Acoustic Simulation of Body-conducted Speech and Its Use to Convert One's Recorded Voices to One's Own Voices

Ruiyan CHEN, Tazuko NISHIMURA, Nobuaki MINEMATSU, Daisuke SAITO
Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
E-mail: {chenry, tazuko, mine, dsk_saito}@gavo.t.u-tokyo.ac.jp, Tel/Fax: +81-3-5841-6662

***Abstract*—When one hears his/her recorded voices for the first time, s/he is probably surprised and maybe disappointed at the differences in voice quality between the recorded voices and his/her own voices. Conversion from recorded voices of a speaker to his/her own voices was technically investigated in previous studies, and in the current study, we propose a novel framework for conversion. Here, four new ideas are introduced and some of them are tested experimentally: a) multiple pathways of in-body voice transmission from the oral cavity to the inner ear are taken into account for recording, b) body-conducted speech, not bone-conducted speech, is defined and simulated, c) a special device is prepared to avoid habituation effects in listening tests, and d) a network-based voice conversion technique is applied to generate one's own voices from his/her recorded voices by using a parallel corpus developed with the above three ideas. Experiments show that the proposed framework can generate one's own voices with higher quality compared to a conventional method, even in cross-language contexts. It is very interesting that body-conducted speech has an unexpectedly larger role to simulate one's own voices compared to air-conducted speech.**

I. Introduction

When people hear their recorded voices, they will feel uncomfortable, and even doubt whether the voices are really their own [1]–[3]. In psychology, this phenomenon is called voice confrontation, which is caused by unexpected differences in the voice quality between recorded voices and own voices [4]. The former consists of only Air-Conducted Speech (ACS), while the latter contains not only ACS but also Bone-Conducted Speech (BCS) transmitted to the inner ear via bone vibration [5]–[7]. As a result, ACS generally lacks energy at low frequency bands, compared to own voices (=ACS+BCS).

Scientists have a great interest in own voices [1]–[7] and recently, brain activities of participants were monitored while they were listening to voices close or not close to their own voices [8]. It was shown that human listeners have unique sensitivity to their own voices. Engineers also have a good interest in own voices [9]–[13], because they may be used effectively in some speech applications, where users imitate given utterances for language learning, voice training, etc. What kind of utterances, or whose utterances, can be imitated better than others? Teachers say that if model utterances are given to a student which have a similar voice quality to the student's own voices, s/he will imitate the model utterances more easily and precisely [10], [11], [14].

In previous studies, many technical attempts were made to simulate own voices, and to the best of our knowledge, their attempts are classified into two approaches. In the first approach, various types of *time-invariant* filters were designed and applied to ACS to simulate own voices [9], [10], [15]–[18]. In these works, it was often stated that the ideal transfer function of the filters should be time-variant, where the function should depend on the individual phonemes observed in input. Probably because the filters actually applied were time-invariant, however, the quality of the simulated voices was not satisfactory. In the other approach, a special device for recording was used to detect BCS, which was added with an adequate weight to ACS [12], [13]. To detect BCS, a bone conduction microphone was used, which can sense and record vibrations of bones. If listeners' impression of their own voices can be simulated well as weighted sum of ACS and BCS, this approach will work well. However, the experiments seem not to show a high validity of the above hypothesis.

How to realize a flexible mapping between recorded voices (ACS) and own voices? One possibility is network-based voice conversion [19], where non-linear and time-variant mapping can be modeled with large enough training data, which are often parallel between source and target. Here, we can point out an essential and probably impossible-to-solve problem. Own voices of a speaker cannot be recorded physically and objectively because they can be heard only in mind and only by that speaker. Own voices are purely mental phenomena in mind and others cannot hear the voices. In this paper, we will make a radical attempt to prepare a parallel data between recorded voices and own voices in a scientifically valid way.

For this attempt, in this paper, a novel technical framework is proposed to finally make it possible to convert ACS to own voices with a network-based conversion technique. Here, the following four issues are addressed especially.

a) Recent findings of hearing showed that, inside the body, there exist multiple pathways of voice transmission from the oral cavity to the inner ear [20], [21]. BCS detected with a single bone conduction microphone may not be sufficient.

b) With the above fact in mind, instead of BCS, we define and introduce a new notion of boDy-Conducted Speech (DCS)[1], and we aim to simulate it with multiple microphones.

[1]BCS is Bone-Conducted Speech and DCS is boDy-Conducted Speech. Readers should be careful.

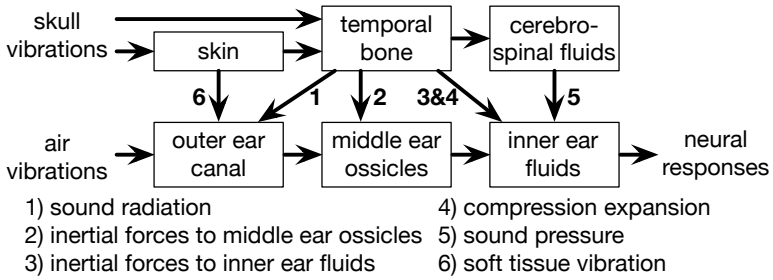

Fig. 1: Overview of the transmission pathways of ACS and BCS, modified from [20], [21]

c) To have listeners hear only their real DCS while they are speaking, and to avoid notorious habituation effects in listening tests, a special device is prepared for audio presentation.

d) By following the above steps, a parallel corpus between ACS and manually simulated own voices is built, which is used to train a network-based voice converter.

