Perceptive strangeness in virtual reality

Artistic research residency at IRCAM in 2019-2020

Vincent Isnard & Trami Nguyen

Abstract

360° videos in virtual reality (VR) present certain limits, such as the impossibility of carrying out movements in the environment by the spectator's own will or direct physical interactions with virtual objects. The voice of the spectator could be a means of increasing this interactivity since it does not require physical interaction but could be enough to produce physical consequences in the virtual world. In this artistic research developed as part of a residency at IRCAM in 2019-2020, we sought to assess the added value of interactivity when the spectator uses his own voice, by including real-time transformations of timbre and spatialization to integrate it into a scenario with a futuristic context through a dialogue with an artificial intelligence having taken human form. Indeed, a fictitious inverted Turing test serves as a pretext for this dialogued interaction and must allow to fictitiously assess our own degree of humanity. A scientific test on perception in VR is carried out simultaneously in order to assess whether the quality of the spectator's voice transformed in real time allows him to further embody his character in the fiction in VR, playing on the strangeness effect that these transformations can generate.

Preamble

This article details the scientific issues raised in the context of a multisensory perception experience in VR, the artistic proposals pertaining to a futuristic and anthropocenic context inserted in a narration in VR, as well as the technological choices (immersive 360° video, ambisonic and binaural 3D sound) made during our artistic research residency at IRCAM in 2019-2020. The installation developed in VR at the end of the residency and presented at the IRCAM Forum from March 4 to 6, 2020 is the first outcome of the project. It currently continues to be developed, on the one hand in a scientific framework (perceptive test), on the other hand in an artistic framework (interactive installation and autonomous film in VR), not in a compartmentalized manner but by a strong and inspiring emulation between science and art.

The Pieter Musk test

Scientific and technological issues

Three major issues, both scientific and technological, have fueled the perceptive experience developed in VR: the perceptive strangeness, the perceptive adaptation of sound and visual content in VR and the use of the voice to interact in VR. Three lines

of research ensued and inspired us alternately or simultaneously during the development of our installation.

The perceptive strangeness:

The perceptive strangeness is a concept whose stakes are becoming more and more significant at the present time with the advent of smart speakers, embodied conversational agents or even robots.

To easily grasp this concept, one can for example rely on fantastic literature: in *The sandman* (1815) by E.T.A. Hoffmann, a young boy is in love with a young girl. However, he is confused because of many aspects (coldness in contact with her skin, impassive face...). At the end of this short story, he will realize that the girl was not a human being but a robot developed to perfection by her "father", physicist. The strangeness or worry that this human familiarity can provoke has been studied subsequently, notably in psychology (e.g. Freud, 1919).

In 1970, the roboticist Mori proposed the hypothesis of the uncanny valley. According to his hypothesis, the less an entity is similar to a human being, the weaker the reaction it provokes (e.g. emotional reaction...), and conversely, the more the similarity increases, the stronger the reaction. But this relation is not linear: it presents a hollow called "uncanny valley" (Fig. 1). This hollow indicates that when the similarity is strong enough but nevertheless imperfect, the quasi-human entity can provoke a very negative reaction (e.g. aversion). There are many interpretations to explain such a negative reaction: one may wonder whether the person is dead, subject to a pathogen, to a disturbing absence of defects, or even from a cognitive point of view, it could be a conflict between the involved perceptive cues, we would believe to recognize a human being but we would not be sure and our perceptive system would be disturbed.

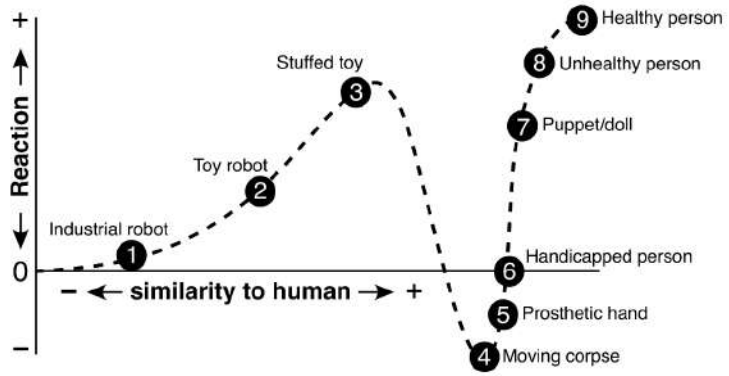


Figure 1. The uncanny valley. Adapted from Mathur & Reichling (2016).

