Exploiting Voice Transformation to Promote Interaction in Virtual Environments

Vincent Isnard, Trami Nguyen, Isabelle Viaud-Delmon

Exploiting Voice Transformation to Promote Interaction in Virtual Environments

Vincent Isnard[*]
French Armed Forces Biomedical Research Institute (IRBA), Brétigny-sur-Orge, France

Trami Nguyen[†]
Ensemble Links, Paris, France

Isabelle Viaud-Delmon[‡]
CNRS, Ircam, Sorbonne Université, Ministère de la Culture, Sciences et Technologies de la Musique et du Son, STMS, F-75004 Paris, France

Abstract

We propose a new type of interaction dedicated to virtual reality (VR). We have designed a futuristic scenario promoting a dialogue with a humanoid. The participant is submitted to a reverse Turing test and has to answer questions addressed by the humanoid. The participant uses his/her own voice to respond, which is transformed in real time in terms of timbre and spatialization, in correspondence with special effects applied to the 360-degree video. Behavioral assessments are proposed after each vocal answer given, to determine the quality of the vocal interaction. Overall, the potential of this new narrative vector is confirmed by the very positive feelings of the participants, while promoting the extension of interactivity in 360-degree VR.

Keywords: Sonic interaction, virtual reality, embodiment, voice transformation, 3D sound, art.

Index Terms: H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

1 Introduction

Among the latest virtual reality (VR) technologies, immersive 360-degree monoscopic video offers a good compromise thanks to a relatively simple and accessible implementation to the general public. But this technology has strong limitations of interactivity, such as the absence of translational movements in the virtual environment, with only 3 degree-of-freedom (DOF) tracking corresponding to rotational motion, instead of full stereo with full 6-DOF of head motion. However, the result, obtained from natural images and sounds, may be sufficient to increase the intensity of a specific emotion and the feeling of presence [1]. In addition, some devices attempt to restore the participant's movements, for example by generating a parallax effect with light field camera networks [2] or by synthesis [3], [4].

A vocal interaction with real-time transformations, in the manner of "embodied conversational agents" (e.g., [5]; see also *Virtual Corporation*, 1996: a video game played with the voice), could also make it possible to compensate for these limits while exploring new sound narrative forms. Without having to add visual sources to the virtual environment, the use of voice could trigger events that promote interactivity and hence the immersion and quality of the VR experience [6]. Finally, the participant may be endowed with original vocal transformations to play directly a fictional character with his/her own voice (e.g., a robot, a monster, etc.). This new narrative source available to the fiction in VR could also amplify the intensity of the experience by introducing self-perception to the virtual world to be explored.

Also, while traditionally embodiment in VR is realized through the use of representations of virtual body parts or entire virtual bodies [7], it has been shown that the experience of body ownership can be influenced by other factors like internal physiological state [8]. Introducing the own voice of the participant in a virtual world might constitute a new powerful way to allow embodiment in the virtual scene, even in the absence of a visual representation of the body. Moreover, as the voice can be compared to an "auditory face" [9], the use of own voice in VR could represent an auditory analogy of the avatar, linking the body of the participant to the virtual world. Indeed, the variability of human voices makes them unique (and even used today as biometrics, e.g., [10]). However, it was also shown that self-voice recognition remains very robust even with severe acoustic degradation [11]. Thus, in a playful or artistic context, creators of VR experiences could take advantage of the potential of embodiment of the participant's voice, while including real-time transformations in imaginary virtual environments.

Here, a futuristic scenario serves as a pretext for a dialogue exchange between a humanoid, embodied by an actor, and the participant. The participant's own voice is transformed in real time to test whether the interaction remains believable even when his/her voice has sound qualities that completely diverge from a natural voice. Moreover, in addition to vocal timbre transformations, the contribution of own voice internalization to VR immersion is assessed by externalizing the participant's own voice. Indeed, these new types of sound practices, in the context of immersive 3D sound applied to VR, still remains to be studied and promoted in the VR community [12], especially when combined with the recent technological developments in ambisonic microphones and binaural restitution [13]. Furthermore, all sound transformations are performed in coherence with visual transformations, given that temporal, spatial and semantic levels can contribute to the audiovisual perception of complex environments, as they can be offered in VR [14].

