spectral transformations described above, the WORLD vocoder uses the transformed spectrum as the basis for synthesizing a pseudo-whispering voice by driving it with white noise.

3. APPLICATION OF IMPROVED PHANTOM SILHOUETTE METHOD TO SINGING VOICE

3.1. Specification Changes

In the original phantom silhouette method, the amount of F_1–F_2 shift and high-frequency compensation usually are adjusted following the median value of f_o of the original voice. However, in a singing voice, the behavior

of f_0 is defined by the song's melody. As a result, the distinction between male and female voices, the shift setting, and compensation amounts may not work well.

To address this problem, the specifications of the phantom silhouette method were changed, and all parameter adjustments were opened to the user. The initial values of both the amount of F_1–F_2 shift and high-frequency compensation were set at temporary values, and fine-tuning was left to the user. The cut-off frequency of the spectral low-frequency suppression was also made adjustable. Users can thus freely manipulate the parameters while creating the desired voice tone.

A single-function module program was created to enable the independent use of various applications. It can be run by specifying a wave file and the parameters on a command line.

3.2. Application to Actual Human Singing Voices

3.2.1. Source material samples

Three male voice and three female voice samples from the JVS-MuSiC voice corpus [11] were used (High, Middle, Low-Key). The songs were nursery rhymes such as "Spring Creek (Haru no Ogawa)," as shown in Fig. 1. The original singing voices were sampled at 24 kHz, with 16-bit linear quantized speech data, and the pitch and tempo were not modified after the recording.

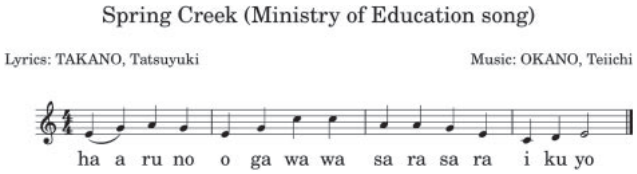


Fig. 1 An example song used to convert from an ordinary singing voice to a whisper voice: score and lyrics for first four bars of "Spring Creek (Haru no Ogawa)," a Ministry of Education song.

3.2.2. Conversion to whisper voices

The actual singing voice was input into the phantom silhouette module to convert it into a whisper voice. When the conversion parameters were appropriately adjusted to whisper voice, it was possible to obtain the same or better quality than with the original phantom silhouette method. The whisper voice generated from the actual singing voice is shown in Fig. 2.

When users utilize whispers in music, the original singing voice data and the whisper voice data are imported as separate tracks into a digital audio workstation, such as the Steinberg Cubase Pro (ver. 10.5.30). The original singing voices are then muted when using the whisper voices.

3.3. Application to Virtual Aingers' Voice

3.3.1. Source material samples

A Yamaha Vocaloid voice synthesizer (ver. 5.6.3), which is based on concatenative synthesis, was used to synthesize the singing voice samples using the Kaori and Ken voice libraries. Songs recorded in JVS-MuSiC and original passages were prepared and used. Singing voice data with 44.1 kHz sampling and 16-bit linear quantization were created using the Vocaloid Editor for Cubase (ver. 4.5).

A whisper voice has no fundamental frequency or overtone structure. Thus, at first glance, it would seem that the sample singing voice does not need a melody. However, in a synthetic singing voice, the timbre of the voice changes in accordance with the pitch of the note. This tendency directly affects the timbre of the whisper voice that is generated. Furthermore, during the production of the whisper part, it is necessary to listen to the song lyrics and adjust the rhythm and timing many times to match the accompaniment. Therefore, a provisional melody for the singing voice to be converted to a whisper voice must be prepared.