A certain number of authors have tested this hypothesis in a scientific framework, in particular to assess whether android robots, generally too slick to be truly human, could generate this type of negative reactions, which would be very harmful for the robotics industry (Fig. 2).

The artistic scenario of our installation includes a conversation with an embodied artificial intelligence. The challenge will be to find the best visual and auditory representation of this entity to make it the most credible and relevant within the story of our fiction, potentially by playing on a strangeness effect perceived by the participant.

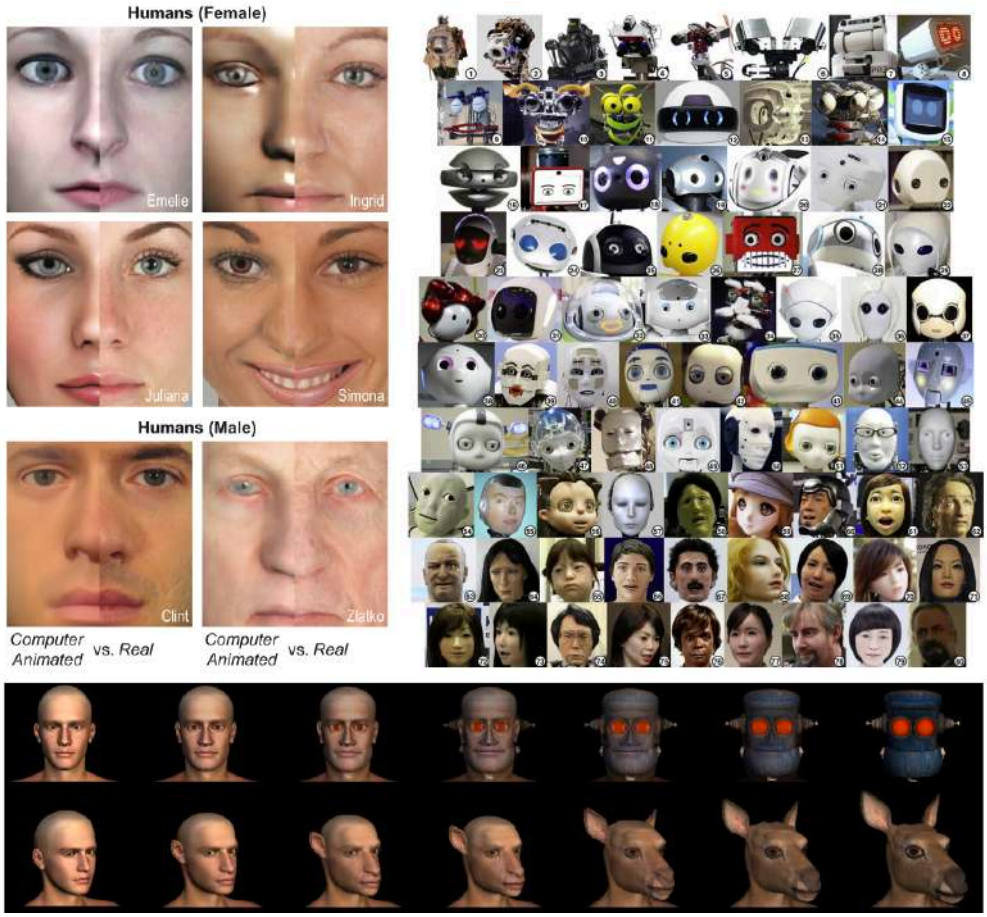


Figure 2. Examples of recent studies on the perception of strangeness. In each of these studies, the authors evaluate the reaction provoked by an android entity compared to a human being. Adapted from: left: Chattopadhyay & MacDorman (2016); right: Mathur & Reichling (2016); below: Ferrey et al. (2015).