In-body voice transmission from the oral cavity to the inner ear is language-independent. To evaluate our converter from ACS to own voices in this respect, listening tests are carried out where two trilingual speakers are recruited as participants. Their Japanese utterances are used to build their converters, which are tested with their utterances in Chinese and English as well as Japanese. It should be noted that, different from assessment tests of artificial voices produced by general voice conversion systems, artificially-created own voices of a speaker have to be assessed by that speaker only, because s/he is the only person who can hear his/her real own voices.

This study has been approved by the research ethics committee of School of Engineering, The University of Tokyo.

II. The mechanism for in-body voice transmission

It seems that people understand naïvely that their own voices are generated with a single type of BCS added to ACS, but recent studies have revealed that the mechanism for in-body voice transmission is not so simple. According to [20], [21], the voice transmission from the oral cavity to the inner ear is hypothesized to take place on multiple pathways, shown in Figure 1, such as 1) sound radiated into the external ear canal, 2) middle ear ossicle inertia, 3) inertia of the cochlear fluids, 4) compression of the cochlear walls, and 5) pressure transmission from the cerebrospinal. In several previous studies, bone vibrations were detected from a participant with a bone conduction microphone attached near his/her ear canal, which is numbered as 1) in Figure 1. Then, the detected signals were treated as BCS in those studies. Taking Figure 1 into account, however, this approach is not valid enough, because real BCS should be a mixture of the multiple signals.

If good sensors are available to detect signals at the individual regions separately on the multiple pathways, those signals should be used effectively to simulate one's own voices. Since the signals of 2) to 5) cannot be detected separately [20]–[22], however, we introduce a new term, boDy-Conducted Speech (DCS), to indicate what remains after removing ACS from one's own speech (OOS). Namely, DCS ≡ OOS − ACS, and it is regarded as sum of the signals on the in-body pathways in Figure 1. It can further be interpreted as the component in OOS that cannot be observed as acoustic signals.

How to observe DCS objectively? As explained in the previous section, DCS cannot be observed acoustically, or objectively. By speaking with ACS suppressed in an adequate way, however, any listener can hear his/her DCS only. In the following section, we prepare a special device for hearing real DCS and simulated DCS at the same time.

To simulate DCS, we use multiple microphones available to detect several signals in Figure 1. Here, we still use the term of BCS, but it is used always as Bone-Conducted Speech detected at a specific position on the body. In the current study, three microphones are available, which are designed to detect BCS at the ear canal, BCS on the skin of the throat, and BCS on the top of the head, i.e. the skull. Through some preliminary testing, we did not use the last one because recording with the skull microphone was not stable. Finally, our simulation of DCS is made by using three sources, ear BCS (eBCS), throat BCS (tBCS) and ACS. eBCS and tBCS are numbered as 1) and 6) in Figure 1, and all of the three kinds of signals are recorded completely synchronously in parallel.

III. Preparations for recording and presentation

A. Recording

ACS was recorded with a general condenser microphone, while the two kinds of BCS were detected with bone conduction microphones. eBCS and tBCS were recorded with an earphone-type microphone (TEMCO EM20N-T3 [23]) and a laryngeal microphone (TEMCO TM80N-T [24]), respectively. Figure 2 shows how these two microphones were attached on

Fig. 2: Attachment of 2 types of BC microphones

the body. The left-hand side shows the microphone for eBCS and the other is for tBCS. For the current study, the cut-off frequency of low-pass filtering in the microphones was modified from 2 kHz, which is the original value, to 5 kHz.

To realize completely synchronized recording with the three microphones, they were connected to a laptop computer with the same type of USB audio adapters. All the recordings were obtained in a soundproof room and then, the synchronized recordings were edited using Adobe Audition 2020. As for reading-aloud material, following previous studies [9], [10], the five vowels of Japanese were used. Further, as more practical material, a specially designed paragraph was used, which contains all the kinds of Japanese morae[2]. Each mora appears only once, but the resulting paragraph is very meaningful [25].

As explained in Section I, in previous studies, OOS was simulated in two ways: filtering ACS, and weighted sum of ACS and eBCS. In these studies, researchers concluded that the ideal transfer function of filtering depended on phonemes, and that so did the weights of ACS and BCS. These findings imply that, to simulate OOS, time-variant filters and time-variant weights should be adequately designed. In this paper, time-variant characteristics of filters are implemented technically by using three microphones and two-step simulation of OOS, which are explained in detail in Section IV.

B. Soundproof earmuffs with small bluetooth speakers

In Section II, we defined DCS as OOS−ACS, and in the following sections, we firstly simulate DCS only, and after that, we try to simulate OOS by adding ACS with an adequate weight to the simulated DCS. To realize a good simulation of DCS, a special listening device is introduced.

