PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Voice conversion model for estimation of transfer characteristic in auditory feedback

Shota MORITA[1]; Daiki KAWAMOTO[1]; Teruki TOYA[2]

[1] Fukuyama University, Japan

[2] Japan Advanced Institute of Science and Technology, Japan

ABSTRACT

We propose a voice conversion model for estimation of transfer characteristic in auditory feedback. A speaker feels a sense of discomfort when the speaker listens to recorded speaker's speech sound. The phenomenon is caused by the different transfer pathways. The recorded speech sound contains just air conducted (AC) sound from the speaker. On the other hand, the speech sound by the speaker has AC and bone conducted (BC) sounds in auditory feedback. There is a problem that the sound source position and the listening point cannot be underspecified. In this paper, we proposed a voice conversion model of AC and BC voices for estimation of transfer characteristic in auditory feedback. Converted voice was obtained by synthesizing AC and BC voices with time delay, AC/BC ratio, and reverberation time as parameters. BC voice was obtained by filtering from AC voice. The transfer characteristic in auditory feedback was estimated by subtracting spectrum of converted voice from that of AC voice.

Keywords: Voice conversion, Auditory feedback, Air-conducted speech, Bone-conducted speech,

1. INTRODUCTION

When we listen to our recorded speech, we have slightly different listening sense. The different sense is caused by auditory feedback (AF) in air-conducted (AC) and bone-conducted (BC) speech. We listen to our own AC and BC speech simultaneously when we speak, which means speakers perceive their own voices from their production systems while communicating through speech [1]. This mechanism is referred to as AF. The AF has an important role of the monitoring loop for pronunciation in speech chain [2]. The speech chain is considered as a mechanism of information pathways in human speech communication. Delayed auditory feedback (DAF) is one of the techniques for investigating the importance of such monitoring loop. We cannot smoothly speak in DAF experiment. It is caused by time delay of AC speech [3]. Here, the recorded own speech contains just AC speech.

The transfer function between AC and self-perceived own (SPO) voices cannot be measured. AC speech can be easily record by ordinary microphones. However, the s SPO voices at the listening point in auditory perception cannot be measured by microphone as physical measurement. Therefore, the transfer function should be estimated by other approaches. If we can obtain the transfer characteristics between AC and SPO voices, the transfer characteristics will be breakthrough for hearing of tone deaf.

The transfer function from AC to SPO voices were estimated using equalizers by peak and shelf filters [4]. The SPO voice was estimated from AC speech by voice conversion using equalizers. Thus, the transfer function was estimated from AC and SPO voices. This estimate approach from AC and SPO voices are very useful to estimate the transfer function in AF. However, the approach cannot clarify the detail of the pathway between AC and SPO voices. Voices in AF for estimating the transfer function should be converted from AC voice based on the model by AC and BC voices using conventional studies. The studies on different pathways with AC and BC voices can be hints to clarify the transfer characteristics with AC and SPO voices. Computer simulation can be one of the useful approaches to estimate the transfer function of BC voice in human head. Fujisaka et al. analyzed BC sound wave propagation using FDTD method for computer simulation [5]. The transfer

[1] s-morita@fukuyama-u.ac.jp
[2] yattin_yatson@jaist.ac.jp

function between AC and BC was estimated using singing voice [6]. In the experiment, AC and BC singing voices were recorded by Supercardioid type and piezo contact microphones, respectively. They showed the frequency characteristics of AC and BC singing voices. The AC voices were delayed about 0.4-0.8 ms from BC voices in AF [7]. In the experiment, BC voice was recorded on skin around left mastoid process by acceleration sensor, vocal-fold vibration was recorded on skin around left thyroid by acceleration sensor, and AC voice was recorded on left pinna by small microphone. The delay time between AC and BC voices in AF was calculated using the times on three recorded times.

In this paper, we propose voice conversion model based on AC and BC voices for estimation of transfer characteristic in AF. The voice conversion model can be used to obtain the SPO voice from AC voice.