Sound and visual coherence of artistic content in VR:

VR currently requires powerful and expensive computers. In particular, as with video games, visual rendering by synthesis of virtual environments is complex to implement (this is less true for 360° immersive video), while sound is generally much simpler to generate, transform and transmit in VR to obtain a rendering consistent with the artistic intention. The lack of realism in the virtual environment is not necessarily a problem in itself, because the narration is often enough to be fully immersed in the VR experience. The challenge therefore lies above all in the coherence of the visual and auditory virtual environment with the artistic intention.

However, due to the great divergence of the creative processes in image and sound, there is a risk of generating an annoying perceptive incongruence in VR (the borderline case being a non-realistic visual rendering, because complex to implement, associated with a realistic sound content, because easily obtained and handled). This is why, instead of proposing more and more complex processes of the image creation, often at the expense of artistic intention, we can propose a "degradation" of the sound content to tend towards a better perceptive visuo-auditory congruence and a better immersive VR experience.

In a previous work, we proposed "auditory sketches" of complex sounds as counterpart of visual sketches, created on the basis of auditory recognition cues (Fig. 3; cf. Isnard, 2016; Isnard et al., 2016).

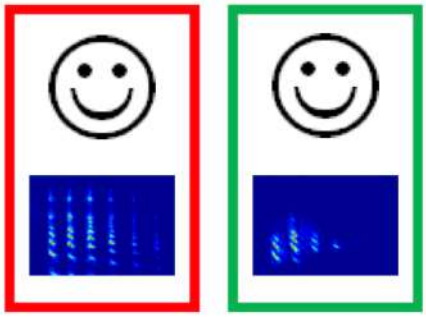


Figure 3. Schematic representation of visuo-auditory incongruence (left) and congruence (right) that can be generated depending on the level of adaptation of image and sound content. Few visual features are enough to recognize a visual object (here a face). In the same way, few auditory features are enough for the human auditory system to recognize a sound. We hypothesize that visuo-auditory congruence is better in the case where the complexities of visual (e.g. face) and auditory (e.g. voice) objects are adapted.

For our VR installation, the image and sound contents will be transformed to correspond to the artistic intention and to the futuristic and anthropocene context of the scenario. The image and the sound will have to be adapted in correspondence to favor visuo-auditory congruence and a better experience of immersion in VR.

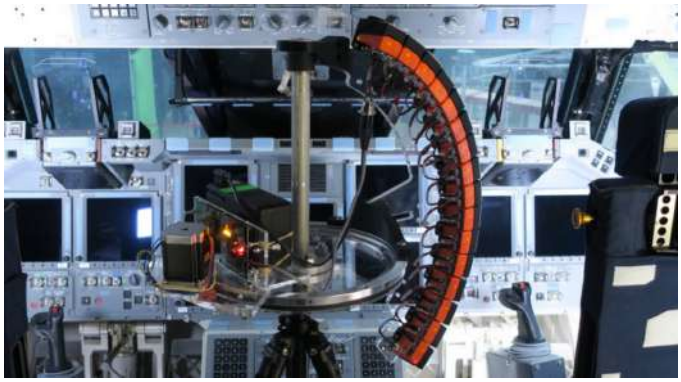
The use of voice to interact in VR:

360° VR videos do not allow movements or interactions. The image is captured in 360° and the participant who views it can only rotate in 360° to observe the entire immersive scene. Its advantage is that it is relatively simple to implement (compared to the development of a 3D synthesis environment) and that it allows obtaining a perfectly realistic quality as raw material before artistic processing.

Some proposals are under development to overcome these limits. For example, light-fields camera networks, or even a reconstruction of the parallax effect on an initially monoscopic image using computational techniques (Fig. 4).

We can also draw inspiration from the interactivity offered by the "embodied virtual agents", for whom the participant's voice, gestures or gaze can be involved in an interaction with this type of virtual agents (Fig. 5).

In our installation, we opted to use the participant's voice, which we believe represents an inexpensive, simple and effective way to improve interactivity in 360° VR. In addition, we propose to add real-time transformations of the participant's own voice to allow him to best embody a fictional character (for example, if the participant had to embody a monster in a video game, he would be offered to transform his own voice in real time into a monster voice so that he would best embody his character). However, one may wonder if such transformations will not be likely to generate a strangeness effect on the participant and if the visual-auditory congruence will always be respected.