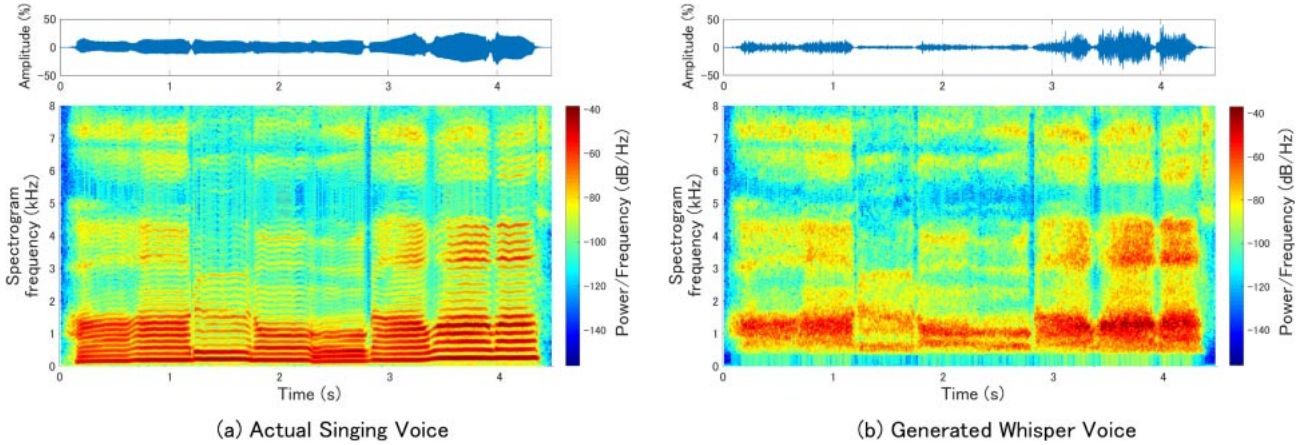


Fig. 2 Waveforms and spectrograms of original human singing voice and generated whisper voice [jvs089: male-voice (Haru no Ogawa wa)].

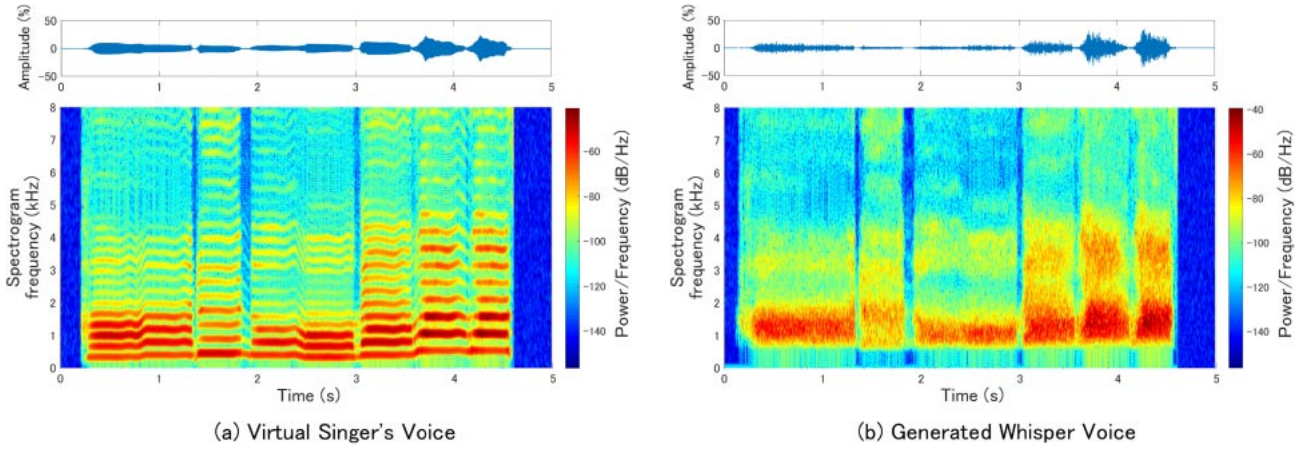


Fig. 3 Waveforms and spectrograms of virtual singer's voice synthesized using VOCALOID and generated whisper voice [Kaori: female-voice (Haru no Ogawa wa)].

3.3.2. Whisper voice generation from virtual singers' voice

The virtual singers' voices were input into the phantom silhouette module and converted into whisper voices. In the conversion, parameters such as a moderate amount of F_1–F_2 shift and high-frequency compensation value were first set. These parameters and the cut-off frequency for low-frequency suppression were adjusted to generate desirable whispers for users. Figure 3 shows the waveforms and spectrograms of a virtual singer's voice and the generated whisper voice.

The parameters were further adjusted during music creation to change the timbre so that the sound would neither be buried in the accompanying instrumental sound nor stand out too much.

3.4. Application to AI Singers' Voice

3.4.1. Synthesis of AI singers' voice samples

The NEUTRINO neural singing synthesizer (ver. 1.0.0), which is based on a neural source-filter (NSF) vocoder [12], a neural network, and the Merrow and Nakumo voice libraries were used to create the AI singers' voice samples. The same songs as described in the previous section were used.