2. ESTIMATION OF TRANSFER CHARACTERISTICS IN AUDITORY FEEDBACK

2.1 Estimation of Transfer Characteristics Using Voice Conversion Model

We mention that our concept for the estimation of transfer characteristic in auditory feedback in this section. One of the approaches to estimate the transfer characteristics in AF is the estimation using voice conversion from AC voice [4]. Estimation approaches of transfer characteristics using AC and SPO voices are very useful. In the research, SPO voices were obtained by an equalizer. In this research, we obtained SPO voice based on the voice conversion model of pathway from AC and BC voices. We explain the model in the next subsection.

2.2 Voice Conversion Model

Pathways for SPO voice by AC and BC voices should be considered to establish the model of voice conversion in AF. We propose a voice conversion model from AC and BC sounds. The outline of the model is shown in Fig. 1.

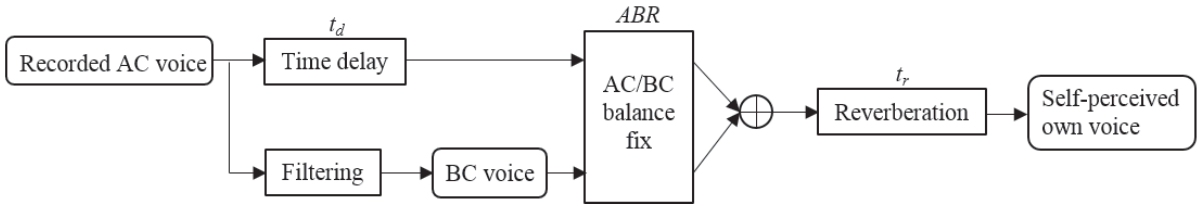

Figure 1 Voice conversion model

BC voice is converted from recorded AC voice by filtering. The filter is constructed based on spectra of AC and BC singing voices that was referred from Won & Berger [6]. The filter characteristic is obtained by subtraction AC spectrum from BC spectrum as referred in Fig. 1 in [6]. The filter characteristic for conversion from AC spectrum $S_{ac}(f)$ to BC spectrum $S_{bc}(f)$ was obtained as follows,

$$F = S_{bc}(f) - S_{ac}(f) \quad (1)$$

where f is frequency, F is the obtained filter characteristic. The obtained filter characteristic was shown in Fig. 2. The filter for conversion to BC voice consist of a low pass filter with cutoff frequency 200 Hz, and two bandpass filters of 250-500 and 750-1500 Hz, based on the Fig. 2. BC voice was obtained by these three filters.

Three parameters of time delay t_d, AC/BC power balance ABR, reverberation time t_r were used in the model to convert SPO voice from AC and BC voices. The time delay was set as the time delay of AC voice from BC voice, based on the results in [7]. The AC/BC power ratio ABR was set to fix the power difference of AC and BC voices. The ABR was calculated as follows,

$$ABR = 10 \log 10 \left(\frac{P_{ac}}{P_{bc}} \right) \text{ [dB]} \quad (2)$$

where P_{ac} is the power of AC voice and P_{bc} is the power of BC voice. The artificial reverberation using reverberation time t_r was set to assumed that the effect of reflection in head. Schroeder's room