Finally, by involving the participant in the fictional process through his/her voice, all these sound and visual alterations seem to have favored interaction and VR immersion.

2 Artistic Scenario

The scenario is based on an inverted Turing test modeled on the Voight-Kampff fictitious test [15]. It features Pieter Musk, a humanoid interviewing the participant to determine his/her degree of humanity and make him/her think about his/her human condition at a time of Internet, artificial intelligence, new technologies, robotic prosthesis, and other technological augmentations [16].

[*] vincent.isnard@def.gouv.fr
[†] traminguyen80@gmail.com
[‡] isabelle.viaud-delmon@ircam.fr

978-0-7381-1367-8/21/$31.00 ©2021 IEEE
DOI 10.1109/VRW52623.2021.00021

The 16 questions proposed are intentionally disturbing in order to elicit an emotional reaction specific to behavior considered as normal for a human. Here is an example of a question and the corresponding answers asked by the character: *“Your 7-year-old child comes home with a jar filled with dead frogs [...]. He also hands you the knife still bloodied which he used to cut the frogs [...]. What do you say to him? Answer A: wonderful! I'll get rid of all that [...]. Answer B: you act as if nothing had happened [...]. Answer C: you roll your eyes, dizzy [...].”* (see [17] for an online video excerpt of this installation presented during the IRCAM Forum in 2020).

The 3 possible answers correspond respectively to a “human”, “post-human” or “machine” behavior (e.g., in the previous example: answer A = “machine”; answer B = “post-human”; answer C = “human”). A “humanity score” is displayed at the end of the experience based on the participant’s answers, for a playful conclusion that conforms to the framework of the scenario, without entering into the scientific analysis.

3 Scientific evaluation

3.1 Participants

Pilot data were obtained from 8 participants (4 women, 4 men, mean age: 36.0 ± 6.4 years). All participants reported normal hearing and normal or corrected-to-normal vision. Participants gave written informed consent before passing the experience.

3.2 Experimental conditions

The vocal content is controlled with exchanges calibrated in numbers of syllables (174.0 ± 7.7 syllables per question; 16.6 ± 1.4 syllables per answer to be read by the participant), for a video lasting a total of 18’28’’. The quality of the voice interaction is tested according to 2 visual and sound processing conditions (see Table 1 and Fig. 1): (1) “Timbral” transformations, based on vocoded robotic voices and visual distortions, at low or high intensity; (2) “Spatial” transformations, with low or high consistency between original and perceived sources. For “Spatial” transformations concerning the participant: in the “low” condition, his/her voice is internalized; while in the “high” condition, his/her voice is moved 3m in front of him/her (i.e., with fixed and zero azimuth and elevation angles). Regarding the actor, the image rendering includes low or high dynamic colorimetric dissociations, thus leading to further dissociations between the image of his body and the sound of his voice, without having to apply additional spatial sound transformations.

These transformations are crossed by groups of 4 question and answer exchanges, so as to obtain 4 repetitions of perceptual evaluations for each combination of conditions. The order of the exchanges and the transformations are the same for all participants and the answers given by the participant do not influence the scenario.

Vocal exchanges	“Timbral”	“Spatial”
Introduction	*Low*	*Low*
1 to 4	Low	Low
5 to 8	Low	High
9 to 12	High	Low
13 to 16	High	High
Conclusion	*High*	*High*

Table 1: Visual and sound transformations according to the successive vocal exchanges: introduction, 16 questions and answers, and conclusion. The participant does not intervene in the introduction and the conclusion.