MuseScore (ver. 3.6.2) was used to input lyrics and melody and to create a MusicXML file. The MusicXML-to-label program module was used to convert the MusicXML file for the song into a label format. The NEUTRINO neural singing synthesizer estimated speech parameters such as fundamental frequency, mel-cepstrum, and aperiodicity indices from the label file.

NEUTRINO used the NSF neural vocoder and the voice parameters to synthesize AI singers' voice samples with 48 kHz sampling and 16-bit linear quantization. The NVIDIA T600 desktop GPU graphics card was installed in the computer to accelerate computation.

3.4.2. Whisper voice generation from AI singers' singing voices

The AI singers' voices were input into the phantom silhouette module and converted into whisper voices. The amount of F_1–F_2 shift, high-frequency compensation value, and cut-off frequency for low-frequency suppression were adjusted to generate desirable whispers. Figure 4 shows the waveforms and spectrograms of the AI singers' voice and the generated whisper voice.

3.5. Comparison and Evaluation of Converted Whispers

The virtual singing voices generated using the Vocaloid voice synthesizer were generated by concatenative synthesis. The whisper voices generated from the synthesized singing voices exhibited characteristics different from the whisper voice converted from the actual human voice, as seen in Figs. 2 and 3. In the virtual singing voice (left side of Fig. 3), the pitch changes between the first and second beats (0.3–1.5 s), where the vowel sound continues, and the harmonic structure shifts upward concurrently. However, there are few formant transitions and amplitude changes in that interval. Thus, in the whisper voice (right side of Fig. 3), the same spectral envelope continues throughout both the first and second beats. Listening to the generated whisper voice reveals sounds resembling long connected tones. If it is crucial to distinguish the note at each beat, adding a short rest at the end of the first beat may be necessary instead of following the score strictly.

In the AI singing voice generated by NEUTRINO (left side of Fig. 4), the vocalization restarts from the beginning of the second beat. Therefore, the volume and formant transition change. In the whisper voice (right side of Fig. 4), spectral changes are evident after the start of the second beat. Listening to the generated whisper voice reveals a break at the second beat.

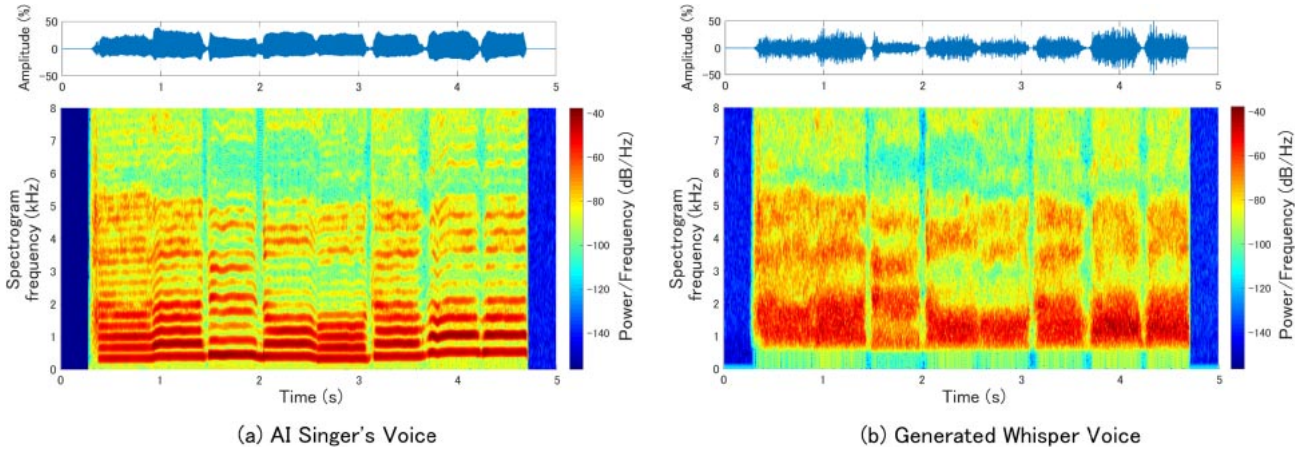


Fig. 4 Waveforms and spectrograms of AI singer's voice synthesized using NEUTRINO neural singing synthesizer and generated whisper voice [Merrow: female-voice (Haru no Ogawa wa)].