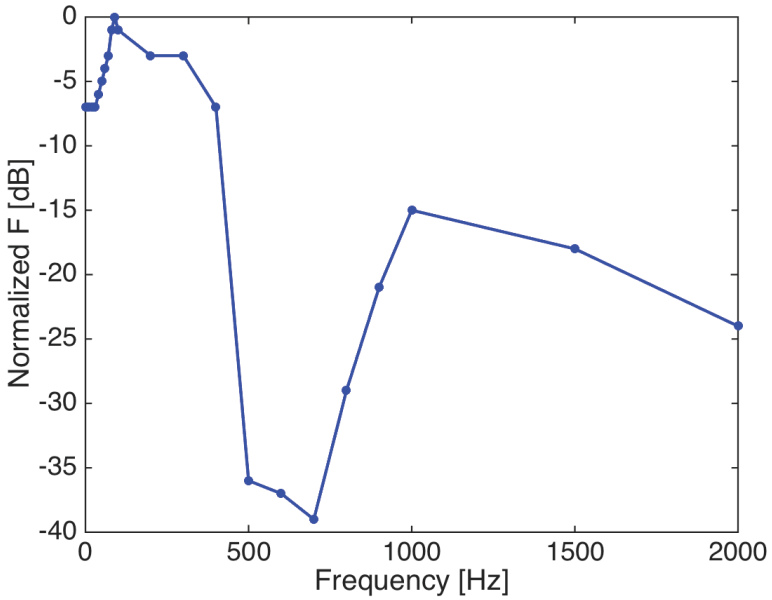

Figure 2 Filter characteristics for conversion from AC voice to BC voice

impulse response [8] were implemented as artificial reverberation. However, the enormity of the effect of the reflection is not cleared. Thus, the effect of the reflection in head will be clarified by an experiment using the model.

Arrival time of recorded AC voice was delayed using time delay t_d from arrival time of BC voice. In the next process, the delayed AC and converted BC voices were fixed these power balance, then these AC and BC voices were summed. After that, the SPO voice was obtained by convolving the impulse response as reverberation.

2.3 Application of Voice Conversion

Applications of voice conversion based on voice conversion model in subsection 2.2 was created by MATLAB. The applications consisted of recorded and conversion parts in Fig. 3 and 4, respectively. Speakers' vowels of /a/, /i/, /u/, /e/, /o/ can be recorded each three times when they select the name in the left box in Fig. 3, and push the button of "Recording Start." Recording time was determined up to 3 s. Waveform window is shown on the top for checking a waveform of the recorded own vowel after recorded. Recorded condition is shown as checked box on the right bottom in Fig. 3.

