IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 25 March 2024, accepted 30 March 2024, date of publication 12 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3388195

RESEARCH ARTICLE

The Effect of Laryngeal Vibratory Stimuli and User's Vocal Style on Sense of Vocal Agency

YUKI SHIMOMURA, YUKI BAN, AND SHIN'ICHI WARISAWA, (Member, IEEE)
Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 2778563, Japan

Corresponding author: Yuki Shimomura (shimomurayuki@lelab.t.u-tokyo.ac.jp)

This work was supported by JST SPRING under Grant JPMJSP2108.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The University of Tokyo Ethics Committee under Application No. 21-92.

ABSTRACT Vocalization significantly impacts daily human activities by affecting cognitive, physical, and emotional aspects. Despite its importance, physical and social barriers can restrict individuals' ability to express themselves vocally. To address this issue, the authors have proposed an innovative method to enhance vocal agency through auditory and laryngeal vibratory stimuli. However, prior research has predominantly examined the effectiveness of this technique in scenarios involving loud vocalizations, and the validity of this method has thus yet to be sufficiently verified. In this study, we examined the effects of vibratory stimuli on the sense of agency in the case of general vocalizations. Findings revealed that vibratory stimuli applied to areas not directly involved in vocalization, such as the wrist, failed to foster a sense of vocal agency, in stark contrast to the effects observed with stimuli applied to the larynx. These results underscore the importance of directing vibratory stimuli to anatomical regions integral to vocalization to enhance vocal agency effectively. This research has substantial potential to influence the development of interactive and virtual reality technologies and support individuals experiencing speech difficulties.

INDEX TERMS Interaction, sense of agency, vibrotactile stimuli, vocalization.

I. INTRODUCTION

Vocalization serves as a critical mechanism for disseminating information in daily interactions, exerting considerable influence on the speaker's cognitive, physical, and emotional aspects. One notable phenomenon, the production effect, delineates how vocalization can enhance the articulation and recollection of words, thereby improving memory retention [1]. In addition, vocalization, especially loud vocalization, has been confirmed to augment muscle output performance by activating the central nervous system [2], and loud vocalization is also expected to reduce mental stress.

Vocalization has emerged as a significant interactive technology. Its application in various domains, notably in human-computer interaction, saw rapid expansion during the 2000s, fostering new forms of community engagement [3]. Given its operation solely through the respiratory system, vocal interaction is the optimal mode of communication in eyes-free or hands-free environments [4]. In addition, language modality is familiar and well-developed, and it can thus be easily used to collaborate with others in a virtual reality (VR) environment [5]. Furthermore, vocalization significantly enhances VR experiences by deepening the sense of presence and immersion owing to the intimacy that it facilitates.

However, vocalization faces notable challenges, including poor performance of speech recognition in noisy environments [3] and severe limitations in employing spoken language for human-computer interactions [6]. Nonetheless, these impediments are being incrementally mitigated with improvements in speech-recognition technologies such as Siri (Apple Inc.) and Alexa (Amazon.com, Inc.).

Moreover, vocalization is constrained by various physical, psychological, and environmental factors. Issues such as

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Chen.

organic dysphonia, exemplified by laryngitis from infections, and functional dysphonia, often linked to psychological conditions, exemplify internal limitations. Excessive vocalization risks the vocal cords [7], primarily due to tensile stress [8]. These problems should not be ignored, as dysphonia is known to affect the health of individuals adversely [9]. Social norms also dictate volume control to avoid disturbing others and prevent the dissemination of private information, particularly in public settings such as transportation or public facilities.

Approaches addressing these constraints are being developed to mitigate sound leakage due to vocalization and to facilitate voice generation without actual voice. Some products that suppress voices are available in the market. It applies to microphones such as mutalk [10] for VR metaverse users and PHASMA [11] for game chat. Despite their efficacy in reducing sound leakage, they do not provide a solution for speech constraints caused by internal factors. As a voice-generation approach, research on silent speech interaction is underway to enable speech communication in situations where audible speech cannot be used. Silent speech interaction is currently in the experimental phase and considers sensing at various stages of the human speech production system. Moreover, software such as Aquestalk2 from AQUEST Corp. (http://www.a-quest.com/) can convert text-based sentences into vocal information. For example, using this software, Hayashi et al. [12] generated rap lyrics from textual information. These technologies have many potential applications, are expected to help assist people with vocal organ disorders, and can be used in noisy environments.

However, vocal experiences constituted by generated voices often lack a sense of agency (SoA). Here, the term ‘‘sense of agency’’ is defined as ‘‘The sense that I am the one who is causing or generating an action’’ [13]. The SoA is critical for assuming responsibility for one's actions [14] and is involved in the presence of experience [15]. This sense plays a pivotal role in health [16], with its disturbances linked to conditions such as schizophrenia [17]. Additionally, in a rapidly evolving digital society, the design of interfaces is becoming increasingly important for human–computer interaction, human–human interactions, and VR. To forge ahead in developing innovative interfaces or refining existing ones, a profound understanding and incorporation of the agency user experience is imperative, as highlighted by Limerick et al. [18].

Therefore, we proposed creating a sense of vocal agency (SoVA) by presenting auditory and laryngeal vibratory stimuli [19]. This method aims to provide a pseudo-vocal experience with an SoA by vicariously presenting vocal-related sensations, even when the actual vocalization is suppressed or restricted. In a previous study [19], we demonstrated that the proposed method could create a specific vocal agency of loud vocalization for a person who vocalizes in a soft voice or whisper. However, it is still being determined whether the proposed method, which presents tactile stimuli that are physically stronger than the vibrations of normal vocalizations, can be applied to a general SoVA because normal vocalizations have less sensory experience than loud vocalizations.

This study aimed to assess the applicability of the proposed method for generating a generalized SoVA by applying vibratory stimuli to a specific body part, i.e., the larynx. The rationale behind this purpose was the possibility that the results of the previous study [19] may have been caused by a priming effect due to the presence of the vibratory stimuli. Priming refers to the tendency of an agent who had previous thoughts relevant to an action to attribute the agency of the action to oneself [20], [21]. Therefore, the vocal intention that preceded the vocal experience and the vibratory stimuli, which were temporally synchronized with the experience, may have caused the priming and created an SoVA. If priming alone is sufficient for creating an SoVA, then the proposed method can be replaced by presenting stimuli to body parts, such as the wrist, where vibrations are more easily presented. Therefore, it is essential to verify that priming alone does not cause SoVA in the proposed method.

Formulating a methodology enhancing vocal agency benefits users' communicative experiences in settings where speaking is conventionally constrained. This approach empowers users to maintain their distinctive vocal agency even in silent venues like trains and libraries, mitigating environmental concerns. The integration with VR technology further broadens its applicability, enabling participation in digital gatherings or conferences irrespective of the physical locale, provided minimal VR resources are accessible. Additionally, the approach holds particular expectations for individuals with speech difficulties, potentially facilitating daily interactions that closely mirror those experienced by individuals without such conditions.

II. RELATED WORKS

This chapter first describes the theories explaining SoA and their correspondence to vocalizations. Next, research on SoA with voice, which is the result of vocalization, is described.

A. SENSE OF AGENCY

1) COMPARATOR MODEL

The comparator model is renowned within classical frameworks for explaining SoA. This model encompasses a process for monitoring intentional actions, comparing the results of actions with those of external events, and distinguishing between them [22]. Numerous comparator models have been proposed, and the one that appears to contribute the most to SoA is the comparator that compares the state predicted by the motor system with the actual state estimated by sensations [23]. A negative correlation exists between the attributed sense of agency and the magnitude of discrepancy between these two states. The prediction by the motor system is based on an efference copy. The efference copy is generated from the motion signal and is compared with the reafference [24]. The sensory system's estimations encompass reafferent inputs, including external cues (e.g., visual and somatosensory feedback) and internal cues (e.g., proprioception). The sensations decline when these are considered to be self-generated, resulting in phenomena such as the inability to tickle oneself [25].