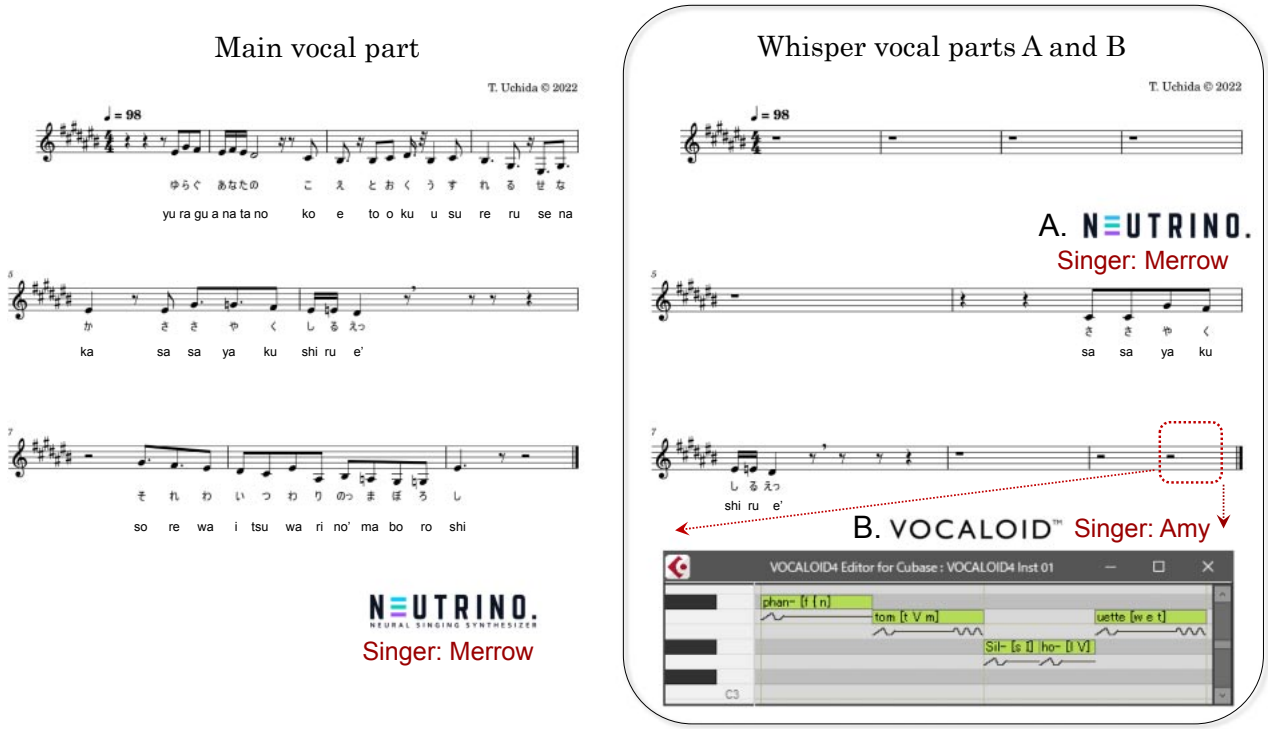


Fig. 5 Example of creating a song in which whisper voices are inserted between ordinary singing voices. Score on left is main vocal melody for ordinary singing voice. Score and piano roll on right are for whisper voice part A and whisper voice part B. Synthesized AI singer voices and virtual singer voices were used as samples.

The AI singing voices were synthesized using the NSF neural vocoder and a neural network. These singing voices were of higher quality than the virtual singing voices generated by concatenative synthesis. Thus, the more realistic the sample singing voice, the higher the quality of the generated whisper voice.

3.6. Example of Song Production Using Whispers Generated from Synthesized Singing Voices

3.6.1. Composition of the music and whispered parts

The following is an example of music production using a synthesized singing voice and a generated whisper voice. For this purpose, the author created a short piece of music consisting of lyrics, melody, and instrumental accompaniment. In the song, whisper parts were inserted between the ordinary singing parts. The scores are shown in Fig. 5. The score on the left is the main melody for an ordinary singing voice. The score and the piano roll on the right are for whisper parts A and B.

The main vocal is the voice of a Japanese female AI singer synthesized using NEUTRINO. Whisper part A is the same singer's voice converted into a whisper voice by

Table 1 Comparison of conversion parameters between previous and improved versions for whisper vocal part A (NEUTRINO: Japanese female "Merrow") and part B (Vocaloid: American female "Amy").

