Does Synthetic Voice alter Social Response to a Photorealistic Character in Virtual Reality?

Katja Zibrek
Mimetic team, Inria Rennes
France
katja.zibrek@inria.fr

João P. Cabral
Trinity College Dublin
Ireland
cabralj@tcd.ie

Rachel McDonnell
Trinity College Dublin
Ireland
ramcdonn@tcd.ie

ABSTRACT

In this paper, we investigate the effect of a realism mismatch in the voice and appearance of a photorealistic virtual character in virtual reality. While many studies have investigated voice attributes for robots, not much is known about the effect voice naturalness has on the perception of realistic virtual characters. We conducted an experiment in Virtual Reality (VR) with over two hundred participants investigating the mismatch between realistic appearance and unrealistic voice on the feeling of presence, and the emotional response of the user to the character expressing a strong negative emotion (sadness, guilt). We predicted that the mismatched voice would lower social presence and cause users to have a negative emotional reaction and feelings of discomfort towards the character. We found that the concern for the virtual character was indeed altered by the unnatural voice, though interestingly it did not affect social presence.

CCS CONCEPTS

• **Computing methodologies** → **Perception**; • **Human-centered computing** → **Virtual reality**; **Sound-based input / output**.

KEYWORDS

virtual character, multimodal, perception, voice synthesis

ACM Reference Format:
Katja Zibrek, João P. Cabral, and Rachel McDonnell. 2021. Does Synthetic Voice alter Social Response to a Photorealistic Character in Virtual Reality?. In *Motion, Interaction and Games (MIG '21), November 10–12, 2021, Virtual Event, Switzerland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487983.3488296

1 INTRODUCTION

Real-time rendering technology has developed rapidly over the last decade and human likenesses have been represented virtually with increasingly impressive detail. We anticipate that photorealism will become commonplace in VR, and that virtual agents will resemble actual people in their appearance and movement.

However, other attributes of realism, such as the naturalness of an agent's voice, may additionally affect the perception of them. In sound analysis, there has been a lot of progress in creating artificial voices using speech synthesis, in order to generate voices which can modify a voice to sound more female or male, have a specific age, etc. These sound manipulations can create artefacts, and such a modified voice may be perceived unnatural.

While synthesised voices have been explored in relation to the effect they have on human perception, many aspects of social behavior when an unnatural voice is applied to a virtual agent in a highly immersive environment, such as VR, are not known. It has been shown that a mismatch in the realism of human face and voice can create a feeling of unease and discomfort when observing them [Mitchell et al. 2011]. This effect has been associated with the uncanny valley [Mori 1970], where objects which appear increasingly human-like are perceived as more familiar and pleasant. However, when they reach a specific level of human resemblance, remaining aspects of their inanimate nature create a mismatch with its apparent human-likeness, which results in a negative reaction from the observer. It is unknown if a mismatch in the naturalness of voice and realism of the agent's appearance in VR would trigger a negative reaction from the observer. In addition, due to a highly immersive nature of VR, virtual agents who behave realistically create the sense of "being there with another" (social presence, see [Biocca et al. 2001]). An unnatural voice may disrupt this illusion, especially if the appearance of the agent is photorealistic and causes the already mentioned uncanny valley effect.

In this paper, we investigate the effects of synthetic voice on social presence, comfort with the character and emotional response to the character. We designed an animated photorealistic character in virtual reality and manipulated the recorded voice from an actor with a high-quality speech synthesis and voice transformation tool. We used a type of synthesis which we predicted to be perceived unnatural but would still preserve some expressiveness of the human voice so emotions could be identified. We were interested if this mismatch between the synthetic voice and the photorealistic appearance of the character would reduce the comfort with the character, lower social presence and decrease appeal, familiarity and increase eeriness. To investigate this question, we conducted a between-subject experiment in VR, where 229 people's responses were recorded as they were reacting to sad, friendly and unfriendly emotional scenarios where the character had either a natural or unnatural voice. Interestingly, we did not find many effects of the unnatural voice on character perception apart from an interaction effect with the scenario, where participant's concern was affected.