How to suppress ACS? One can hear different types of DCS by suppressing ACS in different ways. DCS heard by inserting the two index fingers into the ear canals and DCS heard by covering the ear canals by the palms are perceived as voices with different voice qualities. These differences are discussed in [26] and some of them are known as occlusion effect. The vibrations of speech cause the outer wall of the external canal to vibrate, which make extra standing waves occur.

Which type of DCS should be used as target of simulation in our study? We want to simulate the DCS with listeners' ear canals open, where extra and unnatural standing waves do not take place. With listeners' ears open, however, how to be able to suppress ACS? We solve this problem by using large and strong earmuffs (3M PELTOR X5A, NRR=31dB), shown in Figure 3. By wearing these soundproof earmuffs, the ears are covered by the big ear cups. It should be noted that urethane sponge is put as sound-absorbing material on every part of the inside wall of the ear cups, which can suppress extra standing waves effectively. With these earmuffs over the ears, speakers can hear their own DCS only. Similar suppression is possible with noise cancelling headphones, but they are so good and smart that they can also suppress a part of DCS, probably 1) in Figure 1. This is why noise cancelling headphones should not be used at all in our study.

Fig. 3: Soundproof earmuffs with small bluetooth speakers for easy-to-compare listening experiments

With these earmuffs, listening to real DCS just by speaking is made possible. For listening experiments, however, we have to prepare an environment where participants compare their real DCS with simulated DCS in a reliable way. If simulated DCS is presented through headphones, they have to exchange the earmuffs and the headphones very frequently. In our study, the transfer function of filtering and the weights for summation have to be optimized manually through repeated listening experiments. It is known that repeated presentation and listening often induce habituation effects, with which subjective judgements will become biased. Thus, reliable judgement is extremely important in our experiments and it will be made possible only with an easy-to-compare listening environment.

To realize this environment, we installed small bluetooth stereo speakers (Bearoam Bluetooth 5.1 wireless earhook-type speakers) inside the ear cups of the earmuffs, shown in Figure 3. With this integrated device, participants can listen to real DCS just by speaking and to simulated DCS just by clicking on a laptop. Synchronized listening is also easy, which we consider enhanced reliability of judgment. In our experiments, participants have to compare real DCS with simulated DCS repeatedly. If a good simulated DCS is presented synchronously with its real DCS, these two stimuli will be perceived as an utterance of a single speaker[3]. After the experiments, participants commented that synchronous listening to the two stimuli realized easy and reliable judgement.

IV. Experiments

The experiments were carried out in three steps. 1) DCS was simulated with ACS, eBCS, and tBCS. 2) OOS was simulated with ACS and the simulated DCS. After these two steps, we built a parallel corpus between ACS and the simulated OOS. Finally, 3) a network-based voice converter from ACS to OOS was built. The overview of these steps is shown in Figure 4. In the first two steps, the values of the parameters required for simulation were tuned manually through listening experiments. This is because DCS and OOS cannot be obtained as acoustic signals and they exist only mentally in mind.

[2]Mora is the fundamental unit of speech production of Japanese. It is an open syllable in the form of V or CV.

[3]Synchronous utterances of identical twins are often perceived as an utterance of a single speaker because of acoustic resemblance. Participants judged whether this happened or not with real DCS and simulated DCS.

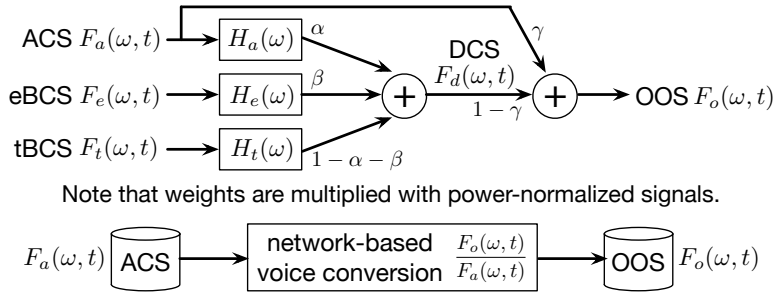

Fig. 4: Overview of the experiments

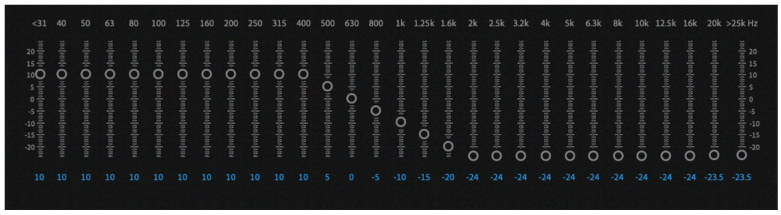

Fig. 5: The graphic equalizer of Adobe Audition 2020 with 20 bands, used to determine $H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$