	WORLD: minimal pitch (Hz)	WORLD: maximal pitch (Hz)	F_1–F_2 formant shift amount (Magnification at 400 Hz)	Compensation ratio of high-frequency component	Cutoff frequency for suppression of low-frequency spectrum (Offset in Hz)
Part A: Previous Version: (phantom silhouette 2.0)	71.0	800.0	0.80	1.81	0.0
Part A: Improved Version: (phantom silhouette 3.1)	40.0	2,000.0	1.40	1.40	0.0
Part B: Previous Version: (phantom silhouette 2.0)	71.0	800.0	0.79	1.82	0.0
Part B: Improved Version: (phantom silhouette 3.1)	40.0	2,000.0	1.00	1.60	0.0

the phantom silhouette method. Whisper part B is based on the voice of an American female virtual singer synthesized using Vocaloid. It was also converted into a whisper voice.

The song was accompanied by an electronic piano (Modartt Pianoteq 7), an electric bass (WAVES Bass Fingers), and drums (Toontrack EZ DRUMMER 2). The timbre of the whispers converted from singing voices was adjusted many times to match the tone of the accompaniment.

3.6.2. Comparison of parameters in the previous version of phantom silhouette (PS-2) and the improved version of phantom silhouette (PS-3)

Here, we compare the parameters set in the previous version of the phantom silhouette (PS-2) and those in the improved version of the phantom silhouette (PS-3). The search range of f_o by WORLD, the amount of F_1–F_2 shift, the amount of high-frequency compensation, and the cutoff frequency for low-frequency suppression are shown in Table 1. There, we show the parameters for whispered voice part A generated NEUTRINO's AI singer and whispered voice part B generated from Vocaloid's virtual singer.

PS-2 was designed for ordinary speech, so the f_o search range was from 71 Hz (D_2) to 800 Hz (G_5). However, the f_o pattern of the singing voice is generally higher and the transition range is larger than those of the ordinary voice. Therefore, the search range was extended from 40 Hz (E_1) to 2,000 Hz (B_6) in the PS-3. Since the f_o of the singing voice is higher than that of the ordinary speech, the amount of F_1–F_2 shift is set excessively small in the conventional version of PS-2. In contrast, in the improved version of PS-3, a larger F_1–F_2 shift is set. On the other hand, the high-frequency compensation is set too high in the conventional PS2 because of the high f_o. In the improved PS-3, however, it is set more moderately. The cutoff frequency for low-frequency suppression, which has a significant impact on timbre, was not changed in this song.

Figure 6 shows the final 3D spectrogram of the singing voices. This example illustrates the potential use of generated whispers.

4. CONCLUSION AND DISCUSSION

4.1. Phantom Silhouette Method and Restrictions When Applying Singing Voice

4.1.1. Advantages of using singing voice as source material

A singing voice needs to convey the pitch transitions of the melody accurately. Therefore, the vocalization is performed so that the fundamental frequency transitions are transmitted clearly. Stable fundamental frequencies and spectral envelopes were obtained for the human singing voice, the Vocaloid virtual singing voice, and the NEUTRINO AI singing voice.

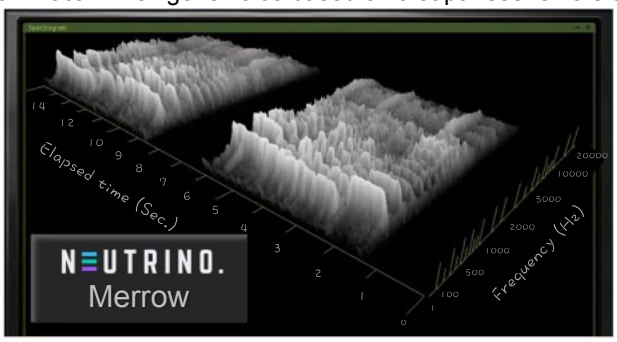
In this experiment, the maximum value of the fundamental frequency search range with the WORLD vocoder was changed from the maximal default value of 800 Hz to 2 kHz, which can occur in singing voices. As a result, the fundamental frequency extraction using WORLD was also stable, and reasonable estimates of the spectral envelope were obtained. The initially expected difference in the transition range of the fundamental frequency between the ordinary speech and the singing voice had little effect on the sound quality. Therefore, the whisper voices generated on the basis of the fundamental frequency were also of good quality.