2 BACKGROUND

In VR studies, autonomous virtual characters can induce a very strong sensation of being actually present and alive with the user,

MIG '21, November 10–12, 2021, Virtual Event, Switzerland

ACM ISBN 978-1-4503-9131-3/21/11…$15.00
https://doi.org/10.1145/3487983.3488296

Figure 1: Realistic virtual character in a living room environment used in the experiment.

commonly referred to as 'social presence', which is apparent by users' response to them. Particularly, maintaining personal distance from the character, similar to real life encounters with people, reveals user's comfort with the character and can be observed and measured in VR (proximity measure, see [Bailenson et al. 2003]). Sensory modalities, such as appearance, haptics and sound could play an important role in this illusion. There is some indication that the self-reported social presence is higher when observing photorealistic characters as opposed to more stylised rendered characters in VR [Zibrek et al. 2017; Zibrek and McDonnell 2019], and can also be increased when the appearance of the character matches its behavior [Zibrek et al. 2018]. There is some evidence showing the importance of haptics [Sallnäs 2010], while a study investigating audio in VR found a positive relationship between audio quality and the sense of social presence [Skalski and Whitbred 2010]. While there are many other determinants of social presence [Oh et al. 2018], not much is known about the importance of quality and naturalness of the character's voice.

There is also the question of how the realism of character's appearance impacts the perception of its voice and vice versa. Computer generated characters which appear almost human can sometimes induce negative emotional response such as disgust, eeriness, or fear, in humans. This aversive response was first described by Mori [Masahiro 1970] as the "uncanny valley". So far, research has identified some possible reasons for the uncanny valley, one of them being the mismatch in fidelity between different elements of character design [Saygin et al. 2012; Seyama and Nagayama 2007; Zell et al. 2015]. For example, disproportionately large eyes will appear more disturbing in realistic photographs than in images of an artificial character [Seyama and Nagayama 2007] and a realistic skin texture will appear less appealing on a character with exaggerated, unrealistic proportions [Zell et al. 2015]. Following this premise, a mismatch in fidelity between the voice and appearance would produce a similar uncanny effect. This was shown in the study of Mitchell et al. [2011]. However, this study only used a real human and a robot as the comparison. A recent study using animated virtual humans by Ferstl et al. [2021] demonstrated that the realism of voice is more preferable than the realism of appearance, confirming the importance of voice realism in character perception. Interestingly, this study also showed that maximizing voice naturalness is beneficial, even when it produces perceptual mismatches.

In addition to the uncanny valley, some voices may be more suitable to specific visual features of the characters. A study investigated voice attribution to robots of different appearance [Torre et al. 2021] and found that people assign voices according to social constructs, e.g., a male voice would be be assigned to a robot with more mechanical, metallic visual features.

Few studies can be found on the evaluation of expressive synthetic speech in the context of virtual characters, e.g. [Cabral et al. 2017; Potard et al. 2016]. There is a need for more research in this topic. Nevertheless, Cabral et al. [2017] report that the voice can have an impact on the avatar's communicative characteristics, in that participants perceived the character as more understandable, expressive and liked their voice more when using a human rather than a synthetic voice.

The work in this paper follows the line of the previous work by Cabral et al. [2017]. However, our work is not focused on evaluating expressive speech synthesis. Instead, we use a voice-morphing program to reconstruct expressive recorded human speech samples, so that the expressiveness aspects of the voice are preserved while the voice is manipulated so that it clearly sounds computer-generated. The idea is to study the possible uncanny effect on the perception of the virtual character. Additionally, our study is conducted in VR, and investigates social response. We also push the level of appearance realism higher than the work of Cabral et al.

3 EXPERIMENT DESIGN

3.1 Stimuli Creation

We chose the same photorealistic character as the work by Zibrek et al. [2019] which was obtained from Epic Games freely accessible Paragon character assets [UE4 2018a]. The environment and scenarios were also the same, built in a highly realistic environment from the Unreal Marketplace [UE4 2018b], using the same recordings of friendly, unfriendly and a sad performances. The sad scenario depicted a tragic situation, intended to induce empathy in the participant. The friendly scenario was intended to create a comfortable situation and a positive emotional response of the participant, while the unfriendly scenario intended to create an uncomfortable situation and a negative emotional response, which we believed would affect the proximity comfort.