A. Simulation of DCS with ACS, eBCS, and tBCS

In previous studies, when OOS was simulated manually by filtering ACS, its transfer function was found to depend on the phonemic identity of the input segment in ACS, indicating that the transfer function should be time-variant. In this study, we firstly attempt to simulate DCS only by filtering ACS, eBCS, and tBCS separately with different but time-invariant transfer functions ($H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$ in Figure 4). If we denote the spectrum of ACS, eBCS, tBCS, and DCS as $F_a(\omega,t)$, $F_e(\omega,t)$, $F_t(\omega,t)$, $F_d(\omega,t)$, respectively, the three transfer functions are manually designed so that the following equations are approximately satisfied,

$$F_d(\omega,t) \approx H_a(\omega)F_a(\omega,t) \quad (1)$$

$$\approx H_e(\omega)F_e(\omega,t) = H_e(\omega)G_{ea}(\omega,t)F_a(\omega,t) \quad (2)$$

$$\approx H_t(\omega)F_t(\omega,t) = H_t(\omega)G_{ta}(\omega,t)F_a(\omega,t), \quad (3)$$

where $G_{ea}(\omega,t)$ and $G_{ta}(\omega,t)$ are defined as

$$G_{ea}(\omega,t) \equiv F_e(\omega,t)/F_a(\omega,t) \quad (4)$$

$$G_{ta}(\omega,t) \equiv F_t(\omega,t)/F_a(\omega,t), \quad (5)$$

which are interpreted as in-body filters representing how ACS should be modified in a *time-variant* way to approximate eBCS and tBCS, respectively. As claimed in previous studies [9], [10], [15]–[18], the spectral changes observed from ACS to BCS or OOS are phoneme-dependent, i.e. time-variant, the two transfer functions of G_{ea} and G_{ta} should be treated as dependent on t. Since the above approximations will not be precise enough, however, we sum the three outputs in Figure 4 with adequate weights of α, β, and $1-\alpha-\beta$ to realize better approximation. Finally, DCS is simulated as

$$F_d(\omega,t) = \{\alpha H_a(\omega) + \beta H_e(\omega)G_{ea}(\omega,t) + (1-\alpha-\beta)H_t(\omega)G_{ta}(\omega,t)\} \times F_a(\omega,t). \quad (6)$$

This is how we realized time-variant characteristics of the transfer function to convert ACS ($F_a(\omega,t)$) to DCS ($F_d(\omega,t)$), although $H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$ are time-invaraint.

How to optimize $H_a(\omega)$, $H_e(\omega)$, $H_t(\omega)$, α, and β for approximation? In this study, as explained above, we did not realize global and simultaneous optimization, but we estimated these parameters sequentially. The three transfer functions were designed separately by a participant of the experiments using the equalizing module of Adobe Audition 2020 with 20 bands, so that each of the filter outputs of ACS, eBCS, and tBCS sounded like his/her own DCS. Figure 5 shows the interface of the equalizing module. Here, his/her reading-aloud of the special paragraph containing all the kinds of morae was used as stimulus. After $H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$ were fixed, the three output signals were summed with weights of α and β. They were determined through listening experiments again, where the most adequate set of weights was selected out of ten candidate sets prepared. After fixing the two weights, simple listening tests were carried out with all the participants, where the summed signals with the selected weights were compared with each of the filter outputs. The former was found to be always better in quality. All the listening experiments for designing the filters and fixing the weights were carried out with the integrated earmuffs explained in Section III.

B. Simulation of OOS with ACS and the simulated DCS

Once the quasi-optimal set of the three transfer functions and the two weights were fixed, the simulated DCS can be generated from ACS, eBCS, and tBCS of any input utterance. As shown in Figure 4, to simulate OOS, $F_o(\omega,t)$, ACS was added to the simulated DCS with weight γ.

$$F_o(\omega,t) = [\gamma + (1-\gamma)\{\alpha H_a(\omega) + \beta H_e(\omega)G_{ea}(\omega,t) + (1-\alpha-\beta)H_t(\omega)G_{ta}(\omega,t)\}] \times F_a(\omega,t). \quad (7)$$