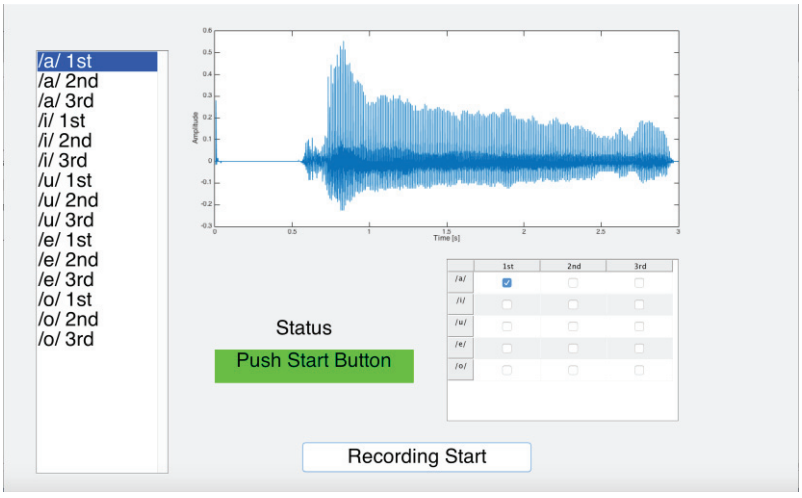

Figure 3 Screenshot of application for recording

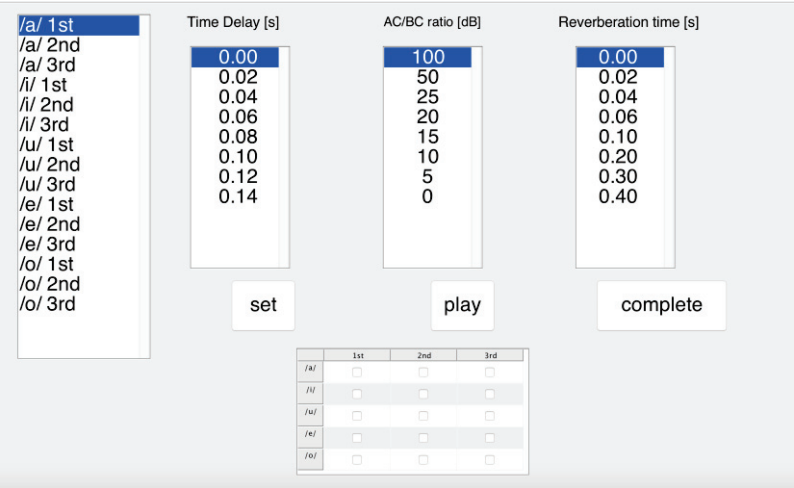

Figure 4 Screenshot of application for voice conversion

The recorded own vowels by the recording application in Fig. 3 are converted to obtain the self-perceived vowels by the application of voice conversion in Fig. 4. Delay times t_d were set 0.00-0.14 s at a certain interval 0.02 s. AC/BC ratios ABR were set 100-0 dB. ABR = 0 dB means AC and BC power level are equal to each other. Reverberation times t_r were set 0.00 to 0.40 s. A recorded own vowel are selected from left panel, then three parameters are fixed on the top panel. After clicking the "set" and "play" buttons, converted vowel was played. After fixing these parameters to best self-perceived vowel and clicking the "complete" button, the values of parameters and converted vowel are saved automatically to new .mat and .wav files, respectively.

3. EXPERIMENTS

3.1 Experimental Condition

The proposed voice conversion model from AC voice in AF was tested in the experiment.

Four male speakers aged 19-22 participated in the experiments. They are native Japanese speakers.

The AF experiments consisted of recording and voice conversion sessions. Figure 5 shows a schematic diagram of the recording system in AF experiments.

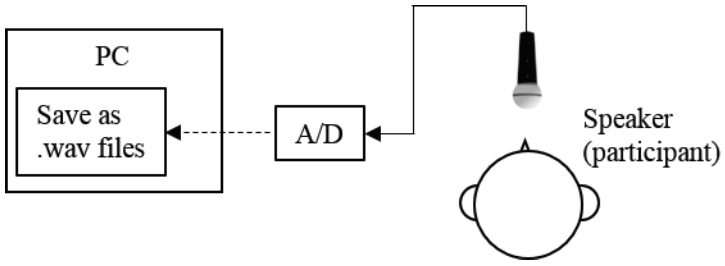

Figure 5 Schematic diagram of the recording system in AF experiments

Experiments were conducted in a small room. Speaker's voices were recorded through a microphone (Rode NT1 KIT) and routed through an audio interface (Roland QUAD-CAPTURE) to a PC (Apple MacBook Pro, with OS macOS High Sierra 10.13.8). The recording application in Fig. 3 was used to record the vowels. The sampling frequency was 8 kHz and the number of quantizing bits was 16. Figure 6 shows a schematic diagram of the voice conversion system in AF experiments. The vowels were routed through the interface, presented to the listeners (participants) through an open headphone (AKG K712) The application of voice conversion as shown in Fig. 4 was used to convert vowels. The system latency did not concern because sessions of recording and voice conversion were

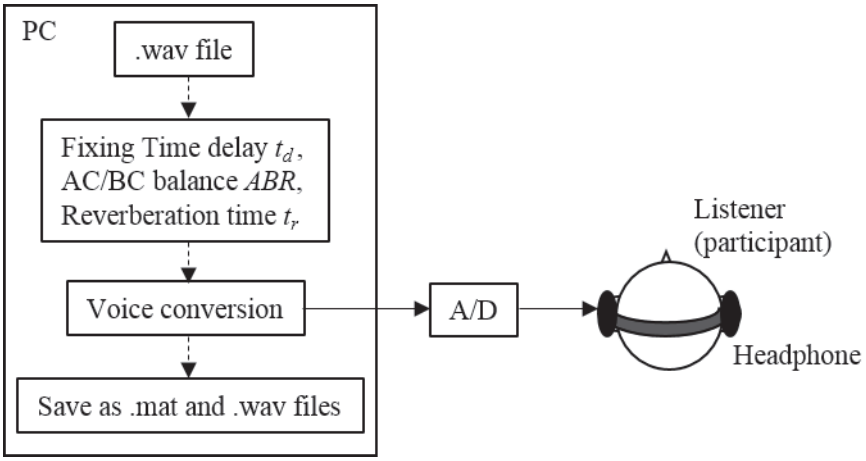

Figure 6 Schematic diagram of the voice conversion system in AF experiments

separated.