3.2 Voice Synthesis

The synthetic speech stimuli were generated by using an high-quality speech manipulation system/vocoder called TANDEM–STRAIGHT [Kawahara et al. 2010; Kawahara et al. 2008]. This system enables the transformation of a number of voice features, including the pitch frequency, pitch range, speech rate, and vocal tract length. The voice transformation process is divided into three stages: analysis of the speech features, the feature transformation and reconstruction of the speech waveform. Each recorded speech signal was analysed to extract the STRAIGHT speech features: the fundamental frequency (F0), energy, aperiodicity, and spectrogram

features. Next, a step of manual processing of the speech features was performed using the TANDEM–STRAIGHT visual interface available for the Matlab development environment. The goal of the speech transformation performed in this work is to produce a voice that clearly sounds unnatural and computer-generated.

This approach for voice synthesis was chosen over using a text-to-speech (TTS) synthesis with expressiveness, such as in Rachman et al. [Rachman et al. 2018] as our system used motion capture for the body and facial animation, and aligning a TTS voice to the original audio to ensure synchrony with the animation and lip movements caused artifacts in the voice, making it unintelligible at points.

One type of transformation was to remove breathing and other non-vocalic sounds related to spontaneous voice by decreasing the energy to a very low value in each of the segments corresponding to those sounds. The next transformation was to draw lines over the voiced contours of the F0 plot to produce a smoother pitch contour. This manipulation has the effect of reducing the prosody variation and producing more artefacts in the synthetic speech due to variation of the F0 parameter. Another transformation was to decrease the vocal tract length (VTL) by a factor of 1/5. For example, this is a similar effect to that of transforming the adult female voice towards a child voice, and gave a cartoon-like effect. Note that this VTL transformation was only used to make the synthetic signal sound more artificial due to signal processing of transforming this parameter. Finally, the speech rate was reduced by a factor of 1/10. These two transformation factors were chosen by listening to the synthetic speech for different values within the allowed range of variation of the parameters. The criteria was to choose values that produced the desired effect of transforming the voice to sound more synthetic, but without introducing too much distortion so that the synthetic speech quality was still high and the speech intelligible.

The smoothing of the pitch also results in a "less human" sounding voice, because it reduces the expressiveness and richness in intonation of the recorded human voice.

3.2.1 Environment. This experiment was developed in Unreal Engine 4 and we used a HTC Vive system for virtual reality, with a tracking area of 2×1.5 metres. As in Zibrek et al. [2019], the character recited pre-recorded sequences but had interactive eye-gaze behaviour, to maintain eye-contact with the user, which is known to increase social presence [Bailenson et al. 2001]. Spatialized audio was also used to ensure that the sound recording came from the exact location of the agent's mouth.

4 EXPERIMENT

Our aim is to investigate peoples' responses towards a photorealistic virtual character with a synthetic voice. We formed the following hypotheses:

- **H1:** ***Synthetic voice will reduce participant's social presence with the character.*** We expect voice to be an important indicator of a character's believability, thus a synthetic voice will negatively impact social presence.
- **H2:** ***Synthetic voice will impact the perception of the character's traits and affect the emotional response of participants to the character expressing different types of emotions.*** We expect the synthesized voice to increase the mismatch with the realistic appearance of the character, affecting the perception of its traits and emotional expression, and dampen the empathetic response to a character in distress.
- **H3:** ***Synthetic voice will increase the discomfort with the character.*** We predict that the synthetic voice will make the character more uncanny (less appealing, less friendly, less realistic and more eerie) and increase the discomfort of standing in front of it in close proximity.