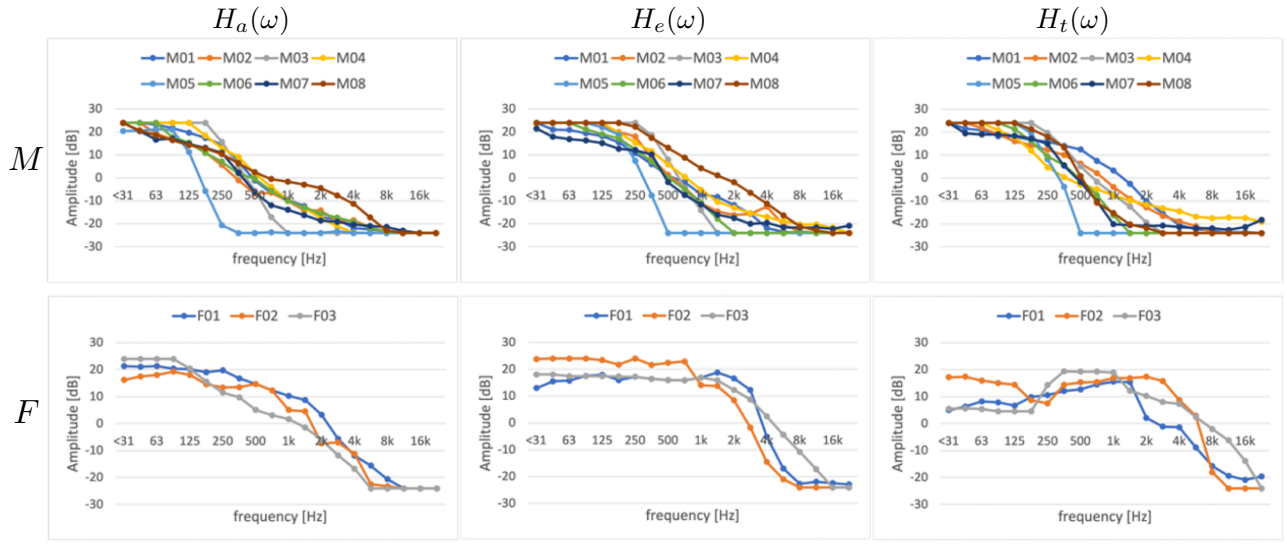

Fig. 6: Amplitude characteristics of $H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$ designed for males (up) and females (down)

The value of γ was also determined again through listening tests. Here only the earhook-type speakers were used, not with the earmuffs, for audio presentation. Since all the optimizations in Figure 4 had to be made manually and repeatedly through listening tests, the easy-to-compare environment prepared for the experiments in Section III was very crucial.

C. Network-based conversion from ACS to OOS

Now, all the parameters were fixed carefully, separately for each participant in the experiments. With the three types of microphones and these parameters, any input utterance can be converted to its OOS, which is simulated OOS to be exact. Since ACS and the simulated OOS were completely synchronized, it was easy to build their parallel corpus, with which, a network-based converter was trained. Functionally speaking, we expected the network to learn the following time-variant transfer function through network training.

$$F_o(\omega,t)/F_a(\omega,t) = \gamma + (1-\gamma)\{\alpha H_a(\omega) + \beta H_e(\omega)G_{ea}(\omega,t) + (1-\alpha-\beta)H_t(\omega)G_{ta}(\omega,t)\}. \quad (8)$$

As explained in Section I, the main effect that should be realized by this converter was enhancement of the energy at low frequency bands as well as modification of some other spectral properties. Taking this expectation into account, we used a voice conversion method that can modify source waveforms based on the log-spectral differences between the source (ACS) and its target (OOS) [27]. Here, for the source spectrum, DNN-based prediction of its target was conducted. The log-spectral differences between the source and the target were calculated and the amplitude characteristics of the differences were approximated by the Log-Magnitude Approximation (LMA) filter [28]. It can convert ACS directly to OOS in a time-variant way, even without a vocoder. No use of any vocoder can improve the quality of the converted speech.

D. Participants and material

To evaluate our framework, we recruited seven male and three female adults as participants, all of whom did not have any difficulty in hearing and reading aloud. From them, we recorded a few instances of each of the five Japanese vowels, and a reading-aloud of the special paragraph. The vowels were used only to examine the conventional methods, and for our framework, we used recordings of the special paragraph to simulate DCS with ACS, eBCS, and tBCS and to simulate OOS with ACS and the simulated DCS.

The transfer functions and the weights manually designed separately by each of the ten participants were used to evaluate the performance of our method to simulate DCS and OOS manually. Further, our automatic converter from ACS to OOS was evaluated by selected participants, who were two trilinguals. Since the in-body voice transmission in Figure 1 is language-independent, we evaluated the converter in a cross-language context. We asked the selected participants to read aloud a phonemically-balanced set of 40 Japanese sentences [29], which were used to train a network-based voice converter. For testing, other 5 Japanese sentences as well as 5 Chinese sentences and 5 English sentences were read aloud. Testing sentences were extracted from a multilingual speech corpus [30], which contains 11 sentence sets of the story of "the North Wind and the Sun" in 11 languages.

V. Results and discussion

A. Amplitude characteristics of $H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$

$H_a(\omega)$, $H_e(\omega)$, and $H_t(\omega)$ were designed separately by each participant, shown in Figure 6. Regardless of the types of microphones, LPF was always obtained. The transfer functions of the designed filters show clearer differences between the two genders than among the types of microphones. As the vocal tract is shorter in female, the spectral features at higher

TABLE I: 10 candidate weight sets prepared for filtered ACS, eBCS, and tBCS to simulate DCS

α	ACS	0	0	0	0	0.2
β	eBCS	0.2	0.4	0.6	0.8	0.2
$1-\alpha-\beta$	tBCS	0.8	0.6	0.4	0.2	0.6
α	ACS	0.2	0.2	**0.4**	**0.4**	**0.6**
β	eBCS	0.4	0.6	**0.2**	**0.4**	**0.2**
$1-\alpha-\beta$	tBCS	0.4	0.2	**0.4**	**0.2**	**0.2**

frequency bands should be retained even in their DCS. If the filters designed for male were applied to female recordings, on some speech segments, their phoneme identity was perceived as different from the original phoneme identity.