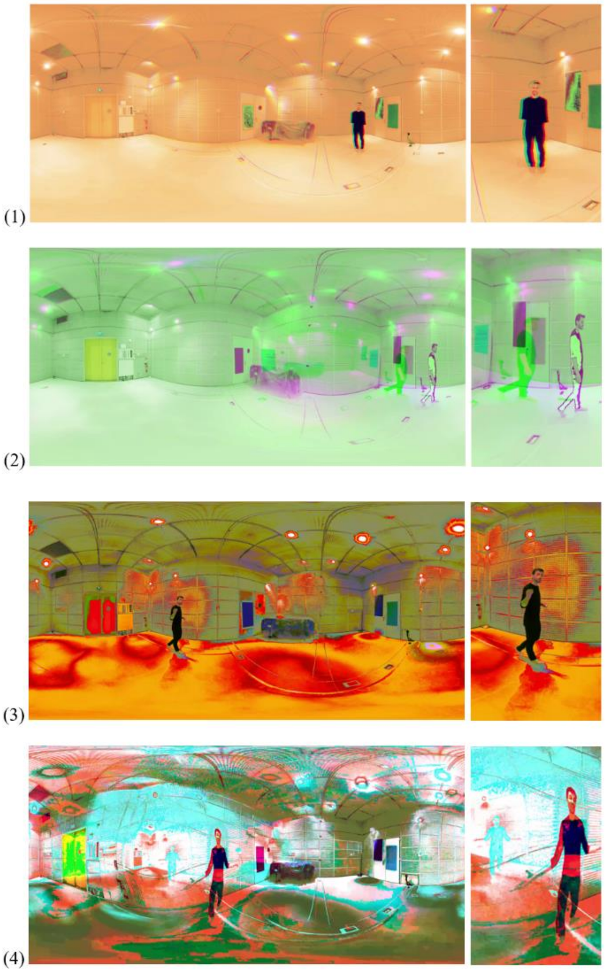

Figure 1: 360-degree visual rendering for each condition of audiovisual transformations. (1) Low “Timbral” transformation and low “Spatial” transformation; (2) low “Timbral” transformation and high “Spatial” transformation; (3) high “Timbral” transformation and low “Spatial” transformation; (4) high “Timbral” transformation and high “Spatial” transformation. The figures on the right are an enlargement of the virtual character that the participant had to follow in the virtual environment.

3.3 Materials

The videos were recorded with an actor in a large empty studio, at the Institute for Research and Coordination in Acoustics/Music (IRCAM), with an Insta360 Pro 2, which is a 360-degree monoscopic camera (7680 x 3840 resolution). Then, the images were stitched, edited and processed in Adobe Premiere Pro 2019. In parallel, the sound was captured in ambisonics with an MH Acoustics Eigenmike spherical microphone (32 channels, 24 bits, Fe = 44.1 kHz). Then, it was edited and normalized in Reaper 5. All parasitic noises related to the actor (e.g., breaths) were cut to accentuate the fictional digital aspect of the character. To note that we used here professional equipment: the camera was rented while the microphone was loaned by IRCAM, and we used professional software (some of which, or equivalents, are available for free). However, as mentioned in the introduction, hardware and software tools for VR, and in particular for immersive 3D sound adapted to VR, are now accessible to the general public at moderate costs (see [13]).

During the VR experience, the synchronous reproduction of the image and the sound is carried out, on a PC dedicated to VR, thanks to a custom Max 8 patch, configured with a I/O vector size and a signal vector size of 512, as a compromise to reduce processing latencies as much as possible while maintaining the sound quality of "Timbral" and "Spatial" transformations. Note that delays induced by real-time voice transformations remain acceptable under 20 ms for natural speech, but possibly few tens of ms in other settings [18]. By informally questioning the participants, the real-time processing on their own voice did not seem to affect their own voice perception in this playful setting.

This patch integrates the "VR" library for the 360-degree image reproduction, while the sound is converted in real time from ambisonic format to binaural format with the "Spat 5" library, so as to maintain a consistent spatialization with the orientation of the head during the experience. Indeed, the ambisonic 3D sound scene is first rotated according to the rotation data from the head-mounted display (HMD), then decoded through 32 virtual speakers (see Fig. 2). Finally, the binaural rendering is obtained with generic KEMAR dummy-head head-related transfer functions (HRTFs). Participants did not report having difficulty locating sound with these generic HRTFs, and certainly benefited from multisensory integration effects between image and sound [14].