Vocalization necessitates the engagement of the respiratory system and laryngeal dynamics to facilitate the vibration of expelled air, notably through processes of abduction and adduction of vocal cords. The genesis of the efference copy in vocalization is attributed to the primary motor cortex, which is located in the precentral gyrus and responsible for voluntary control of the larynx, and from the brain areas responsible for vocal-related movements such as expiration and articulation [26]. In contrast, the sensory consequences of vocalization are derived from a conglomerate of systems encompassing auditory, proprioceptive, and somatosensory modalities. Voice resulting from vocalization is one of these and is used to control vocalization. One of the roles of voice is explained by delayed auditory feedback (DAF) [27]. In addition, as demonstrated by Tremblay et al. [28], somatosensory information is a significant component of speech targets.

2) MULTIFACTORIAL TWO-STEP ACCOUNT MODEL

The comparator model provides cues related to an SoVA, yet it fails to encompass instances of SoVA without vocalization. Individuals not engaged in vocalizing partly lack the efference copy pertinent to such actions, and the comparator detects significant mismatches when the vicarious sensory consequences of vocalizations are presented. Therefore, it is necessary to consider both predictive and retrospective inferences like apparent mental causation [20].

Upon synthesizing these inferences, Synofzik et al. [23] advanced the "multifactorial two-step account model." Within this framework, SoA bifurcates into the feeling of agency (FoA) and judgment of agency (JoA). FoA represents a low-level feeling of agency comprising cues mainly used as input for the comparator, such as sensory feedback. In contrast, JoA represents a conceptual judgment of agency, referring to intentions and external contextual cues.

B. SENSE OF AGENCY FOR VOICE

The SoA for voices resulting from vocalization has been examined. Franken et al. [29] revealed that even when listening to a voice modulated by 100 cents in real-time, the modulated voice was accepted as a reference signal. However, the modulated voice decreased the subjective SoA. Lind et al. [30] demonstrated that participants who heard their voice speaking something different from their actual utterance experienced the sounds they heard as self-generated. Similarly, Zheng et al. [31] showed that the coherence of vocal cord movements and the resulting sensory events might cause participants to classify a stranger's voice as their own perceptually. Collectively, these investigations affirm that variations in vocal output, whether in timbre or content, do not inherently disrupt the SoA.

Furthermore, Banakou and Slater [21] demonstrated that applying vibratory stimuli to the thyroid cartilage can induce an illusion of agency over vocalizations emitted by a virtual avatar. This finding was significant as it occurred without feedforward prediction, prior thinking, or uniqueness of reasoning [21]. However, this study was also focused on whether the participant produced the voice, and to the best of our knowledge, no prior studies have focused on the agency throughout the entire vocalization process.

C. LARYNGEAL VIBRATION OF VOCALIZATION

Vocal cord vibration emerges as a crucial element in vocal perception. Khosravani et al. [32] administered vibrotactile stimuli to the larynx of spasmodic dysphonia sufferers, modifying afferent proprioceptor inputs to the sensorimotor cortex responsible for speech production. The findings indicated the potential of laryngeal vibrotactile stimuli as a therapeutic intervention for spasmodic dysphonia. Thus, laryngeal vibratory perception is a factor of vocal control. Banakou and Slater [21] also applied vibratory feedback to the thyroid cartilage to enhance speaking sensation. Based on these observations, we hypothesized that laryngeal vibratory information, which constitutes vocal feedback and is considered specific to speech, contributes to the creation of SoVA.

III. SENSE OF VOCAL AGENCY

This study delineates the SoVA as an SoA pertinent to the vocal experience, encompassing both the vocal production mechanism and the consequent acoustic phenomena. This conceptualization is more similar to the framework of SoA as proposed by Synofzik et al. [23] than Gallagher's interpretation. Consequently, this formulation articulates a direct causal linkage between the act of vocalization and its resultant effects, thereby contributing to a more nuanced understanding of vocal agency.

The proposed method for presenting an SoVA employs two vicarious stimuli [19]. The first is laryngeal vibratory stimuli, presented as a proxy for the perception of vocal cord vibration as an action causing vocalization. The second is auditory stimuli representing the voice resulting from vocalization. These two stimuli are expected to reduce the difference between the desired and actual states inferred from sensory feedback, resulting in the creation of SoVA. This agency inference is supported by the perception of laryngeal vibration, which is unique to vocalization.

A. PRESENTING SENSE OF VOCAL AGENCY

This section describes the design of the auditory and vibratory stimuli (Fig. 1) presented to create an SoVA. These design guidelines are similar to those in our previous study [19]; a more detailed study of the design is presented in a previous paper.

1) AUDITORY STIMULI

We used participants' vocalizations of vowel /a/ sustained for approximately 4 s as auditory stimuli. This design is based on two reasons. First, the integration of air-conducted and bone-conducted sounds constitutes the auditory perception during speech [33]. It is known that the timbre of self-heard sounds

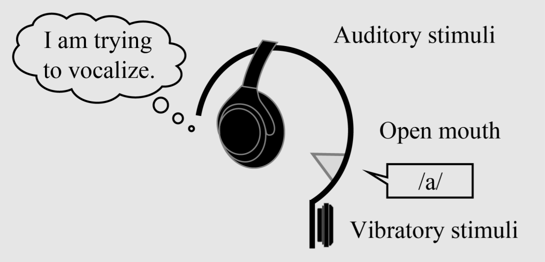


FIGURE 1. The proposed method for creating SoVA by auditory and laryngeal vibratory stimuli (adapted from [19]).

differs from that of recordings, and even recordings of one's voice can be perceived as disharmonious. Specifically, for all vowels, participants perceived their vocalizations as louder in the low-frequency range (approximately 5 dB at 100 Hz) and softer in the high-frequency range (approximately −5 dB at 4000 Hz) compared to the recorded sound [34]. It is considered impossible to implement a universal voice manipulation filter applicable to all people that can reduce this disharmony [35]. Therefore, we used the participant's vocalization of the vowel /a/, which is considered to have a significant resonance component and a relatively minor discrepancy between the self-hearing sound and the recording [33].

Second, the stimuli duration was determined based on previous studies [19], [29]. A vocalization period of 4 s was considered sufficiently long to create an SoVA and sufficiently short to maintain stable vocalization. A maximum error of 0.3 s was allowed for the voice recordings in the experiment. If this requirement was not satisfied, a voice was re-recorded. The volume of speech during the recording and the volume of the sound heard through headphones during the experiment were set as 58 dB, as the volume of a normal conversation is considered to be 50–65 dB [36].

2) VIBRATORY STIMULI

Human vocalization is initiated when exhaled air is vibrated by the vocal cords symmetrically located in the larynx. In this study, the transducers were placed symmetrically on the larynx to mimic the arrangement of vocal cords. If the vibration is strong enough, it is expected to propagate to the vicinity of vocal cords. In addition, vibratory stimuli that are temporally synchronized to nearby areas are expected to generate phantom sensations [37], resulting in a vibratory perception similar to actual vocalization.

In developing the vibratory waveform, the endeavor was to emulate the inherent vibrations of vocal cords during vocalizations. Vocalization is thought to be generated predominantly from dipolar sources at 125 Hz [38], [39]. However, the resonant frequency exhibits variability contingent upon the articulated vowel, alongside notable inter- and intra-individual disparities. Therefore, by presenting each individual's recorded voice as a vibratory waveform, we constructed vibratory stimuli that reflected individual differences with minimal discomfort.

IV. EXPERIMENTS

This experiment aimed to evaluate the efficacy of auditory and laryngeal vibratory stimuli in creating a general SoVA. Factors related to the presentation position of the vibratory stimuli were set to clarify the contribution of vibratory stimuli to an SoVA while considering priming. In addition, to confirm how a user's actions affect an SoVA, we set up factors related to the vocal style assumed by the proposed system.