Speakers uttered five vowels of /a/, /i/, /u/, /e/, /o/ using recording application in recording session. Each vowel was uttered three times in each speaker. The distance between microphone and speaker were set around 0.2 m.

In the voice conversion session, a participant converted own recorded vowels to self-perceived vowels using voice conversion application. Sequential order of vowels in a certain conversion trial was "/a/, /i/, /u/, /e/, /o/." The trial was repeated totally three times. In each stimulus, the participant fixed three parameters to get close to the self-perceived voice. The participant can utter the same vowel while listening to converted own vowel when fixing these parameters in the experiment.

3.2 Results

The results of AF experiments were shown in Tab. 1. These results were calculated by average of participants and trials number. The results of these parameters were different in each vowel. Effects of time delay from BC voice were around 0.07 s. The results showed the possibility of time delay in section of ear canal entrance and auditory periphery because the large time lag was shown between the time delay in [7] and the results in Tab. 1. AC/BC ratios of /a/ and /o/ were smaller than the others. The results implied that the effects of BC voice in /a/ and /o/ were larger than /i/, /u/, and /e/. Reverberation times of /a/, /u/ and /e/ were over 0.1 s. It is assumed that the converted vowel from AC and BC voices was perceived like two separate stimuli because of the effects of time delay in voice conversion. However, it is assumed that the converted vowel, where the room impulse response was convolved, was heard as one unified stimulus. The results implied that the reverberation affects the voice conversion model from two signals like AC and BC voice to a signal like self-perceived voice.

The spectra of recorded and converted voice /e/ by a participant were shown in Fig. 7. For the analysis, the window size was set 2,048 and Hanning window was used. The higher frequency components were drastically attenuated in the converted (SPO) voice compared with the recorded (AC) voice.

Table 1 Results of AF experiments

	Time delay [s]	AC/BC ratio [dB]	Reverberation time [s]
/a/	0.070	37.5	0.115
/i/	0.070	48.3	0.062
/u/	0.077	47.5	0.130
/e/	0.057	67.9	0.153
/o/	0.075	37.9	0.072

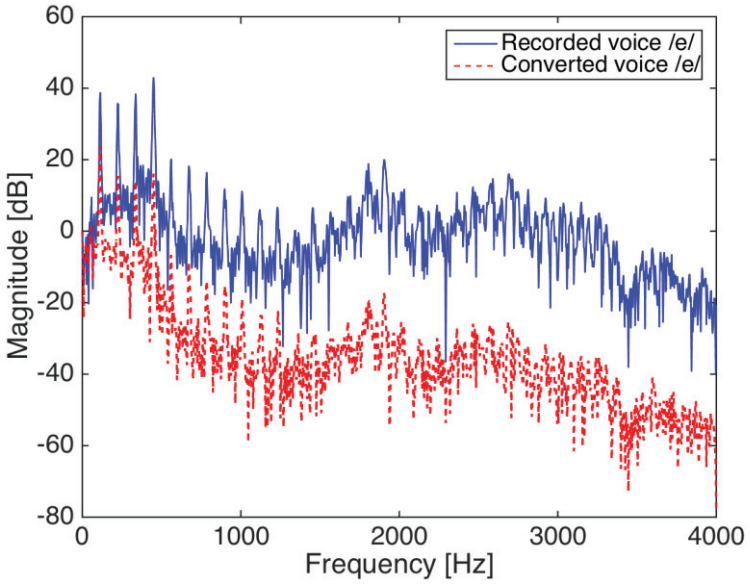

Figure 7 Spectra of recorded and converted vowels /e/ by a participant

The transfer characteristics between recorded and converted (self-perceived own) voice can be obtained by subtracting spectrum of converted voice from spectrum of recorded AC voice. This is almost the same process with estimate transfer function from AC and equalized voices [4]. The calculated transfer characteristics of voice /e/ by a participant were shown in Fig. 8. The transfer characteristics showed low-pass characteristics.