B. Weights for ACS, eBCS, and tBCS to simulate DCS

Table I shows ten candidate weight sets, out of which each participant selected the optimal one that can simulate his/her DCS the best. These candidate sets had been prepared through preliminary tests. All the participants selected a weight set with $\alpha \geq 0.4$, i.e. one of the three sets in bold in the table. The averages are 0.48, 0.28, and 0.24, meaning that filtered ACS and filtered BCSes are equally important to simulate DCS, and that the two types of BCSes are also equally important to simulate DCS. As described in Section IV-A, very simple listening tests were conducted to compare the quality between each of the three filter outputs in Figure 4 and the simulated DCS, i.e. weighted sum of the three outputs. All the participants preferred the weighted sum. In this experiment, however, simultaneous optimization was not made in designing the three filters and fixing the three weights. The simultaneous optimization is approximately possible with interactive genetic algorithms [31], which is one of our future works.

C. Weights for ACS and DCS to simulate OOS

Similar to the experimental setup used in Section V-B, nine candidate values were prepared for γ, which varied from 0.1 to 0.9 with a step of 0.1. After listening to each of the nine stimuli, a participant was asked to judge the acoustic similarity of the simulated OOS to his/her real OOS qualitatively. Here, a simple three-level scale was used and the three levels indicated bad, fair, and good. After listening to all the stimuli, the participant re-evaluated only the good stimuli to select the best one after intensive listening. We used this two-step strategy to reduce the time required for the entire listening tests.

In this experiment, not only the stimuli generated with our proposed method, but also those generated with a previous study [12] were used, where a simple weighted sum of ACS and eBCS was used as simulated OOS. This approach can be realized as a special case of Figure 4, where $H_a(\omega)$=0, $H_e(\omega)$=1, $H_t(\omega)$=0, α=0, and β=1. Table II shows the results, where v_1 and v_2 mean the stimuli generated by the previous study and those generated by our proposal, respectively. It should be noted that the best stimulus was selected out of the entire stimulus sets of v_1 and v_2.

It is found that the best stimulus was always one of the v_2 candidates, which clearly indicates superiority of our method

TABLE II: Quality assessment to fix the value of γ

v_1:previous method [12], v_2:proposed method
×:bad, △:fair, ○:good, ★:best

	M02		M03		M04		M05		M06	
γ	v_1	v_2	v_1	v_2	v_1	v_2	v_1	v_2	v_1	v_2
0.1	×	○	×	★	△	△	×	×	×	○
0.2	×	★	×	○	×	△	×	△	×	★
0.3	×	△	×	○	△	△	×	○	×	○
0.4	×	○	×	△	×	○	△	○	×	○
0.5	×	×	×	△	△	★	△	★	×	○
0.6	×	○	×	△	△	○	△	○	×	△
0.7	×	○	×	△	○	○	△	○	×	△
0.8	△	○	×	×	○	○	○	○	△	△
0.9	△	△	×	×	○	△	○	○	△	△

	M07		M08		F01		F02		F03	
γ	v_1	v_2	v_1	v_2	v_1	v_2	v_1	v_2	v_1	v_2
0.1	○	○	×	★	×	★	×	×	×	△
0.2	○	○	×	△	×	○	△	△	×	○
0.3	△	○	×	△	×	○	△	○	×	★
0.4	○	○	×	△	×	○	△	○	×	○
0.5	△	★	△	△	×	○	△	○	△	○
0.6	○	○	△	△	△	×	△	○	×	○
0.7	○	○	△	△	○	△	△	★	×	△
0.8	○	○	△	△	△	×	○	○	×	○
0.9	○	○	△	△	×	×	○	○	△	△

to the previous method. The averaged value of γ for the best stimuli is 0.35, meaning that the best weight for DCS is 0.65. It is very interesting that the simulated DCS is unexpectedly very dominant in the participants' OOS. This may be a reason why some people even doubt that their recorded voices are generated from their mouths [1]–[3].

D. Performance of the network-based converter

The converter was assessed in two ways, objectively and subjectively, by using ACS and two kinds of simulated OOS, i.e. manually simulated OOS (mOOS) and automatically converted OOS from ACS (aOOS). Objective assessment was conducted by calculating the mel-cepstrum distortions in three cases of (ACS, aOOS), (ACS, mOOS), and (aOOS, mOOS). Here, for two time-aligned sequences of mel-cepstrums, $X=(x_1, x_2, ,, .x_T)$ and $Y=(y_1, y_2, ,, .y_T)$, the distortion is quantified in the following equation, where 24 dimensional cepstrums were extracted with STRAIGHT [32].

$$\text{Mel-CD[dB]} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\ln 10}\sqrt{2||x_t - y_t||^2}. \tag{9}$$

On the other hand, subjective assessment was made by having two trilingual participants, one male and one female, listen to their ACS and the two kinds of simulated OOS. Since real OOS of a speaker can be heard only by the speaker, subjective assessment had to be made in a speaker-closed mode.

The original 15 utterances, 5 for each of Japanese, Chinese, and English, were recorded from the two trilingual speakers, and they were used as ACS. All the utterances were converted to mOOS and aOOS. Figure 7 shows the averaged amplitude of the mel-cepstrum distortion over the 5 utterances for each language. The upper figure shows the results of the male speaker and the lower shows those of the female speaker.