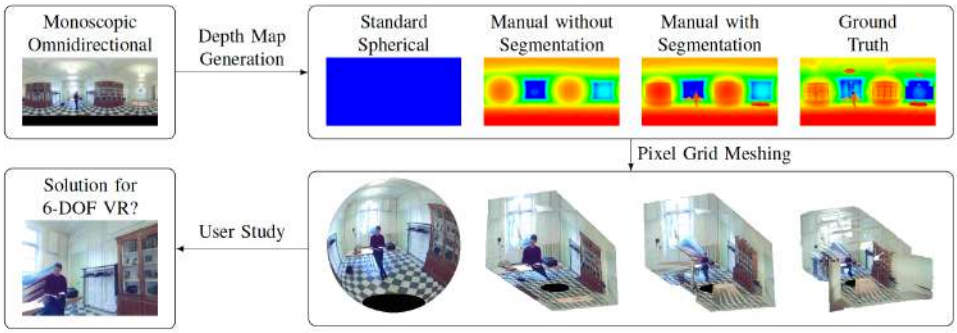


Figure 4. Examples of techniques for proposing a displacement of the body in 360° VR. Above: network of light fields cameras (Google); below: synthesizing process of the parallax effect on a monoscopic image (adapted from Dinechin & Paljic, 2018).

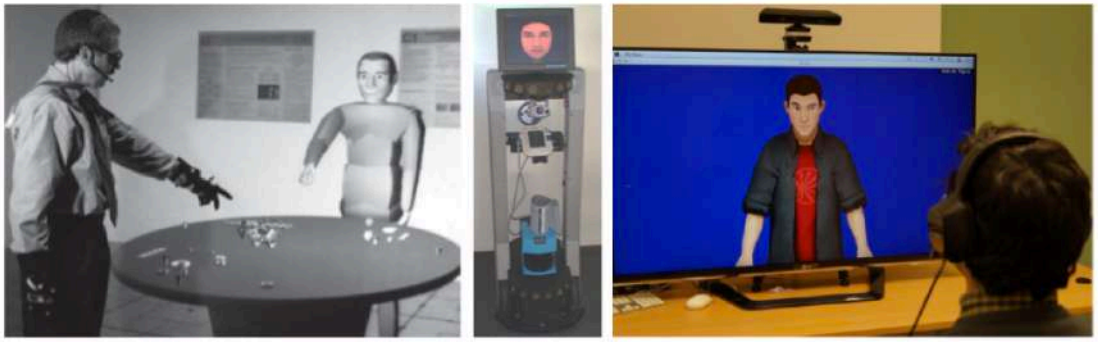


Figure 5. Examples of embodied conversational agents. Adapted from: left: Kopp et al. (2003); middle: Tamagawa et al. (2011); right: Baur et al. (2013).

Artistic scenario

Originally, the Turing test was proposed by the famous mathematician to determine whether a given entity is a human being or a machine (an artificial intelligence), assuming that with the improvement of technologies the border between two would become all the more tenuous and that machines could end up impersonating humans by imitating some of our cognitive abilities. The test essentially consists in asking questions to this unknown entity via an interface. Depending on the given answers, we must deduce whether this entity is a human being or a machine.

Starting from this source of inspiration, the science fiction writer Philip K. Dick, in *Do androids dream of electric sheep?* (1966; taken to the cinema under the

title of *Blade runner* by Ridley Scott in 1982, cf. Fig. 6), imagined the fictitious Voight-Kampff test designed to allow the police to find and unmask escaped androids (the "replicants"), otherwise indistinguishable from the human population. This test consists of asking disturbing questions to examine whether they provoke emotional reactions in the subject, which do not exist in the replicants.

Figure 6. Poster of the film "Blade runner".

Today, however, it is less the question of machines that would pass themselves off as humans that seems really problematic, insofar as it is always humans (for now) who control the machine and who seek to improve this imitation for better and for worse, than the problem of humans who would gradually turn into machines through various technological enhancements. We think for example of all of our electronic assistants (starting with the Internet) up to human enhancement by the use of electronic prostheses (cf. Frischmann & Selinger, 2018).