For Mongolian khoomii singing and growl singing in metal music, it is assumed that the estimation of the fundamental frequency is difficult owing to the specifications of the unsuitability of the WORLD vocoder. However, from a practical application standpoint, such singing is quite unlikely to be converted into whisper voice.

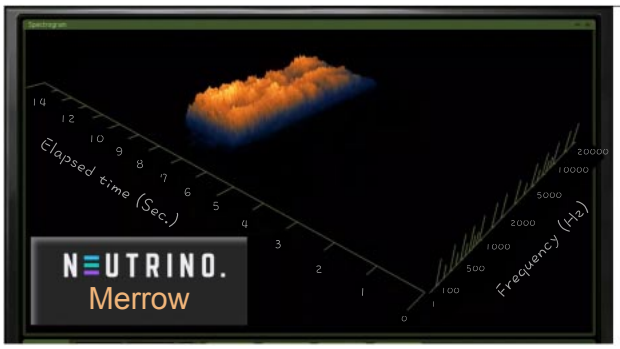
4.1.2. Restrictions due to singing voices

In the original phantom silhouette method, the parameters are adjusted by the median value of f_o, which is also

(a) Main vocal: AI singer's voice based on a Japanese female singer

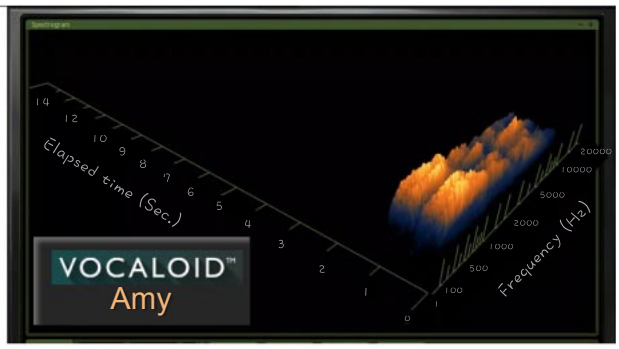


(b) Whisper part A: Japanese lyric "sasayaku shiruetto "



Converted from Japanese female AI singer's voice

(c) Whisper part B: English lyric "phantom silhouette"



Converted from American female virtual singer's voice

Fig. 6 Spectrograms of (a) main vocal, (b) whisper voice part A, and (c) whisper voice part B. Last five bars of score are shown in Fig. 5. Main vocal is voice of Japanese female AI singer synthesized using neural vocoder. Whisper part A is same AI singer's voice converted into whisper voice using phantom silhouette method. Whisper part B is based on voice of American female virtual singer synthesized using concatenative synthesis. It was converted into whisper voice.

used to distinguish male and female voices. However, in singing voices, the transition range of f_o is determined by the melody. Therefore, when the f_o median value is used to estimate parameters in songs, inappropriate parameter estimation and failure to distinguish between male and female voices often occur. This problem restricts the use of singing voices as the source material.

To overcome this restriction, the specifications of the phantom silhouette method were modified. In the improved method, all parameter adjustments are left to the user. Changing the parameter values can greatly change the timbre of the whispers. The user's preferred whispered timbre is obtained by changing the parameters little by little so the user is more actively involved in creating whisper voices. This openness gives the user more freedom to create target whisper voices.

4.2. Possibilities and Future Challenges

4.2.1. Charm of expression of imperfect voices

The processes of voice production in the vocal organs differ between ordinary speech and actual whispered speech. The whisper voices generated using the phantom silhouette method are a deformation of the ordinary speech voices and therefore are different in principle from actual whispered speech. It is thus vital to use the proposed method with the understanding that it is only an alternative method.

Similarly, the singing voices of virtual singers do not entirely imitate that of a human being. Instead, the artificial and mechanical impression they leave gives them their unique appeal. This contradiction means that an artificially colored whisper may be more acceptable than a whisper completely indistinguishable from a human voice. In other words, the fact that a quasi-whisper is different from a real whisper may actually be an advantage. In this case, a whispering module that is intuitively easy to understand and can be parametrically processed might be practical.

4.2.2. Symbiosis with neural vocoder using DNN

Research on deep-learning whisper voice generation based on deep neural networks (DNNs) has recently progressed [13]. It is expected that generating higher-quality whispers using a neural vocoder and other methods will become possible.