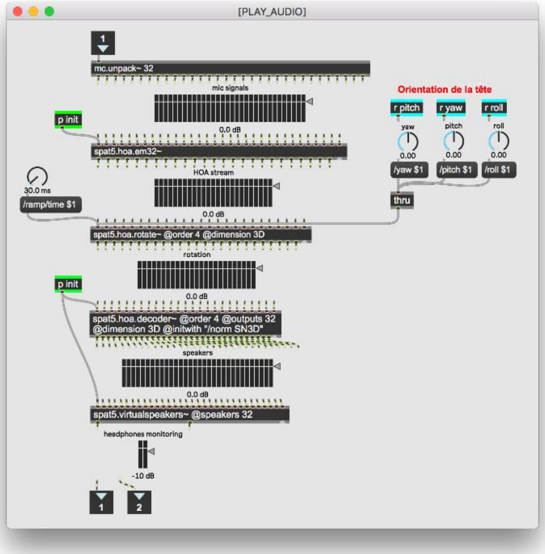

Figure 2: Custom Max 8 patch for the ambisonic to binaural conversion of the 3D sound. The orientation coordinates of the head, obtained from the HMD, are retrieved at the top right to perform the rotation of the 3D sound scene before conversion to binaural format, then reproduced on headphones.

Participants are equipped with an Oculus Rift CV1 HMD (1080 x 1200 resolution per eye, 90-Hz refresh rate, 110-degree field of view) and Beyerdynamic DT 770 headphones, connected to an RME Fireface UC sound card. Moreover, 2 Oculus Touch controllers are also used for the course of the experience, with buttons to lauch the videos and a joystick to perform the perceptual evaluation.

The voice of the participants is picked up using a DPA 4066 microphone, connected to the sound card, to be processed in real time in the same patch. The following sound treatments are carried out in cascade. First treatment: "Timbral" transformations with vocoders (the same transformations used for the voice of the actor, with variable intensity, low or high, depending on the experimental conditions, cf. Table 1). Second treatment: for low "Spatial" transformation only (cf. Table 1): a filtering to simulate own voice listening without microphone, with mainly a filter cutting high frequencies above 5 kHz [19]. Indeed, the choice of closed headphones aims to reduce as much as possible the listening of the natural voice (thus, by air conduction) to privilege the transformed voice, and therefore implies in return to simulate the filtering by the head under natural conditions. An informal test validated the parameters of this filtering as giving a more natural rendering than without filtering when no other transformation is applied. Finally, the last treatment applied on the voice is the binaural spatialization (cf. Table 1). And to the spatialization is added a short artificial reverberation based on a simplified model of a room impulse response from the "Spat 5" library, to add a low room effect more consistent with the virtual space (as the experiment took place in a small studio treated acoustically) and the voice position (or "Spatial" transformation), with early and late reflections delays less than 10 ms, and late reverberation less than 50 ms [20]. Thereby, the two alternatives "Spatial" transformations of the participant's voice (low or high) and the actor's voice are as much as possible associated to the same virtual room.

During the VR experience, participants are sitting on a rotating chair in an IRCAM studio. They are encouraged to explore the environment in 360-degree by following the character appearing successively in different places in space.

3.4 Procedure

After each question and the 3 possible answers pronounced by the character, these are displayed on the HMD. Then, participants have to read their chosen answer aloud before performing the following perceptual assessment: "Do the treatments on your voice promote your interaction with the virtual character?". They then move a cursor, on a vertical analog scale appearing in the HMD. The quality of vocal interaction is rated from 0 ("Not at all") to 100 ("Completely"), according to the position of the cursor on the scale. They were previously told orally that this assessment relates to VR immersion as well as the consistency of their virtual incarnation in the context of fiction. The whole VR experience lasts approximately 30'.

After the VR experience, participants complete a form to give their general appreciation on the experience, any general discomfort that may have been endured during the experience, the duration of the experience, and a free comment on the installation.

4 Results

First, participant feedback reported few symptoms of cybersickness (e.g., "no" symptoms, "very slightly", "a little nausea", "a little disoriented"). However, the length of the VR experience was often judged to be a bit too long by most participants.

The means of the perceptual evaluation of vocal interaction, over the 4 repetitions of each combination of the experimental conditions, were compared on all participants (see Fig. 3). Contrary to our expectations, no glaring differences appeared between the experimental conditions in terms of voice interaction rating scores, even between low and high intensity transformations.

Therefore, participants seem to have generally appreciated the experience, despite the constraints of the experimental evaluation setting added to the fictional scenario, with an average evaluation over all the conditions of 62/100. This interpretation is corroborated by their comments: e.g., "much appreciated", "surprising", "interesting". On the other hand, from informal discussions, the appreciation of the experience seems to be all the stronger for those participants already familiar with VR or new technologies.