A. EXPERIMENTAL DESIGNS

The experiment was conducted using a within-participant design consisting of two factors. Three vibratory stimuli conditions were used:

- "No-vibration condition," wherein participants only vocalize and are not presented with any vibratory stimuli;
- "Wrist condition," wherein vibratory stimuli are presented from the device on the wrist of the dominant hand at the same time as the participant's vocalizations; and
- "Larynx condition," wherein the device vibrates the participant's larynx as the participant vocalizes. This condition corresponds with our proposed method.

The "wrist condition" included investigating whether vibration applied to a specific body part related to vocalization, namely the larynx, contributes to the SoVA. The "larynx condition" and "wrist condition" are equivalent in that the vibratory stimuli are temporally synchronized with the vocalization, and we can thus detect the effect of vibratory stimuli except for the priming effect. The wrist was selected because it has hirsute skin [40] and is less likely to produce the sensation of active touch. This property is identical to that of laryngeal skin. Active touch should not be allowed because it provides a vastly different experience than passive touch [41]. Additionally, the study incorporated a "no-vibration condition" to verify the degree to which an SoVA is generated without vibratory stimuli.

Four vocal-style conditions were used as actions that participants performed during the experiment:

- "Soft voice condition," wherein participants vocalize the vowel /a/ under 42 dB while vibrating their vocal cords;
- "Whisper condition," wherein participants vocalize the vowel /a/ in whispers under 42 dB without vibrating their vocal cords;
- "Mouthing condition," wherein participants open their mouth as they would when vocalizing /a/; and
- "Imagining condition," wherein participants do not vocalize anything but imagine that they are vocalizing the vowel /a/ at 58 dB for approximately 4 s.

The pseudo-vocal experience targeted in this experiment is the vocalization of the /a/ sound at 58 dB for 4 s. The "soft voice condition" was identified as most closely approximating the target vocalization, utilized under circumstances requiring reduced volume. Whispering is one of the five voice qualities defined by the International Phonetic Association [42] and is not accompanied by vocal cord vibration. In voiced utterances, whispers provide the most quietness

TABLE 1. Various sensory correspondences between the four vocal styles and the pseudo-vocal experience that the experiment is aimed at. The circles indicate that the sensations are in general agreement.

	Vocal styles			
	Imagining	Mouthing	Whisper	Soft voice
Image of vocalization	○	○	○	○
Proprioception		○	○	○
Oral sensation			○	○
Vocal cord vibration				○

to ensure privacy during communication. The acoustic characteristics of whispers differ from those of normal vocalizations [43]. Consequently, the "whisper condition" diverges from the "soft voice condition" regarding vocal cord vibration and acoustic output. The "mouthing condition" is a condition wherein lip movements (proprioception) are synchronized with the vocal experience. Because mouthing does not involve actual pronunciation, the sensations of breathing and touch inside the mouth (oral sensation) differ from actual vocalization. In the "imagining condition," the image of vocalization is temporally synchronized with the vocal experience. Participants were requested to maintain this image in all four styles. The characteristics of each style are summarized in Table 1. The circles in the table indicate the approximate agreement between the actual vocalization and the sensation for each style.

B. MATERIALS

1) EQUIPMENT

The experiment was conducted in an experimental area (width: 2.4 m, length: 2 m, and height: 3 m), separated by partitions. Sound level measurements regarding vocalizations were facilitated using a precision sound level meter LA-4441A (Ono Sokki Co., Ltd., Japan). These measurements were conducted with the experimenter positioning the device approximately 1 m away from the participants' mouths. Measurements were performed to maintain the participants' vocalizations within the range of normal vocalizations, and the physical quantities reported were not precise.

Vocalizations were recorded using an ATR2100x-USB microphone (Audio-Technica Corporation, Japan), and the Audacity software was used to save the digital recordings. The audio was digitized to 32 bit/44.1 kHz. The noise-canceling headphones WH1000-XM4 (Sony Corporation, Japan) connected to an audio interface UR-RT4 (Steinberg Media Technologies GmbH, Germany) were used to control the audio.

Participants were equipped with two devices that administer vibratory stimuli implemented using a vibro-transducer Vp210 (Acouve Laboratory, Japan) on their wrist and larynx (the design and evaluation are presented in Section IV-B2). These vibrators were manipulated via the audio interface and the amplifier LP-2024A (Lepy, China). Displacement resulting from the stimuli was quantitatively assessed using a laser displacement meter LK-G5000 and LK-H055 (KEYENCE CORPORATION, Japan).

2) DEVICES FOR VIBRATORY STIMULI

The two implemented devices are shown in Fig. 2. The laryngeal device incorporates two transducers, symmetrically applied to either side of the larynx at the vocal cords' level. In contrast, the wrist device utilizes a single transducer positioned against the back of the dominant hand's wrist. The rubber material is used as a cushioning material between the connectors and transducers to reduce the propagation of vibrations to non-targeted anatomical regions.

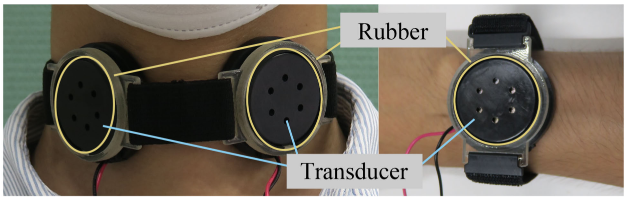


FIGURE 2. Appearance of wearing the laryngeal device (left) and wrist device (right).

Device efficacy was assessed utilizing the laser displacement meter within a specifically constructed measurement environment to replicate vibratory stimuli on the larynx (Fig. 3). An acrylic pipe was covered with an artificial skin sheet, and the laryngeal device was coiled on the sheet. The thickness of the sheet was 3 mm, the hardness was 10–15, and the diameter of the pipe was 130 mm. Displacement metrics were garnered from two critical points: (A) the center of the transducer and (B) a location approximately 2 mm from the transducer's peripheral boundary.

The Fourier transform analysis of the measured waveforms (Fig. 3) indicated that the device could replicate the input waveform's frequency and that the vibrations at the fundamental frequency (F0) of the input waveform propagated through the sheet. However, the device was limited in reproducing input waveform frequencies above 500 Hz. This tendency is consistent with the frequency characteristics of the Vp2 transducer as documented in a previous study [44]. Furthermore, it was observed that vibrations exceeding 200 Hz barely propagated through the sheet. However, although the vibration was presented perpendicular to the seat, the measurement was implemented at a horizontal distance from the transducer; therefore, it is unclear to what extent high-frequency vibration propagates when the device is attached to a person.

3) CONTROLLING THE TIMING OF VOCALIZATION

In the experiment, the timing of the vocalization was taught to participants using a graphical user interface (GUI) (shown in Fig. 4) to match the timing of vocalization with that of the stimuli. Participants could start the GUI by pressing a button at any time. When the GUI was activated, a black indicator slid from left to right, and participants followed the instructions to calibrate their breathing, inhalation, and vocalization.

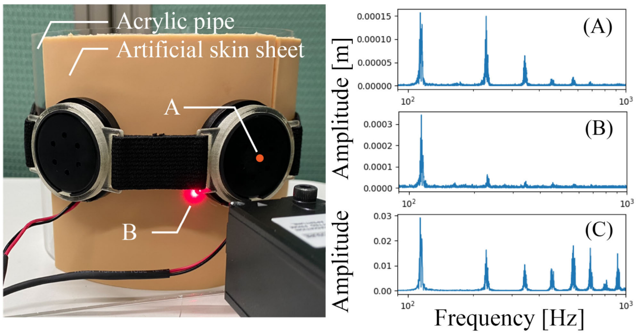


FIGURE 3. Appearance of vibratory measurement with the laser displacement meter (left) and the result of Fourier transform analysis (right). Graphs (A) and (B) present the FFT results of the displacements measured at the center of the transducer and the rubber surface near the transducer, respectively. Graph (C) presents the FFT results of the input audio signal.