In the meantime, the phantom silhouette method, based on the WORLD vocoder, a traditional high-quality vocoder, can be applied to any source material, including human voices and virtual singers' voices generated by concatenative synthesis and AI singers' voices by using a neural

vocoder. This ease of use is an advantage. Moreover, this method has a light computational load, even in a computer environment without GPU support. In this AI singing voice experiment, singing voices synthesized using a neural vocoder were converted into whisper voices using the phantom silhouette method. This process was sequential in that singing voices synthesized using an NSF neural vocoder were analyzed, transformed, and resynthesized using the traditional high-quality WORLD vocoder. In other words, this is an example of cooperative use that takes advantage of these two methods.

The advantage of the parametric transformation of the phantom silhouette method is that it is clear which acoustic attributes are being manipulated, which is vital for applications in auditory research and for creating educational materials for speech education [14,15]. Future work includes investigating the applicability of this method while taking advantage of its merits.

REFERENCES

[1] T. Uchida and M. Morise, "A practical method of generating whisper voice: Development of phantom silhouette method and its improvement," *Acoust. Sci. & Tech.*, **42**, 214–217 (2021).

[2] M. Morise, F. Yokomori and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, **E99-D**, 1877–1884 (2016).

[3] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, **84**, 57–65 (2016).

[4] T. Uchida, "As we speak: Pure culture of the voice timbre," *Proc. 85th Annu. Conv. Jpn. Psychol. Assoc.*, p. 114 (2021) (in Japanese).

[5] A. Kitaoka, "The 10th visual illusion and auditory illusion contest in Japan," http://www.psy.ritsumei.ac.jp/~akitaoka/sakkon/sakkon2018.html (accessed 10 May 2022).

[6] H. Kenmochi, "Recent trend of singing synthesis: Technology that supports 'Hatsune Miku'," *J. Acoust. Soc. Jpn.* (*J*), **67**, 46–50 (2011) (in Japanese).

[7] STUDIO NEUTRINO: "NEUTRINO -Neural singing synthesizer," https://n3utrino.work/ (accessed 10 May 2022).

[8] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, **270**, 303–304 (1995).

[9] M. Matsuda, H. Mori and H. Kasuya, "Formants structure of whispered vowels," *J. Acoust. Soc. Jpn.* (*J*), **56**, 477–487 (2000) (in Japanese).

[10] T. Kishida, Y. Nakajima, K. Ueda and G. B. Remijn, "Effects of factor elimination on intelligibility of noise-vocoded Japanese speech," *Proc. 31st Int. Congr. Psychol.*, PS28A-01-19 (2016).

[11] H. Tamaru, S. Takamichi, N. Tanji and H. Saruwatari, "JVS-MuSiC: Free Japanese multispeaker singing-voice corpus," *arXiv preprint*, arXiv:2001.07044 (2020).

[12] X. Wang, S Takaki and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," *Proc. Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 5916–5920 (2019).

[13] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba and A. Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Process. Lett.*, **27**, 186–190 (2019).

[14] T. Uchida and M. Morise, "Investigation of voice pitch illusion using quasi singing voice and quasi whisper," *IPSJ J.*, **61**, 807–816 (2020) (in Japanese).

[15] T. Uchida, "Reversal of relationship between impression of voice pitch and height of fundamental frequency: Its appearance and disappearance," *Acoust. Sci. & Tech.*, **40**, 198–208 (2019).

Teruhisa Uchida received his B.A., M.A., and Ph.D. degrees in educational psychology from Nagoya University, Nagoya, Japan, in 1988, 1990, and 1996, respectively. From 1994 to 2002, he was a research associate in the Research Division of the National Center for University Entrance Examinations. He became an associate professor in 2002 and a professor in 2017 at the National Center. His work includes developing listening comprehension tests for the annual National Center Test for University Admissions and analyzing trends and characteristics of candidates who have taken the National Center Test. He is a member of the Acoustical Society of Japan, the Acoustical Society of America, the Japanese Psychological Association, the Japanese Association of Educational Psychology, the Phonetic Society of Japan, and the Japan Association for Research on Testing. He received the 2003 Japanese Psychological Association Research Award, was the winner of the 2016 Special Jury Prize, and 2018, 2019, and 2021 Winning Prizes of the Visual and Auditory Illusion Contest in Japan, and received the Best Paper Award from the Japan Association for Research on Testing (JART) in 2014 and 2018, and the Furui prize ASJ Paper Award in 2021.