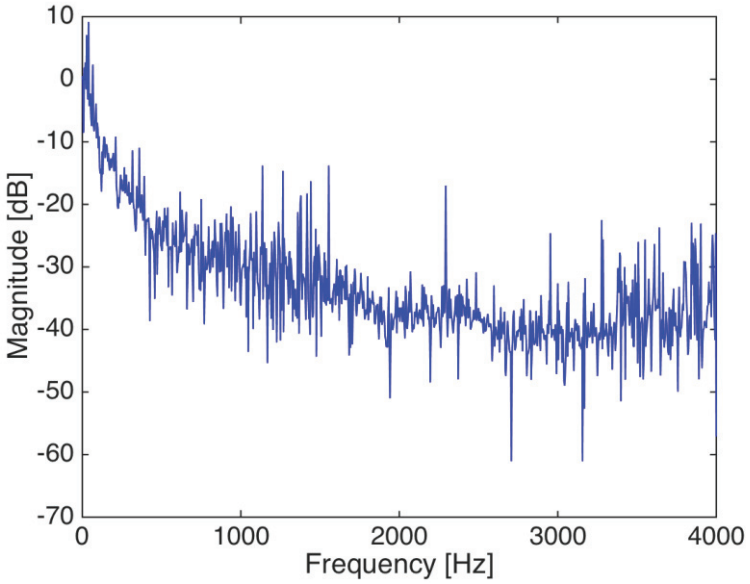

Figure 8 Transfer characteristic of vowels /e/ by a participant

4. CONCLUSIONS

This paper proposed the voice conversion model by AC and BC voices for estimating transfer characteristics between recorded and SPO voices in AF. Time delay, AC/BC ratio and reverberation time were considered as parameter in the model. The results show the possibility that the voice conversion model by AC and BC voices can be used to obtain the SPO voice from recorded AC voice. Moreover, the results showed the possibility to estimate the transfer characteristics. The individuality and affect reverberation should be considered in the future work.

REFERENCES

1. Denes PB, Pinson EN. The speech chain, 2nd ed. New York, USA: Waveland Press, Inc.; 1993.
2. Postma A. Detection of errors during speech production: a review of speech monitoring models. Congnition 2000; 77: p.87-131.
3. Toya T, Ishikawa D, Miyauchi R, Nishimoto K, Unoki M. Study on effects of speech production during delayed auditory feedback for air-conducted and bone-conducted speech. Journal of Signal Processing 2016; 20(4): p.197-200.
4. Won SY, Berger J, Slaney M. Simulation of one's own voice in a tow-parameter model. Proc ICMPC 2014; 4-8 August 2014; Seoul, South Korea.
5. Fujisaka Y, Nakagawa S, Tonoike M. Analysis of bone conducted sound wave propagation in human head. Proc. ICA 2004; 4-9 April 2004; Kyoto, Japan 2004. p. I683-I686.
6. Won SY, Berger J. Estimating transfer function from air to bone conduction using singing voice. Proc. ICMC 2005; 5-9 September 2005, Barcelona Spain 2005.
7. Ochiai Y, MURAKAMI T. Measurement of delay times between air conduction sounds and bone conduction sounds for auditory feedback. Proc. student society of IEICE Tokyo branch; Japan 2018. p. 140. (in Japanese)
8. Schroeder MR. Modulation transfer functions: definition and measurement. Acoustica 1981; p.179-182.