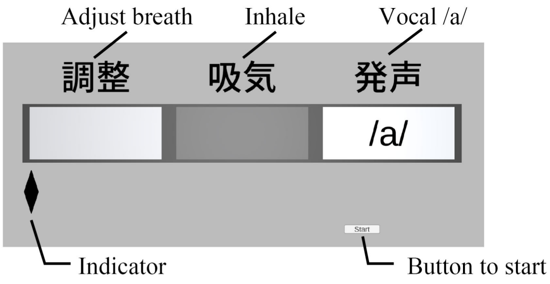


FIGURE 4. Graphical user interface used to teach vocal timing to participants.

C. SUBJECTIVE RATINGS

Before the experiment, participants completed a questionnaire to record basic information such as gender and age. All the participants were confirmed to have no hearing impairment.

Upon completion of each trial within the experiment, participants were requested to fill out a questionnaire (shown in Table 2). The questionnaire was designed with four questions, focusing on the experiential quality of vocalization. Two related to the quality of the vocal experience: vocal sensation (Q1) and vocal agency (Q2). Vocal sensation, which corresponds to the FoA defined by Synofzik et al. [23], is a nonconceptual feeling of vocalization. In contrast, the SoVA is construed as an overarching SoA for the entire vocal act, obtained after the judgment for attribution of agency. The third question (Q3) explored participants' identification with the heard voices as their own. The final question (Q4) probed whether participants felt as if they had vibrated their larynx. Participants were asked to respond to Q1 using a visual analog scale (VAS) (0 = not at all; 1 = same as actual vocalization) and Q2–Q4 on a seven-point Likert scale (1 = not at all; 7 = strongly). In instances of non-perception of laryngeal vibration, participants were instructed to denote "no-vibration" instead of using the Likert scale in Q4.

Participants were also asked to indicate the degree of discomfort they experienced while wearing the devices on a seven-point Likert scale.

D. PROCEDURE

1) INFORMED CONSENT AND VOCAL RECORDING

Participants first read experimental instructions and signed a consent form if they agreed to participate. Subsequently, they recorded their voice while vocalizing the vowel /a/ for approximately 4 s at a target intensity of 58 dB. Before the recording, the experimenter facilitated a brief training session to instruct participants on maintaining a consistent vocal volume at 58 dB. The preparatory practice was deliberately confined to the period before the recording session. It was because requiring volume adjustment with volume feedback during recording would affect the acoustics of the voice, interfere with natural vocalization, and impair the sensation of vocalization.

2) DECIDING THE INTENSITY OF VIBRATORY STIMULI

In this experiment, two stimuli intensities were designed so that each participant perceived the vibratory intensities at the larynx and wrist to be similar. This is because skin's sensitivity to vibrations differs among individuals, and the difference in the sensitivity of skin at the wrist and larynx is also considered to differ among individuals. Consequently, the perceptual threshold for vibration at the wrist of each participant was established as a benchmark for adjusting stimuli intensity. The following two-step procedure was implemented for each participant to calibrate the vibratory intensities.

Initially, participants were tasked with identifying their threshold levels for wrist vibratory stimuli using an adjustment method. This process was conducted three times, and 3.6 times the mean of the measured values was applied as the intensity of the wrist vibratory stimuli. Subsequently, to determine the point of subjective equality where the two stimuli intensities were perceived to be of equal magnitude, the intensity of the laryngeal vibratory stimuli was adjusted using an adjustment method in response to the determined wrist stimuli. This determination was repeated three times, and the mean of the measured values was employed to set each participant's laryngeal vibratory intensity.

The adoption of a 3.6 times intensity in the current experiment was informed by outcomes from a preliminary study involving 21 participants. This initial investigation sought to evaluate the SoVA with 1.2 times, 2.4 times, and 3.6 times vibratory intensities. The findings indicated no difference between the tested vibratory intensities. Consequently, the 3.6 times intensity was selected for its superior strength among the options, facilitating easier detection of vibrations by participants.

3) PRACTICE FOR EACH VOCAL CONDITION

The volume levels of the recorded voices were calibrated using Python so that the overall root mean square value of the signal could be controlled among participants. Participants were then acquainted with and practiced the four distinct vocal styles (soft voice, whisper, mouthing, and imagining). Participants were reminded not to vibrate their vocal cords

TABLE 2. Four questions answered regarding each trial of the experiment. Participants were asked to answer the first question on the VAS and the remaining questions on a seven-point Likert scale.

Variable names	Questionnaire statements
Q1. Vocal-sensation	It felt as if I got the same sensation as when I actually vocalized at normal volume.
Q2. Vocal-agency	It felt as if I was the one who vocalized the voice, and I felt the vocal experience.
Q3. Own-voice	It felt as if the voice I heard was my own voice.
Q4. Vibratory-agency	It felt as if I vibrated my larynx.

when whispering and to maintain vocalizations under 42 dB for whispering and the soft voice.

After the practice session, participants were instructed on utilizing the GUI, as shown in Fig. 4, and they practiced vocalization based on the GUI after wearing the device. Each of the four vocal styles was executed twice, culminating in eight iterations. The initial phase of the session consisted of the "no-vibration condition," and they practiced sustaining each vocal style for 4 s according to the GUI. In this phase, we repeated the practice as required. The session's latter phase consisted of the "wrist/larynx condition." The combination and sequence of the vocal and vibratory conditions were counterbalanced to prevent order effects. This series of practice sessions would help participants become accustomed to the recorded voice and thus stabilize the generation of an SoVA in the experiment. Finally, participants performed normal vocalizations to confirm the sensation that served as the basis for their responses in subsequent trials.

4) PRESENTING SENSE OF VOCAL AGENCY

Participants were seated and observed the GUI on a monitor, wearing the devices (Fig. 5). Participants could interact with the GUI at any time to commence the trials. Upon receiving cues from the GUI, they executed the instructed vocal actions and were presented with the stimuli concurrently. Under all conditions, participants were asked to imagine a normal vowel /a/ vocalization. After each trial, participants answered the questionnaire, as shown in Table 2.

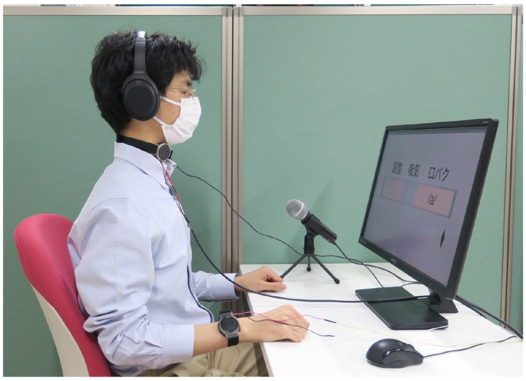


FIGURE 5. Appearance of the experiment.

The experimental framework consisted of 24 trials (three conditions for vibratory stimuli × four conditions for vocal styles × two repetitions). The arrangement of trials was designed to prevent order effects using a Balanced Latin Square [45].

TABLE 3. Statistical tests used for analysis.

Test type	Parametric test	Non-parametric test
Normality	Shapiro-Wilk test	-
Homoscedasticity	Bartlett test	Fligner-Killeen test
ANOVA	ANOVA	ART-ANOVA [46]
Multiple Comparison	Paired T-test	Wilcoxon signed-rank test

E. STATISTICAL ANALYSIS

Statistical tests were conducted utilizing Python and R libraries, and the significance level alpha was established at 0.05. The R library was employed for the analysis of variance (ANOVA). For nonparametric data, ART-ANOVA [46] was used. ART represents a transformation process that facilitates the application of the ANOVA to data in which normality cannot be assumed. Other tests were performed using Python software. The Benjamini & Hochberg correction was used for multiple comparisons after the ANOVA to adjust the significance level. Table 3 lists the different test methods used depending on whether the data has normality.

Q4 was not answered using a simple Likert scale; some responses indicated that the participants did not perceive any vibration in their larynx. Thus, we did not perform ANOVA or other statistical tests for Q4. The discomfort experienced with each device, as identified in the questionnaire at the end of the experiment, was analyzed in conjunction with the vocal agency (Q2). The median responses to Q2 for each participant were calculated, and Spearman's correlation coefficient was calculated along with the ratings of discomfort with the laryngeal device to investigate the possibility that discomfort with the device had a significant effect on the sensation in the experiment.