For our installation, we therefore imagined a subversive test to determine if the participant would not be, to a certain degree, a machine. In the VR experience, the test is presented by our fictional character Pieter Musk, an artificial intelligence who presents himself as the son of Elon Musk, the famous entrepreneur. Using his reverse Turing test, Pieter Musk seeks to identify the spectator, by determining his degree of humanity, to allow him or not to access his father's confidential data. The questions are intentionally disturbing to elicit an emotional response. An example: "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 [...]."

In the installation, the answers given by the participant are part of the fiction and do not count towards scientific analysis.

Experimental protocol for the scientific test

To test the participant's interactivity and immersion in the VR installation, several parameters are successively modified during the test:

- the sound processed in real time in timbre (human voice or robotic voice) and in spatialization (voice co-located with the voice source or delocalized);
- the image processed correspondingly, respectively in distortion and RGB dissociation.

For each fictitious question asked by Pieter Musk, the participant must answer aloud. Following this, he performs a perceptive evaluation on a visual scale to determine whether or not the treatments on his own voice favor his interaction in VR.

We hypothesize that congruent treatments between image and sound, and between the fiction environment (futuristic) and the participant's voice (made robotic), favor this interaction and improve interactivity.

The whole VR experience lasts around 30 min.

Design of the VR installation

The actor chosen for this experience is Piersten Leirom, performer and dancer. The filming took place in a studio at IRCAM in 2018. The filming equipment was as follows (cf. Fig. 7):

- for the image, we used a 360° Insta Pro 2 camera (rental) which has the advantage of having an 8k resolution, a remote management of the fan (to limit noise in the sound recording) and video capture, and also an automated stitching before importing 360° images to PC;

- the sound was recorded in ambisonics using an Eigenmike 32-capsules microphone (belonging to IRCAM). Note that more accessible microphones exist such as the Zoom H3-VR or the Zylia ZM-1; the same goes for the image with a wide range of consumer cameras.

As a reminder, the ambisonic sound (obtained by recording or synthesis) can be decoded on any type of sound reproduction system (ambisonic dome, 5.1 system, etc.), and in particular in binaural, that is to say in 3D sound in any headphones, keeping all of the spatialization information from the original sound scene.

Figure 7. Filming equipment. Left: Insta Pro 2 360° camera; right: Eigenmike 32-capsules ambisonics microphone.

For editing, we used Adobe Premiere Pro which supports images from Insta Pro 2. For sound, we used Reaper which easily manages files with a large number of channels. The start and end cuts of each shot have been adjusted thanks to filming claps and timecodes.

Note that there is a panoply of (free) tools, the Facebook 360 Spatial Workstation, for producing VR. The Spatialiser allows you to manage spatialized sound in Reaper (or other DAWs) with video monitoring. The Encoder allows you to combine a VR image with spatialized sound into a single video file. We still opted to perform this "encoding", at least the simultaneous reading of the image and sound, in Max 8 (see next paragraph), especially because these tools do not currently allow to manage files of high-spatial resolution (8k for the image, 32 channels for the sound).

Playback and real-time image and sound processing were therefore performed in Max 8. Connection with our VR headset, an Oculus Rift, was made possible thanks to the "vr" library developed by Graham Wakefield (Fig. 8; the Oculus Rift is not the only headset supported by this library). This library is extremely practical and efficient because it makes it possible to retrieve all the spatial data from the Oculus headset but also from the controllers. And it obviously allows the display of an image in VR in the VR headset.