4.1 Measures

We used the same questionnaire as in Zibrek et al. [2019] to measure people's emotional response and other observations after interacting with the character. While we used a rather large tracking space for VR, we placed the participant intentionally close to the character in the room to investigate if they felt uncomfortable with the close proximity to it. This was the measure of proximity, a slight variation to the more usual measures of minimum distance towards the character (see [Bailenson et al. 2003]). The measure originates from anthropology [Hall 1966], where people stand further away from unfamiliar people and closer to familiar, pleasant people. Close proximity with an unpleasant character in VR may therefore cause discomfort, and participants indicated if this was the case by answering the question "When I first saw the girl in the room, I felt I was standing too close, I was in her intimate space" either with "Yes" or "No".

The next set of questions measured emotional response in the form of a 7-point Likert scale from 1 – Not at all to 7 – Extremely. First, participants were asked to what extent they felt "Concerned" [Davis 1983], "Excited", "Afraid" [Golan and Hill 2006] and "Calm" [Golan and Hill 2006] after observing the character. The next three questions asked about Affinity and Realism, based on measures previously used by McDonnell et al. [2012] ("Eeriness", "Appeal", "Familiarity" for Affinity, and "Overall", "Movement", "Appearance" and "Behaviour" Realism) . Finally, we tested for social presence based on the questionnaire by Bailenson et al. [2003] which consists of 5 questions related to the social presence with the character in VR. Finally, we asked about the place illusion [Slater 2009], a subjective response, where the participants were asked if they felt as if they were in a "living room". This question is related to the concept of presence, or "being there" in a virtual space [Skarbez et al. 2017; Zibrek et al. 2019].

4.2 Participants and Procedure

Our experiment was installed in a public setting in a Science Gallery museum, which is an international chain of art exhibition centres with a science outreach. Participants were members of the public attending an exhibition from diverse backgrounds who were introduced to the experiment by the gallery mediators. Participants were first asked to read an electronic consent form and fill out a demographics questionnaire. They were then shown to a VR booth, where the HMD was placed on their head and they were given the motion controller. The other instructions were the same as in Zibrek et al. [2019] where the participant was placed in a virtual living room with a virtual television, instructing them about the task.

The character started to speak when the participant directed his/her gaze towards it. The character's speech was either the pre-recorded or synthesised versions of the voice, for either the friendly, unfriendly, or sad scenario. Following the sequence, the participant was asked to answer the questions about proximity comfort, character's traits, their own emotional response, perceived realism, affinity towards the character and their feeling of being present in space and with the character, and the experiment terminated.

4.3 Analysis

375 volunteers participated in the experiment. We first reviewed the data for possible exclusion. We had two major exclusion criteria: under 18 years of age and missing answers. Following the general data protection policy, where underaged participants need guardian approval for using their data, such data was immediately deleted and excluded from analysis. Additionally, if the participant had equal or less than 50% of the questions answered, his or her data were not included in the analysis. The reason for so many exclusions can be attributed to the setting of the experiment, which was a public gallery and the participants were its visitors. The visitors may have been less motivated to finish the experiment and would leave before the experiment was over.

Following this procedure, we included responses of 229 participants in the final analysis, which was approximately 38 participants for each scenario/voice condition combination.

Participants were aged from 18 to 77 (average age = 27), of which 107 were male, 113 were female and 9 did not provide an answer. Gender was approximately balanced in each scenario/voice group (average number of females = 18, males = 17).

5 RESULTS

To explore the effects of voice type (*Real, Synthetic*) and scenario (*Sad, Friendly, Unfriendly*) on people's subjective responses, we analyzed the subjective scales separately. The measured scales were: Proximity, Concerned, Excited, Afraid, Comfortable, Appeal, Eerie, Familiar, Overall Realism, Movement realism, Appearance Realism, Social Presence and Place Illusion.

We analyzed the results by using ANOVA with between–subject factor Scenario and Voice type. ANOVA is generally robust for violations of normality and our sample per each factor group was larger that 20. Therefore, we chose to use a parametric ANOVA. A non-parametric equivalent (Kruskal-Wallis one-way ANOVA) was used with categorical data (Proximity answers) and when homogeneity of variance was breached, which was identified by performing the Levene's test. We used Tukey's HSD for the post-hoc tests. The most important results are shown in Figure 2.