F. PARTICIPANTS

The power analysis tool PANGEA [47] was used to determine the necessary sample size. Cohen's d from a previous study [19] was calculated using the following equation (Eq. 1), which represents the effect size for comparing two corresponding groups. The effect size of the vibratory factor has been widely estimated to be 0.16–1.07. The sample size of this study was 20. This sample size was considered appropriate because a medium effect size of 0.6 could be detected with a power of 0.817.

$$\mathrm{d} = \frac{\mathrm{mean}(x) - \mathrm{mean}(y)}{\sqrt{\frac{\mathrm{std}(x)^2+\mathrm{std}(y)^2}{2}}} \tag{1}$$

The participants were recruited via a mailing list. Twenty-two individuals participated in the experiment; however, the analysis excluded two due to missing or incomplete data. The participants consisted of 13 males, six females, and one undisclosed gender, and their average age was 25.35 years old (SD = 7.00).

Each participant received 1080 yen for participating in the experiment. This study was approved by the Research Ethics Committee of the University of Tokyo (No. 21-92). Each participant provided written informed consent before the experiment.

V. RESULTS

A. ANALYSIS OF RECORDED VOICES

The mean fundamental frequencies of vocalizations recorded were 133.2 Hz (SD = 25.8 Hz) and 215.8 Hz (SD = 13.3 Hz) for males and females, respectively. Because the fundamental frequency of vocalization of Japanese aged 14 years or older is approximately 120 Hz for males and 240 Hz for females [48], [49], the F0 of the participants in this study was generally considered average. The average duration of the recorded voices was 4.12 s (SD = 0.15 s).

B. ANSWERS FOR THE QUESTIONNAIRE

Fig. 6 presents the aggregated outcomes for responses to each question. Except for Q4, all the results were averaged over two replicates for each condition. Due to data size and correspondence disparities, Q4 was not averaged over the replicates, and the box plots only represent the responses on the Likert scale.

1) HYPOTHESIS TESTS FOR VOCAL EVALUATION (Q1, Q2) AND VOICE TIMBRE (Q3)

A normality test was conducted on responses to Q1, which utilized a parametric VAS, and only one of the 12 null hypotheses was rejected. Therefore, a parametric test was deemed suitable for Q1. The homoscedasticities between the 12 groups for each of Q1-Q3 were tested, and the homoscedasticities were not rejected for any of the questions; therefore, there were no problems in applying the tests listed in Table 3. The results of the ANOVA conducted for Q1–Q3 are presented in Table 4. The results show that the interaction between the two factors was insignificant for any of the questions and that the main effect of the vocalization factor was significant. Significant differences were observed in the vibration factors for the two questions, Q1 and Q2. Table 5 lists the results of multiple comparisons conducted as a post-test for the factors for which significant differences were found.

TABLE 4. Results of ANOVA for the questions. A general ANOVA is used for vocal sensation, and ART-ANOVA is used for the others.

	Vocal sensation	Vocal agency	Own voice
Vibration	🟧 $F_{2,38} = 6.67$	🟩 $F_{2,209} = 3.17$	$F_{2,209} = 0.16$
Vocalization	🟧 $F_{3,57} = 33.0$	🟧 $F_{3,209} = 78.8$	🟧 $F_{3,209} = 10.6$
Two-way interaction	$F_{6,114} = 1.35$	$F_{6,209} = 0.53$	$F_{6,209} = 0.36$

🟧: $p < 0.01$, 🟩: $p < 0.05$, □: n.s.

TABLE 5. Results of multiple comparisons for Q1, Q2, and Q3.

	Vocal sensation	Vocal agency	Own voice
None	0.33	2.5	
Wrist	0.36	3	
Larynx	0.39	3.5	
Imagine	0.23	2.3	4.5
Mouth	0.3	3	5
Whisper	0.28	2.5	4.5
Soft voice	0.63	5	6

$**$: $p < 0.01$, $*$: $p < 0.05$, +: $p < 0.10$

2) TOTALIZATION FOR VIBRATORY QUESTION (Q4)

Table 6 delineates the probabilities of participants reporting the absence of vibratory sensations in their larynx. Participants answered "no-vibration" in less than 30% of the trials, even when vocal cords were not actually vibrated or vibration was not presented to the larynx. Analyzing the data for each individual, we found that 13 participants answered "no-vibration" in only 10% or less of the trials, five in under 30%, and the remaining two in under 50%. This trend may be attributable to participants' limited knowledge about the vibratory stimuli, predisposing them to anticipate vibrations in all conditions. Such an assumption might have forced the participants to respond that they perceived the vibration even though they did not perceive it. Furthermore, the assumption could have created a virtual vibratory perception of the larynx. Under the "soft voice condition," participants reported perceiving the vibration in almost all trials. It may be due to the inference that vibrations occurred in the larynx because they intentionally vibrated their vocal cords.

TABLE 6. Probabilities of the participants had answered that they did not perceive any vibration on their larynx for each condition.

	None	Wrist	Larynx
Imagine	30	22.5	12.5
Mouth	27.5	20	17.5
Whisper	20	22.5	22.5
Soft voice	5	5	5

3) ANALYSIS FOR DISCOMFORT TOWARD DEVICES

Fig. 7 delineates the results of discomfort ratings associated with the laryngeal and wrist devices. Notably, the aggregate response count diverges because one participant did not respond to the question regarding the wrist device. The figure shows that the discomfort with the wrist device was generally low, with approximately half of the participants reporting minimal to no discomfort. In contrast, discomfort associated with the laryngeal device was dichotomized at an approximate rating of 4, which confirms that approximately

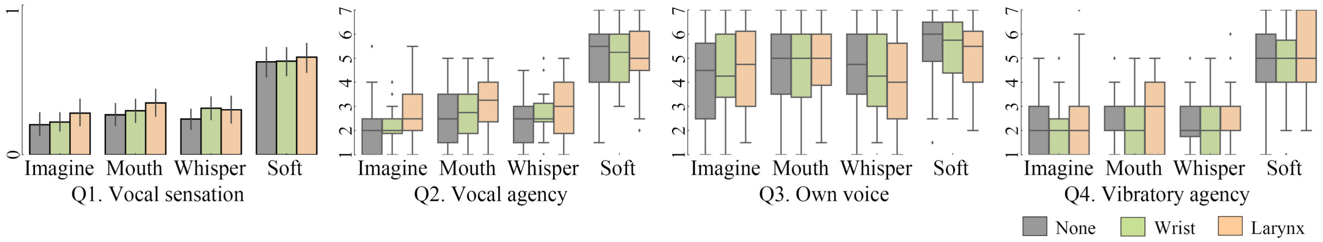
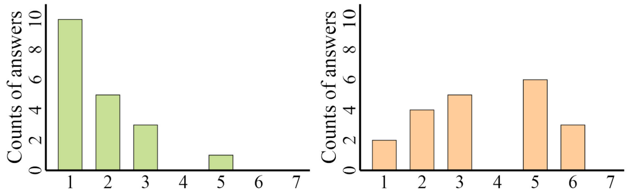


FIGURE 6. Results of vocal sensation, vocal agency, own voice, and vibratory agency. The error bars indicate standard errors.

half the participants felt slightly more significant discomfort. The correlation coefficient between the median of vocal agency (Q2) and discomfort caused by the laryngeal device was 0.043, indicating no significant relationship (p=0.86). Therefore, this issue requires further attention and a solution; however, the device does not significantly affect the results of this experiment.



(a) Discomfort for wrist device.
(b) Discomfort for larynx device.

FIGURE 7. Results of answers on a seven-point Likert scale regarding the degree of discomfort they experienced with each of the laryngeal and wrist devices.

VI. DISCUSSION

The study aimed to test the role of laryngeal vibratory stimuli in creating an SoVA. The underlying hypothesis was that stimuli to the larynx, a specific body part associated with vocalization, would aid in making inferences regarding vocal agency. Another purpose of the study was to investigate the differences in sensation among the four vocal styles.