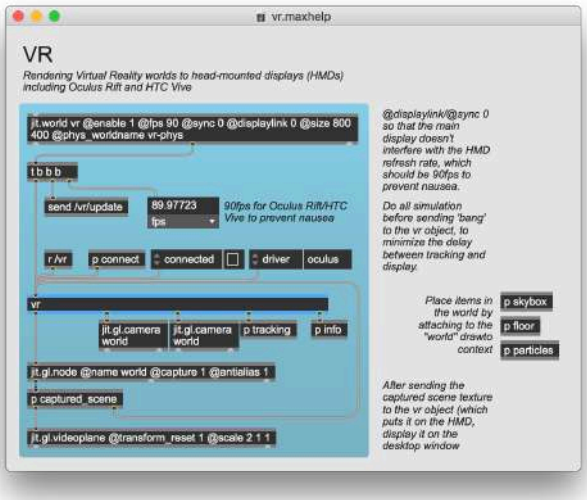


Figure 8. Help window from the "vr" library in Max 8 allowing connection to the Oculus Rift and display of an VR image in the VR headset.

For the spatialized sound, we used the "spat" library developed within the Acoustic and Cognitive Spaces team at IRCAM. This library is particularly complete and flexible to use for all aspects of 3D sound that exist today (Fig. 9).

Ultimately, the video is played back using the VR headset in which the image is displayed in 360°, while the sound is reproduced on the headphones in binaural. The participant equipped with VR equipment can observe the immersive scene all around him by turning his head and body, and the visual and auditory renderings are then reproduced in a coherent manner and updated in real time (simultaneous rotations of the 360° image and the 3D sound scene).

Finally, to improve interactivity, the participant's own voice is captured using a headset microphone to be processed in real time through vocoders to generate a robotic timbre, as well as in spatialization again using the Spat. However, we preferred not to modify the evolution of the scenario of the fiction according to the answers pronounced by the participant so as not to disturb or burden the scientific test. The questions and answers are therefore linked in a pre-established order.

Note, for this last aspect, that when we speak into a microphone and listen to ourselves through headphones, we hear our voice as can be heard by people around us. However, this does not correspond to the timbre that we hear ourselves, because we hear two sound streams distinct from the timbre heard by people around us (or as it is picked up by the microphone): first the sound that comes out of our mouth is filtered by our head before reaching our ears, second the sound that is produced by our vocal cords also passes, by a second path, directly by bone conduction in our ears with specific filtering. We therefore carried out a global filtering simulating these two concomitant filters, as proposed by Porschmann (2000). It corresponds to a filtering

of high frequencies above 5 kHz, which therefore allows you to hear yourself through the microphone and the headphones as when you hear yourself speaking naturally without all the setup requested here for the VR.

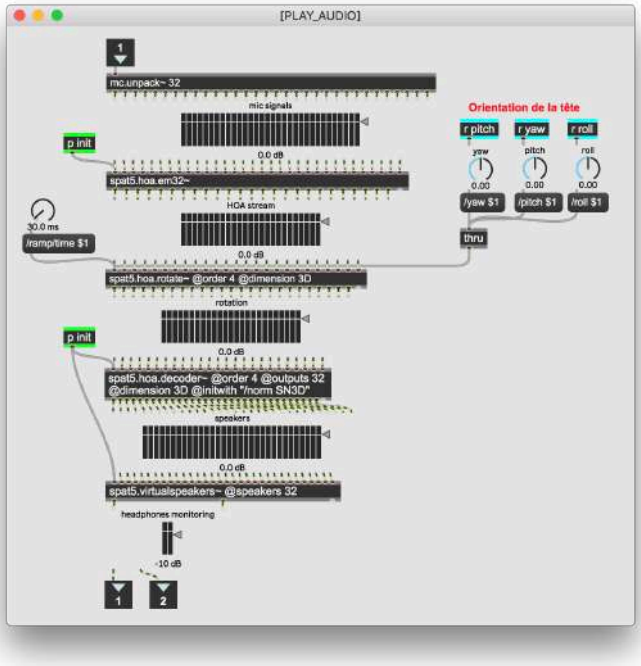


Figure 9. Ambisonic to binaural conversion using the "spat" library in Max 8 for a VR playback of the spatialized sound on headphones. The spatial coordinates of the VR headset are transmitted to the Spat object for the rotation of the sound scene captured in 32-channels ambisonics before conversion to binaural.

First results of the scientific test

Thirteen participants agreed to pass the experience and perform the perceptive assessment on interaction and immersion in VR with a system simulating a dialogue. Overall, on all the experimental conditions, the participants noted a good interaction with their own voice (rated around 4.5/7 on average on all the conditions combined).