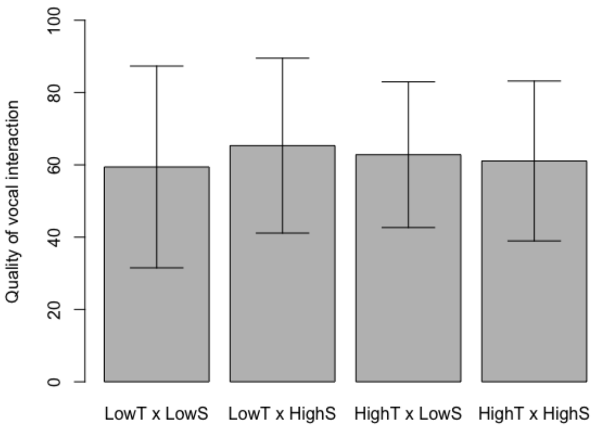

Figure 3: Subjective evaluation of the quality of the vocal interaction. The 4 combinations of the experimental conditions are noted: LowT = low "Timbral" transformation; HighT = high "Timbral" transformation; LowS = low "Spatial" transformation; HighS = high "Spatial" transformation. The bars correspond to the means and the error bars to the 95% confidence intervals.

5 Discussion

These data must be supplemented by new evaluations to confirm the results. In addition, even for this VR experience comprising a priori few events likely to cause cybersickness, the time spent in a HMD remains a key factor in its acceptability, although it was already the subject of a compromise with a strong limitation of the tested conditions (e.g., only one perceptual evaluation under HMD). The use of techniques synthesizing 6-DOF tracking from monoscopic 360-degree video could also reduce motion sickness effects [3].

Despite the experimental constraints, the integration of the experience into a fictional setting is highly appreciated and tolerated by the participants. And the obtained results allow us to give a first estimate of the possibilities of vocal interaction as a new vector of interactivity in VR. This experience must evolve with freer vocal interactions, controlling a posteriori the amount of speech to also give an indication of the participant's involvement; as well as enhanced interactivity by modifying the course of the scenario according to the participant's answers. Other factors will be explored: e.g., familiarity, eeriness, and the sense of own voice [21]; acceptance of the interaction with a virtual character [22].

More surprisingly, there was no difference observed here depending on the "Timbral" or "Spatial" transformations applied to the voice and their intensity. The intensity range of the transformations will therefore be increased in an attempt to remedy this. In particular, a check will be carried out with the own voice of the participants without transformation, and vice versa, with strong transformations, to examine to what extent these are tolerated without abolishing the sense of own voice. In addition, movements in sound spatialization will be added in order to promote the externalization of their own voice, to examine the contribution of internalization to immersion or to the feeling of presence in VR, and the role of 3D sound in these cognitive treatments [6], [23].

Moreover, the reverberation, added to the participant's voice, was here configured to always associate it with the same virtual environment in which the virtual character was evolving, i.e., the same acoustic space independently of the "Spatial" transformations. On the one hand, the estimation of the virtual acoustic space could be optimized using impulse responses measured at the time of filming or modeling techniques on 360-degree images [24]; on the other hand, the reverberation could be an additional parameter on which to play to measure its impact on immersion in the context of a narrative fiction [25].

Another interpretation of these results concerns the influence of the fictional scenario in VR and the choices of artistic transformations. Here, the evolution of visual and sound transformations was gradual with an overall increase in the intensity of the effects. The initial artistic intention was to imagine a gradual breakdown of the digital environment where the fictitious test takes place, with an increase in the strangeness of the questions, thus aiming at an awareness of the growing role of technologies in the natural environment [16]. Thus, as the transformations became more extreme, participants may have benefited from a habituation effect during the test, and become more involved in the VR experience, assigning generally constant scores despite an increase in intensity of the effects, while the effects at the end of the experience, in particular, might have been deemed too extreme at first glance. This interpretation would signal a flexibility in the tolerance of the transformations made to the own voice in a VR experience, and should be tested by comparing their increase in intensity, like here, or constant transformations.

Finally, according to the profiles of the participants, those with a knowledge of VR or new technologies, even minimal, seem to have appreciated the experience more than the more naive participants, while their expertise could on the contrary have led them to criticisms of the constraints of the experimental setting. This therefore confirms the high potential of the vocal interaction as a new narrative vector in VR. The profile of the participants according to their familiarity with VR will be more systematically studied. In particular, control will be carried out with naive participants subjected to a preliminary stage of familiarization with VR, to check if their appreciation increases by being made more available for the immersive experience and the new features of the vocal interaction.