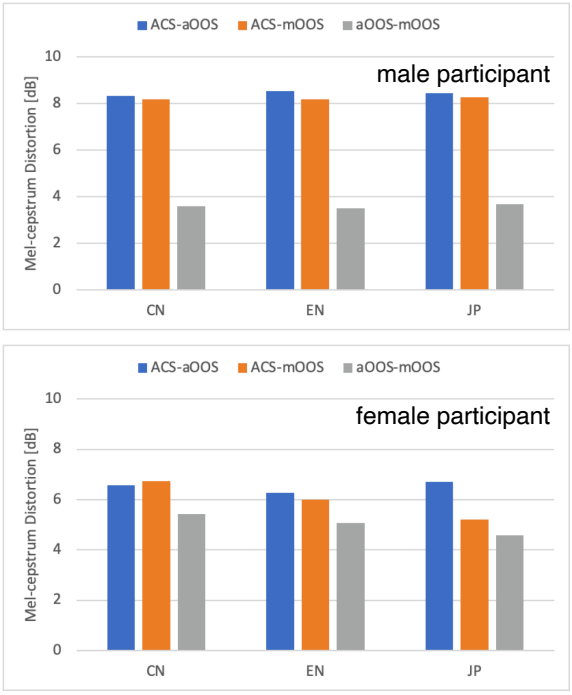

Fig. 7: Objective assessment of the converter

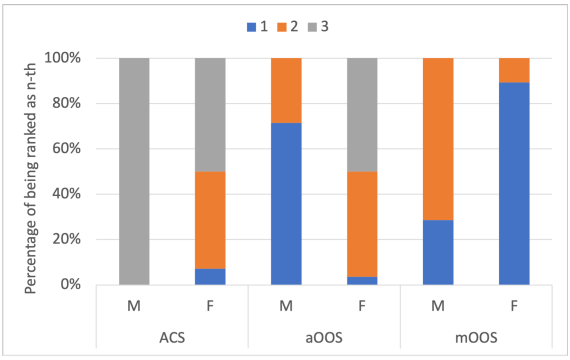

Fig. 8: Subjective assessment of the converter

As was highly expected, the results are totally language-independent, but clear differences are found between the two genders. The male speaker shows larger distortions in (ACS, aOOS) and (ACS, mOOS) compared to the female speaker. This is reasonable because, as shown in Figure 6, in male speakers, a larger potion of the ACS spectrum has to be suppressed to simulate DCS, while it has to be retained in female speakers. The performance of approximation by the network-based conversion is assessed objectively by aOOS-mOOS. The Mel-CD of the male speaker is so small as about 3.5 [dB], and we can say that a good enough performance of approximation was realized for this speaker. According to t-test, however, the performance was found to be significantly worse with the female speaker (p<0.001, See Figure 7). We have to examine this tendency in the future work by using a larger number of speakers of both genders.

From the objective fact that ACS-mOOS is much larger and aOOS-mOOS is significantly smaller in the male speaker, it is expected that better scores of subjective assessment will be obtained from the male speaker. To confirm this, for each of the 15 utterances, ACS, aOOS, and mOOS were randomly presented and the two participants were asked to order the triplet in terms of similarity to their real OOS. Figure 8 shows the percentage of ACS, aOOS, and mOOS being ranked as n-th. As expected, the male participant preferred aOOS, while the female participant preferred mOOS. The network-based conversion has to be tuned more to the female speaker.

From these results, we can claim at least that our framework of simulating OOS as weighted sum of DCS and ACS, where DCS is simulated as weighted sum of filtered ACS and BCSes, is more effective than the previous method [12], but fine tuning is still needed, which may be made possible by simultaneous optimization of the filters and the weights, and by network topology optimization for conversion even with a larger number of training speakers.

VI. Conclusions

In this paper, based on recent findings on in-body transmission of voices from the oral cavity to the inner ear, a novel framework was proposed and tested. Firstly, the framework simulated body-conducted speech, which is theoretically defined as what remains after subtracting air-conducted speech from one's own speech. The simulation used three source signals, i.e. air-conducted speech, bone-conducted speech detected near to the ear canal, and bone-conducted speech detected on the larynx. The three source signals were processed with three individual filters which were manually designed so that the filtered speech can approximate the body-conducted speech. To realize better approximation, the three kinds of the filtered speech were summed with adequate weights. Secondly, the proposed framework simulated one's own speech by adding air-conducted speech to the simulated body-conducted speech with an adequate weight. For designing the filters and fixing the weights, a special listening device was prepared to make repeated listening tests highly reliable. After building a parallel corpus between air-conducted speech and its corresponding simulated one's own speech, finally, we built a network-based voice conversion system that converts the former speech to the latter speech. Although development of the system was preliminary, validity of the proposed framework was shown experimentally even in cross-language contexts.

As future work, we're interested in fine tuning of each step conducted in this study and in applying the system to speech applications. For detail, we will carry out 1) optimizing the filters and the weights simultaneously, 2) optimizing the network topology of the conversion system with a larger amount of training speakers, 3) training the conversion system so that it can run in a speaker-independent way, 4) assessing the system using a larger number of participants, and 5) introducing the system to speech applications, where users imitate model speakers in language learning and drama training, and model singers in vocal training.