However, the results were not very variable depending on the experimental conditions, which corresponded to a variation of the real-time transformations. For the coming tests, it will therefore be a question of widening the range of effects in an attempt to observe results that are more differentiated depending on the conditions.

Furthermore, the results varied widely depending on the participants. It seems that participants who are more familiar with VR enjoyed the experience more, probably because they had more control over the system and were able to focus more on the voice interaction. It will therefore be for the coming tests to take into account more systematically the different profiles of the participants (naive, video game players, etc.) and to propose a familiarization step or a natural vs. transformed voice comparison to better reflect the interest of the system if necessary.

Finally, the participants told us in general comments that they generally greatly appreciated the experience, the originality of the scenario and the voice interaction. This experience generated relatively few symptoms of cybercinetosis despite its fairly long duration (30 min; it is generally recommended to limit a VR experience to

around 20 min maximum), probably because the movements of the image and sound were quite limited.

Outcomes

The feedbacks on this VR installation were very positive and the first results of the scientific test very encouraging. Our system is fully functional for carrying out perceptive tests with artistic components which provide material for scientific issues. Several improvements are nevertheless envisaged, in particular: the real-time transformation conditions, the variety of the vocal interactions to assess from when and to what extent we appreciate the effect obtained by the transformations on our own voice, while revising the experimental protocol to limit the time spent in VR. In addition, the whole system and the perceptive results obtained will allow us to feed our thinking for the rest of our work on this system: on the one hand, the extension of the scientific test, on the other hand, the production of an autonomous film in VR strongly inspired by this installation.

Acknowledgments

We would like to thank all of the people who were involved in this project: Piersten Leirom; Isabelle Viaud-Delmon, Olivier Warusfel and the whole Acoustic and Cognitive Spaces team of IRCAM; Jérémie Bourgogne, Cyril Claverie and all of IRCAM's Production; Greg Beller, Markus Noisternig, Paola Palumbo and the entire IRC department of IRCAM; Sebastian Rivas, Anouck Avisse and the GRAME team for a work residency at GRAME in 2019 and complementary to that of IRCAM.

References

- Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., & André, E. (2013). A job interview simulation: Social cue-based interaction with a virtual character. In *2013 International Conference on Social Computing* (pp. 220-227). IEEE.
- Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision*, *16*(11), 7-7.
- Dick, P. K. (1979). *Les Androïdes rêvent-ils de moutons électriques?*. JC Lattès.
- de Dinechin, G. D., & Paljic, A. (2018). 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)* (pp. 1-8). IEEE.
- Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. *Frontiers in psychology*, *6*, 249.
- Freud, S. (1919). L'inquiétante étrangeté et autres essais ([1985] éd.). *Paris: Folio*.
- Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

- Hoffman, E. T. A. (1815). *Le marchand de sable.*
- Isnard, V. (2016). *L'efficacité du système auditif humain pour la reconnaissance de sons naturels* (Doctoral dissertation, Paris 6).
- Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. (2016). Auditory sketches: very sparse representations of sounds are still recognizable. *PloS one*, *11*(3).
- Kopp, S., Jung, B., Lessmann, N., & Wachsmuth, I. (2003). Max-a multimodal assistant in virtual reality construction. *KI*, *17*(4), 11.
- Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22-32.
- Pörschmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica united with Acustica*, *86*(6), 1038-1045.
- Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, *3*(3), 253-262.

Presentation of the authors

Vincent Isnard

Vincent Isnard received a PhD in neuroscience from IRCAM, with a specialization in auditory perception, and holds three Masters in sound technologies and a Bachelor's degree in philosophy. He is a researcher, sound engineer and computer music designer. He also developed his contemporary musical practices in the classes of Laurent Durupt and Denis Dufour at the Conservatoire.

Trami Nguyen

Pianist, performer and visual artist, Trami Nguyen holds a master's degree from the HEM in Geneva. Co-founder of the Ensemble Links, she defends contemporary repertoires and creations of participatory, scenographic, immersive and/or multidisciplinary concerts. Her visual projects revolve around performances performed in Europe and extend to the field of virtual reality.