A. ANALYSIS OF VOCAL EVALUATION (Q1 AND Q2)

As delineated in Table 5, there was a significant difference in only one pair of the vibratory factor for both Q1 and Q2; the "larynx condition" was rated significantly higher than the "no-vibration condition." This outcome underscores that vibratory stimuli to the wrist, synchronized with vocalizations, were insufficient to cause an SoVA. This suggests that the SoVA occurs through mechanisms other than the priming effect implied by previous research [19], [21]. A stronger SoVA is achieved when there is alignment between the anticipated and the actual vibratory sensations experienced in the vicinity of the vocal cords.

Multiple comparisons of the vocal styles revealed nearly identical significant differences between Q1 and Q2 (Table 5). It was expected that the "soft voice condition" would be evaluated as significantly higher than the other three styles in both metrics because the desirable state and the afferent and efference copy are most likely to match in this condition. Nonetheless, environmental constraints predominantly limit a soft voice's applicability. Among the three vocal styles, except for the "soft voice condition," a significant difference was found between the "imagining condition" and "mouthing condition" in Q1 (Table 5). This significantly higher evaluation in the "mouthing condition" infers that proprioceptive feedback associated with vocalization potentially enhances the realism of the vocal experience. In fact, Sugimori et al. [50] found that participants were more inclined to believe they had vocalized a voice presented in the "mouthing condition" rather than the "imagining condition." The same tendency was observed in the SoVA (Q2); however, the difference was insignificant. This lack of significance might be attributed to the vibratory stimuli overshadowing the influence of proprioceptive feedback despite its potential contribution to SoVA. Consequently, the contribution of proprioception to the SoVA could not be determined.

Whispering seems to be the most practical of the four vocal styles, attributed to its quietness, high information transmission ability, and the existence of speech-conversion technologies such as WESPER [51]. However, whispering tended to be evaluated almost identically to imagining and mouthing in both Q1 and Q2. This may be mainly due to the following three factors: (i) the low effect of breathing and oral sensation associated with actual vocalization on the SoVA, (ii) the discomfort of intention and hearing a voice that is qualitatively different from the voice that is actually being uttered, and (iii) the discomfort of being presented with laryngeal vibration despite the clear intention of not vibrating the vocal cords.

B. ANALYSIS OF VOICE TIMBRE (Q3)

The analysis revealed no statistically significant differences across the vibratory stimuli conditions and only a weak significant trend in the vocal styles concerning the quality of the voice heard through headphones (Table 5). The participants appeared to hear themselves more in the "soft voice/mouthing condition" than in the other two conditions. Although the soft voice seemed the most favorable in this question, no significant differences existed between the "soft voice condition" and the others. This lack of significance could stem from the acoustic differences between soft voices and normal vocalization. In fact, one participant reported disharmony between the voice through the headphone and the intended voice quality. The preferential evaluation in the "mouth condition" over the "imagining condition" suggests that proprioceptive feedback consistent with vocalization may influence the voice perception evaluation. The tendency of the diminished evaluation in the "whisper condition" suggests that the distinct auditory characteristics of whispering, devoid of clear vocal cord vibration, may introduce a significant qualitative disharmony between the recorded voice and the intended sound.

It should be noted that participants were exposed to identical auditory stimuli across 24 trials, introducing a potential habituation effect to the sound. Furthermore, as elaborated in Section II-B, the agency of voices is easily generated. Consequently, it is possible that an SoA to the voice was generated in many trials and that the SoA hid the disharmony with the voice caused by the vocal style.

C. ANALYSIS OF VIBRATORY PERCEPTION (Q4)

Regarding the sense of vibratory agency, notably higher evaluation tended to be obtained only in the "soft voice condition," wherein the vocal intention and the sensory feedback almost coincided and participants involved active vocal cord vibration. Furthermore, no remarkable differences were observed in the other vocalization and vibratory conditions (Table 5). In addition to the existence of a bias mentioned in Section V-B2, the following may explain the absence of pronounced differences.

Initially, Q4 explicitly asked participants to respond to the sensation of laryngeal vibration alone, and it is thus possible that the question itself drew the participants' attention to the device and caused them to clearly distinguish the vibration of the device from the vibration of the vocal cords. This demarcation undermines the exclusivity necessary for the conscious will [20] and, as a result, prevents a sense of agency. Secondly, this question was the last item in the questionnaire and was answered after a posteriori evaluation of wide sensations, which may have obscured the sensations obtained from experience. Lastly, the tactile perception attributed to the device might be perceived only on the skin surface of the neck. This perception's variability could be attributed to interactions between the thickness and adiposity of the participant's neck skin and the frequency of vibrations from the device. As Rombout and Postma-Nilsenova [52] indicated, external vibratory perceptions during vocalization may detract from the experiential quality. Consequently, developing a vibratory stimuli presentation method that surpasses existing systems' reliability and stability is necessary. This method should ensure all participants can internally perceive the presented vibratory stimuli. Subsequently, it is critical to devise an improved experimental framework to evaluate the vibratory agency precisely.

D. SUMMARY

This investigation demonstrates that auditory and vibratory stimuli to the larynx effectively augment the general SoVA. The significant enhancement of SoVA through laryngeal vibratory stimuli, as opposed to the inadequate response from wrist vibration, validates the proposed method's suitability. However, the current experimental framework needs to be revised to definitively ascertain the degree to which externally applied vibratory stimuli on the larynx are internally perceived, akin to sensations from the vocal cords. This shortcoming primarily arises from the substantial impact of the "no-vibration condition" on participants' subjective evaluation of vibration. Therefore, to advance the understanding of vibratory stimuli perception, future investigations are encouraged to apply vibratory stimuli for assessment consistently.

The evaluation revealed that a soft voice resembling natural vocalization received the highest preference. Nevertheless, there are limitations to integrating a soft voice into the proposed system. Among the other three vocal styles, mouthing is preferred because it can be used in more situations than a soft voice, and it is less likely to cause discomfort during the vocal experience and the vibratory stimuli to the vocal cords. The "imagining condition," devoid of spontaneous action, did not underperform more than the "mouthing/whispering conditions" and may be applicable in the system, albeit the most challenging implementation.

VII. LIMITATIONS AND FUTURE WORK

The current study provided participants with a constrained vocal experience to assess the basic SoVA. Thus, participants were instructed to vocalize a predetermined vowel at a designated time. As Barlas and Obhi [53] suggested, the SoA is undermined when the degrees of freedom of the act are low, and the SoVA in this experiment can also be undermined. This undermined agency may have made the apparent effect sizes of experimental factors smaller than they actually were, thus reducing the power of the statistical tests. Therefore, future research should be conducted to construct and validate a system that provides a freer vocal experience. Furthermore, given the proposed method's intended application for daily use, subsequent research should explore the habituation effects associated with prolonged utilization.

Distinct specifications emerge when developing a real-time system to simulate pseudo-vocal experiences, depending on the input interface. A critical specification involves the system's latency from input to output. Specifically, for inputs through actual vocalization, auditory stimuli should not inhibit vocalization by the DAF. As DAF generally occurs when the feedback is delayed by 50 ms to 100 ms, the response of the system should be less than 50 ms. Conversely, for textual inputs, the awareness corresponding to the SoA can be substituted by presenting the prime immediately before the system-generated vocal experience; therefore, the response time is considered less constrained.

Should the system's input interface not encompass a soft voice, leveraging real-time cord vibrations or the voice generated by the user as vibratory stimuli becomes unfeasible. Therefore, it is necessary to devise a methodology to design the optimal vibratory stimuli bespoke to each user and possibly a methodology that can be applied uniformly to many people. The possibility of generating vibratory stimuli corresponding to each of the various tones of free speech or the possibility that a single vibratory stimulus alone is sufficient to create an SoVA should be considered. In addition, vocal cord vibrations produced during vocalization propagate over a wide area from the head to the abdomen [33], and vibrations at these body parts may also be perceived in association with vocalization. Therefore, further investigation is needed to determine the position of the vibratory stimuli that can assist in SoVA inference.