References

[1] P. S. Holzman and C. Rousey, "The voice as a percept." *Journal of Personality and Social Psychology*, vol. 4, no. 1, p. 79, 1966.

[2] C. L. Rousey and P. S. Holzman, "Recognition of one ' s own voice," *Journal of Personality and Social Psychology*, vol. 6, no. 4, pp. 464–466, 1967.

[3] A. J. Weston and C. L. Rousey, "Voice confrontation in individuals with normal and defective speech patterns," *Perceptual and motor skills*, vol. 30, no. 1, pp. 187–190, 1970.

[4] R. J. Davidson, G. E. Schwartz, and D. Shapiro, *Consciousness and self-regulation*. Springer, 1983.

[5] G. v. Békésy, "Note on the definition of the term: hearing by bone conduction," *J. Acoust. Soc. Am.*, vol. 26, p. 106, 1954.

[6] J. Tonndorf, "A new concept of bone conduction," *Arch Otolaryngol.*, vol. 87, pp. 595–600, 1968.

[7] D. Maurer and T. Landis, "Role of bone conduction in the self-perception of speech," *Folia Phoniatr Logop*, vol. 42, pp. 226–229, 1990.

[8] T. Hosaka, M. Kimura, and Y. Yotsumoto, "Neural representations of own-voice in the human auditory cortex," *Scientific Reports*, vol. 11, no. 591, 2021.

[9] I. Nakayama, "Voice timbre in autophonic production compared with that in extraphonic production," *Journal of the Acoustical Society of Japan (E)*, vol. 18, no. 2, pp. 67–71, 1997.

[10] N. Minematsu, N. Nakamura, and S. Nakagawa, "A study on the generation of speech heard during the autophonic production using analysis-synthesis techniques," in *Proc. Spring Meeting of Acoust. Soc. Jpn.*, 2000 (in Japanese), pp. 211–212.

[11] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors – in search of the golden speaker," *Speech Communication*, vol. 37, no. 3, pp. 161–173, 2002.

[12] M. Mori, C. Yoshida, M. Ogihara, S. Taniguchi, and K. Takahashi, "The rate of air-transmitted and bone-transmitted sounds in autophonic production," *IEEJ Transactions on Electronics, Information and Systems*, vol. 127, no. 8, pp. 1268–1269, 2007.

[13] T. Shimada and M. Imura, "Technical support for vocal imitation by synthesizing one's own speech," in *IPSJ SIG Technical Report*, vol. MUS-122, 2019 (in Japanese).

[14] M. Yoram and K. Hirose, "Language training system utilizing speech modification," in *Proc. ICSLP*, 1996, pp. 1449–1452.

[15] M. Kimura and Y. Yotsumoto, "Auditory traits of "own voice"," *PLoS ONE*, vol. 13, no. 6, 2018.

[16] L. I. Shuster and J. D. Durrant, "Toward a better understanding of the perception of self-produced speech," *J. Communication Disorders*, vol. 60, pp. 1–11, 2003.

[17] A. Vurma, "The timbre of the voice as perceived by the singer him-/herself," *Logop Phoniatr Vocology*, vol. 39, pp. 1–10, 2014.

[18] S. Y. Won, J. Berger, and M. Slaney, "Simulation of one's own voice in a two-parameter model," in *Proc. Int. Conf. Music Percept. Cogn.*, 2014.

[19] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: from statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[20] S. Stenfelt and R. L. Goode, "Bone-conducted sound: physiological and clinical aspects," *Otology & Neurotology*, vol. 26, no. 6, pp. 1245–1261, 2005.

[21] S. Stenfelt, "Acoustic and physiologic aspects of bone conduction hearing," *Implantable Bone Conduction Hearing Aids*, vol. 71, pp. 10–21, 2011.

[22] T. Toya, P. Birkholz, and M. Unoki, "Analysis of transmission characteristics of bone-conducted speech using spoken voice," in *IEICE Technical Report*, vol. SP2017-150, 2018 (in Japanese).

[23] *TEMCO EM20N-T3*, https://www.temco-j.co.jp/products/em20n-t3/.

[24] *TEMCO TM80N-T*, https://www.temco-j.co.jp/products/tm80n-t/.

[25] *Crayon*, http://www.ntt-i.net/IROHA/iro/kureyon.html.

[26] M. O. Hansen and M. R. Stinson, "Air conducted and body conducted sound produced by own voice," *Canadian Acoustics*, vol. 26, no. 2, pp. 11–19, Jun. 1998.

[27] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, 2018.

[28] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Trans. Electronics and Communications*, vol. 66-A, no. 2, pp. 10–18, 1983.

[29] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[30] *University of Tsukuba MultiLingual speech corpus (UT-ML)*, http://research.nii.ac.jp/src/en/UT-ML.html.

[31] H. Farooq and M. T. Siddique, "A comparative study on user interfaces of interactive genetic algorithm," *Procedia Computer Science*, vol. 32, pp. 45–52, 2014.

[32] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.