This VR installation paves the way for other experiences involving the participant's voice for immersive artistic and/or musical creations, in spatialized sound environments transformed in real time.

Acknowledgments

This installation is the result of an art/science project supported by IRCAM and GRAME, French national centers for musical creation, as part of complementary residencies in 2019-2020. Part of the protocol and the results were presented during the IRCAM 2020 Forum.

References

[1] A. Chirico et al. Effectiveness of immersive videos in inducing awe: an experimental study, *Scientific Reports*, *7*(1), 1-11, 2017.

[2] R. S. Overbeck et al. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality, *ACM Transactions on Graphics (TOG)*, *37*(6), 1-15, 2018.

[3] J. Huang et al. 6-DOF VR videos with a single 360-camera, In *2017 IEEE Virtual Reality (VR)* (pp. 37-44), IEEE, March 2017.

[4] G. D. de Dinechin and A. Paljic. *Cinematic virtual reality with motion parallax from a single monoscopic omnidirectional image*. In 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018), IEEE, 2018.

[5] D. Potdevin et al. Virtual Intimacy, this little something between us: A study about Human perception of intimate behaviors in Embodied Conversational Agents, In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 165-172), November 2018.

[6] M. Slater and S. Wilbur. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual

environments, *Presence: Teleoperators & Virtual Environments*, *6*(6), 603-616, 1997.

[7] K. Kilteni et al. The sense of embodiment in virtual reality, *Presence: Teleoperators and Virtual Environments*, *21*(4), 373-387, 2012.

[8] K. Suzuki et al. Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion, *Neuropsychologia*, *51*(13), 2909-2917, 2013.

[9] P. Belin et al. Thinking the voice: neural correlates of voice perception, *Trends in cognitive sciences*, *8*(3), 129-135, 2004.

[10] A. Boles and P. Rad. Voice biometrics: Deep learning-based voiceprint authentication system, In *2017 12th System of Systems Engineering Conference (SoSE)* (pp. 1-6), IEEE, June 2017.

[11] M. Xu et al. Acoustic cues for the recognition of self-voice and other-voice, *Frontiers in psychology*, *4*, 735, 2013.

[12] R. R. A. Faria et al. Improving spatial perception through sound field simulation in VR, In *IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2005,* (pp. 6-pp), IEEE, July 2005.

[13] Zotter, F., & Frank, M. Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality, Springer Nature, 2019.

[14] C. Suied et al. Integration of auditory and visual information in the recognition of realistic objects, *Experimental Brain Research*, *194*(1), 91, 2009.

[15] P. K. Dick. Blade Runner. Les Androïdes rêvent-ils de moutons électriques ?, J'ai Lu, 2014.

[16] B. Frischmann and E. Selinger. *Re-engineering humanity*, Cambridge University Press, 2018.

[17] V. Isnard and T. Nguyen. *L'étrangeté perceptive en réalité virtuelle.* Forum IRCAM, 2020.

[18] L. Rachman et al. DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior research methods*, *50*(1), 323-343, 2018.

[19] C. Pörschmann. Influences of bone conduction and air conduction on the sound of one's own voice, *Acta Acustica united with Acustica*, 86(6), 1038-1045, 2000.

[20] T. Carpentier. A new implementation of Spat in Max. In *15th Sound and Music Computing Conference (SMC2018)* (pp. 184-191), July 2018.

[21] M. Kimura and Y. Yotsumoto. Auditory traits of "own voice", *PloS one*, *13*(6), e0199443, 2018.

[22] F. Eyssel et al. 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism, In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125-126), IEEE, March 2012.

[23] S. Nichols et al. Measurement of presence and its consequences in virtual environments, *International Journal of Human-Computer Studies*, *52*(3), 471-491, 2000.

[24] H. Kim et al. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 120-126), IEEE, March 2019.

[25] D. Västfjäll et al. Emotion and auditory virtual environments: affect-based judgments of music reproduced with virtual reverberation times. *CyberPsychology & Behavior*, *5*(1), 19-32, 2002.