VIII. CONCLUSION

This study explored the impact of vibratory stimuli on the SoVA in the case of general vocalizations. Laryngeal vibratory stimuli and auditory stimuli are confirmed to be applicable and valid methods for general SoVA. Specific vibratory stimuli to the larynx, a body part related to vocalization, are found to contribute to the SoVA in a manner different from that of priming. This is the starting point for a method to extend and enrich vocal acts, which are closely related to and have various effects on daily life. This method has the potential to be applied to improve various interaction techniques and to support people with dysphonia. It was suggested that mouthing is the best input for the system for creating an SoVA, and if not possible, then a soft voice can be used. Although it has been confirmed that a soft voice provides the most favorable sensation regarding vocal agency and voice disharmony, it cannot circumvent many of the restrictions that vocal acts are subject to. However, mouthing, which does not produce actual sounds, was found to be sufficient for creating a certain SoVA.

In the future, it will be necessary to construct a real-time system for vocal agency using this method and verify the benefits of its application to the user. Furthermore, a method for creating a strong SoVA for special vocalizations, such as loud vocalizations and whispers, is expected to facilitate a more immersive and stress-relieving experience.

REFERENCES

[1] C. M. MacLeod, N. Gopie, K. L. Hourihan, K. R. Neary, and J. D. Ozubko, "The production effect: Delineation of a phenomenon," *J. Experim. Psychology: Learn., Memory, Cognition*, vol. 36, no. 3, pp. 671–685, 2010, doi: 10.1037/a0018785.

[2] M. Ikai and A. H. Steinhaus, "Some factors modifying the expression of human strength," *J. Appl. Physiol.*, vol. 16, no. 1, pp. 157–163, Jan. 1961, doi: 10.1152/jappl.1961.16.1.157.

[3] S. H. Kurniawan and A. J. Sporka, "Vocal interaction," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.* Florence, Italy: Association for Computing Machinery, 2008, pp. 2407–2410, doi: 10.1145/1358628.1358695.

[4] M. P. Aylett, P. O. Kristensson, S. Whittaker, and Y. Vazquez-Alvarez, "None of a CHInd: Relationship counselling for HCI and speech technology," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.* Toronto, ON, Canada: Association for Computing Machinery, Apr. 2014, pp. 749–758, doi: 10.1145/2559206.2578868.

[5] J. Wang, A. Chellali, and C. G. L. Cao, "A study of communication modalities in a virtual collaborative task," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2013, pp. 542–546, doi: 10.1109/SMC.2013.98.

[6] B. Shneiderman, "The limits of speech recognition," *Commun. ACM*, vol. 43, no. 9, pp. 63–65, Sep. 2000, doi: 10.1145/348941.348990.

[7] K. W. Altman, C. Atkinson, and C. Lazarus, "Current and emerging concepts in muscle tension dysphonia: A 30-month review," *J. Voice*, vol. 19, no. 2, pp. 261–267, Jun. 2005, doi: 10.1016/j.jvoice.2004.03.007.

[8] I. R. Titze, "Mechanical stress in phonation," *J. Voice*, vol. 8, no. 2, pp. 99–105, Jun. 1994, doi: 10.1016/s0892-1997(05)80302-9.

[9] J. A. Wilson, I. J. Deary, A. Millar, and K. Mackenzie, "The quality of life impact of dysphonia," *Clin. Otolaryngology Allied Sci.*, vol. 27, no. 3, pp. 179–182, Jun. 2002, doi: 10.1046/j.1365-2273.2002.00559.x.

[10] Shiftall *Mutalk*. Accessed: Jun. 1, 2024. [Online]. Available: https://en.shiftall.net/products/mutalk

[11] METADOX. *The Phasma*. Accessed: Jun. 1, 2024. [Online]. Available: https://metadox.pro/

[12] M. Hayashi, S. Bachelder, M. Nakajima, and Y. Shishikui, "Rap music video generator: Write a script to make your rap music video with synthesized voice and CG animation," in *Proc. IEEE 6th Global Conf. Consum. Electron. (GCCE)*. Nagoya, Japan: IEEE, Oct. 2017, pp. 1–2, doi: 10.1109/GCCE.2017.8229189.

[13] S. Gallagher, "Philosophical conceptions of the self: Implications for cognitive science," *Trends Cognit. Sci.*, vol. 4, no. 1, pp. 14–21, Jan. 2000, doi: 10.1016/s1364-6613(99)01417-5.

[14] R. Legaspi and T. Toyoizumi, "A Bayesian psychophysics model of sense of agency," *Nature Commun.*, vol. 10, no. 1, p. 4250, Sep. 2019, doi: 10.1038/s41467-019-12170-0.

[15] G. Herrera, R. Jordan, and L. Vera, "Agency and presence: A common dependence on subjectivity?" *Presence*, vol. 15, no. 5, pp. 539–552, Oct. 2006, doi: 10.1162/pres.15.5.539. https://doi.org/10.1162/pres.15.5.539

[16] P. Haggard, "Sense of agency in the human brain," *Nature Rev. Neurosci.*, vol. 18, no. 4, pp. 196–207, Apr. 2017, doi: 10.1038/nrn.2017.14.

[17] J. W. Moore, "What is the sense of agency and why does it matter?" *Frontiers Psychol.*, vol. 7, pp. 1–9, Aug. 2016, doi: 10.3389/fpsyg.2016.01272.

[18] H. Limerick, D. Coyle, and J. W. Moore, "The experience of agency in human-computer interactions: A review," *Frontiers Human Neurosci.*, vol. 8, pp. 1–10, Aug. 2014, doi: 10.3389/fnhum.2014.00643.

[19] Y. Shimomura, Y. Ban, and S. Warisawa, "Presenting sense of loud vocalization using vibratory stimuli to the larynx and auditory stimuli," in *Proc. 27th ACM Symp. Virtual Reality Softw. Technol.* Osaka, Japan: Association for Computing Machinery, Dec. 2021, p. 10, doi: 10.1145/3489849.3489891.

[20] D. M. Wegner and T. Wheatley, "Apparent mental causation: Sources of the experience of will," *Amer. Psychologist*, vol. 54, no. 7, pp. 480–492, 1999, doi: 10.1037/0003-066x.54.7.480.

[21] D. Banakou and M. Slater, "Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 49, pp. 17678–17683, Dec. 2014, doi: 10.1073/pnas.1414936111.

[22] C. Frith, "The self in action: Lessons from delusions of control," *Consciousness Cognition*, vol. 14, no. 4, pp. 752–770, Dec. 2005, doi: 10.1016/j.concog.2005.04.002.

[23] M. Synofzik, G. Vosgerau, and A. Newen, "Beyond the comparator model: A multifactorial two-step account of agency," *Consciousness Cognition*, vol. 17, no. 1, pp. 219–239, Mar. 2008, doi: 10.1016/j.concog.2007.03.010.

[24] E. von Holst, "Relations between the central nervous system and the peripheral organs," *Brit. J. Animal Behaviour*, vol. 2, no. 3, pp. 89–94, Jul. 1954, doi: 10.1016/s0950-5601(54)80044-x.

[25] L. Weiskrantz, J. Elliott, and C. Darlington, "Preliminary observations on tickling oneself," *Nature*, vol. 230, no. 5296, pp. 598–599, Apr. 1971, doi: 10.1038/230598a0.

[26] M. Belyk and S. Brown, "The origins of the vocal brain in humans," *Neurosci. Biobehavioral Rev.*, vol. 77, pp. 177–193, Jun. 2017, doi: 10.1016/j.neubiorev.2017.03.014.

[27] B. S. Lee, "Some effects of side-tone delay," *J. Acoust. Soc. Amer.*, vol. 22, no. 5, pp. 639–640, Sep. 1950, doi: 10.1121/1.1906665.

[28] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, no. 6942, pp. 866–869, Jun. 2003, doi: 10.1038/nature01710.

[29] M. K. Franken, R. J. Hartsuiker, P. Johansson, L. Hall, and A. Lind, "Speaking with an alien voice: Flexible sense of agency during vocal production," *J. Experim. Psychology: Human Perception Perform.*, vol. 47, no. 4, pp. 479–494, Apr. 2021, doi: 10.1037/xhp0000799.

[30] A. Lind, L. Hall, B. Breidegard, C. Balkenius, and P. Johansson, "Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say," *Psychol. Sci.*, vol. 25, no. 6, pp. 1198–1205, Jun. 2014, doi: 10.1177/0956797614529797.

[31] Z. Z. Zheng, E. N. MacDonald, K. G. Munhall, and I. S. Johnsrude, ''Perceiving a stranger's voice as being one's own: A 'rubber voice' illusion?'' *PLoS ONE*, vol. 6, no. 4, Apr. 2011, Art. no. e18655, doi: 10.1371/journal.pone.0018655.

[32] S. Khosravani, A. Mahnan, I.-L. Yeh, J. E. Aman, P. J. Watson, Y. Zhang, G. Goding, and J. Konczak, ''Laryngeal vibration as a non-invasive neuromodulation therapy for spasmodic dysphonia,'' *Sci. Rep.*, vol. 9, no. 1, p. 17955, Nov. 2019, doi: 10.1038/s41598-019-54396-4.

[33] G. v. Békésy, ''The structure of the middle ear and the hearing of one's own voice by bone conduction,'' *J. Acoust. Soc. Amer.*, vol. 21, no. 3, pp. 217–232, May 1949, doi: 10.1121/1.1906501.

[34] I. Nakayama, ''Voice timbre in autophonic production compared with that in extraphonic production,'' *J. Acoust. Soc. Jpn. E*, vol. 18, no. 2, pp. 67–71, 1997, doi: 10.1250/ast.18.67.

[35] M. Kimura and Y. Yotsumoto, ''Auditory traits of 'own voice,''' *PLoS ONE*, vol. 13, no. 6, Jun. 2018, Art. no. e0199443, doi: 10.1371/journal.pone.0199443.

[36] E. Daniel, ''Noise and hearing loss: A review,'' *J. School Health*, vol. 77, no. 5, pp. 225–231, May 2007, doi: 10.1111/j.1746-1561.2007.00197.x.

[37] H. Kato, Y. Hashimoto, and H. Kajimoto, ''Basic properties of phantom sensation for practical haptic applications,'' in *Proc. Int. Conf. Human Haptic Sens. Touch Enabled Comput. Appl.* Amsterdam, The Netherlands: Springer, Jul. 2010, pp. 271–278, doi: 10.1007/978-3-642-14064-8_39.

[38] Z. Zhang, L. Mongeau, and S. H. Frankel, ''Experimental verification of the quasi-steady approximation for aerodynamic sound generation by pulsating jets in tubes,'' *J. Acoust. Soc. Amer.*, vol. 112, no. 4, pp. 1652–1663, Oct. 2002, doi: 10.1121/1.1506159.

[39] W. Zhao, C. Zhang, S. H. Frankel, and L. Mongeau, ''Computational aeroacoustics of phonation, part I: Computational methods and sound generation mechanisms,'' *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2134–2146, Nov. 2002, doi: 10.1121/1.1506693.

[40] A. B. Vallbo, H. Olausson, J. Wessberg, and N. Kakuda, ''Receptive field characteristics of tactile units with myelinated afferents in hairy skin of human subjects,'' *J. Physiol.*, vol. 483, no. 3, pp. 783–795, Mar. 1995, doi: 10.1113/jphysiol.1995.sp020622.

[41] J. J. Gibson, ''Observations on active touch,'' *Psychol. Rev.*, vol. 69, no. 6, pp. 477–491, Nov. 1962, doi: 10.1037/h0046962.

[42] M. J. Ball, J. Esling, and C. Dickson, ''The VoQS system for the transcription of voice quality,'' *J. Int. Phonetic Assoc.*, vol. 25, no. 2, pp. 71–80, Dec. 1995, doi: 10.1017/s0025100300005181.

[43] T. Ito, K. Takeda, and F. Itakura, ''Analysis and recognition of whispered speech,'' *Speech Commun.*, vol. 45, no. 2, pp. 139–152, Feb. 2005, doi: 10.1016/j.specom.2003.10.005.

[44] Y. Ujitoko, S. Sakurai, and K. Hirota, ''Vibrator transparency: Reusing vibrotactile signal assets for different black box vibrators without re-designing,'' in *Proc. IEEE Haptics Symp. (HAPTICS)*. Washington DC, USA: IEEE, Mar. 2020, pp. 882–889, doi: 10.1109/HAPTICS45997.2020.ras.HAP20.80.00957e94.

[45] J. V. Bradley, ''Complete counterbalancing of immediate sequential effects in a Latin square design,'' *J. Amer. Stat. Assoc.*, vol. 53, no. 282, pp. 525–528, Jun. 1958, doi: 10.1080/01621459.1958.10501456.

[46] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, ''The aligned rank transform for nonparametric factorial analyses using only Anova procedures,'' in *Proc. SIGCHI Conf. Human Factors Comput. Syst.* Vancouver, BC, Canada: Association for Computing Machinery, May 2011, pp. 143–146, doi: 10.1145/1978942.1978963.

[47] J. Westfall. (2016). *Pangea: Power Analysis for General Anova Designs*. Accessed: Jun. 1, 2024. [Online]. Available: http://jakewestfall.org/pangea/

[48] T. Hirahara and R. Akahane-Yamada, ''Acoustic characteristics of Japanese vowels,'' in *Proc. 18th Int. Congr. Acoust.*, Kyoto, Japan, 2004, pp. 3287–3290.

[49] R. Terasawa, Y. Kakita, and M. Hirano, ''Simultaneous measurements of mean air flow rate, fundamental frequency and voice intensity–Results from 30 normal male and 30 normal female subjects,'' *Jpn. J. Logopedics Phoniatrics*, vol. 25, no. 3, pp. 189–207, 1984, doi: 10.5112/jjlp.25.189.

[50] E. Sugimori, T. Asai, and Y. Tanno, ''The potential link between sense of agency and output monitoring over speech,'' *Consciousness Cognition*, vol. 22, no. 1, pp. 360–374, Mar. 2013, doi: 10.1016/j.concog.2012.07.010.

[51] J. Rekimoto, ''WESPER: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2023, pp. 1–12, doi: 10.1145/3544548.3580706.

[52] Lisa. E. Rombout and M. Postma-Nilsenova, ''Exploring a voice illusion,'' in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 711–717, doi: 10.1109/ACII.2019.8925492.

[53] Z. Barlas and S. S. Obhi, ''Freedom, choice, and the sense of agency,'' *Frontiers Hum. Neurosci.*, vol. 7, p. 514, May 2013, doi: 10.3389/fnhum.2013.00514.

YUKI SHIMOMURA received the master's degree in environmental from The University of Tokyo, Tokyo, Japan, in 2023, where he is currently pursuing the Ph.D. degree with the Department of Frontier Sciences. His current research interest includes presenting a sense of vocal agency.

YUKI BAN received the M.S. and Ph.D. degrees in information science and technology from The University of Tokyo, Japan, in 2013 and 2016, respectively. From 2016 to 2017, he was a Researcher with Xcoo Inc. Research. Since 2017, he has been with The University of Tokyo. He is currently a Project Lecturer with the Department of Frontier Sciences, The University of Tokyo. His current research interests include cross-modal interfaces and biological measurement.

SHIN'ICHI WARISAWA (Member, IEEE) was an Assistant Professor with Tokyo Institute of Technology, from 1994 to 2000. Since 2000, he has been with The University of Tokyo. From 2010 to 2011, he was a Visiting Researcher with Massachusetts Institute of Technology. He was a Visiting Professor with Universite Jean Monnet, in 2016. He is currently a Professor with the Department of Frontier Sciences, The University of Tokyo. His current research interests include wearable/ambient human health monitoring, nano/micro sensing devices fabrication, and sensing information technology applications for human well-being.