The 5 measured items of the Social Presence scale were tested for reliability and due to sufficient correlation (Cronbach's alpha: $\alpha = 0.63$), we used a cumulative score of all 5 items as the final result and treated it as a continuous scale, as is custom.

5.1 H1: Voice type and social presence

We were first interested if the character's Voice type will affect the feeling of social presence with it. We did not find support for this, as no main or interaction effects were found with Social Presence.

5.2 H2: Voice type, traits and emotions

We found a significant main effect of Scenario for the variable Concerned ($F(2, 223) = 7.695, p = 0.001$). Participants were more concerned when watching a Sad character as opposed to a Friendly one ($p < 0.001$). This was expected, since the Sad scenario featured a character in distress, while the character in the Friendly scenario was happy and free of concern.

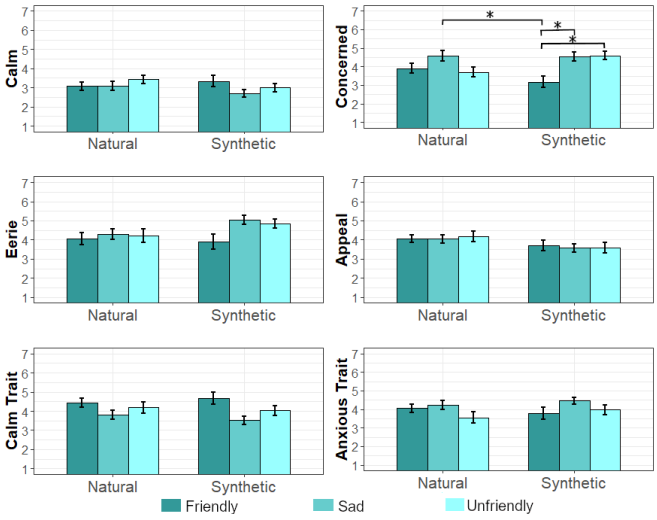

Figure 2: Interaction between Voice type and Scenario for selected dependent variables. Lines above bars signify significant differences, $* = p < 0.05$.

More importantly, the effect was further influenced by Voice type ($F(2, 223) = 4.651, p = 0.011$) - participants were least concerned when the Friendly character had a Synthetic voice when compared to other Synthetic voice scenarios (all $p < 0.005$) and compared to Natural voice / Sad scenario combination, see Figure 2. It would appear that the Synthetic voice made a stronger polarisation of the concern participants felt after watching the character, especially between Friendly and negative scenarios (Sad, Unfriendly) in the Synthetic voice condition. Interestingly, we did not find any other effects of Voice type on our dependent variables.

There was also a significant main effect of Scenario on the perception of the character's trait Calm ($F(2, 223) = 6.287, p = 0.002$). The character in the Sad condition was perceived as least Calm, especially when compared to Friendly character ($p < 0.001$).

Unexpectedly, we did not get any other differences in emotional responses according to Scenario. Unfriendly character did not increase fear (Afraid) or decrease Calm, Friendly did not increase excitement. In addition, the character in the Sad scenario was not perceived to be significantly more Anxious than Friendly and Unfriendly, even though it received higher ratings.

5.3 H3: Voice Type and discomfort

We did not find any effects of Voice or Scenario on the variables Appeal, Familiar. Figure 2 shows higher ratings in the Synthetic Voice condition for Eerie, however, the effect was not significant.

The analysis on Proximity revealed differences according to the Scenario (Kruskal-Wallis test: $H(2, N = 229) = 7.173, p = 0.027$), where the participants reported higher discomfort with the closeness of the character in the Sad condition, and least for the Friendly condition, however, the pairwise comparisons did not reveal any significant differences.

5.4 Other results

Unexpectedly, there was a main effect of Scenario on Movement Realism ($F(2, 221) = 4.566, p = 0.011$), where the character in the Sad scenario was perceived to have more realistic motion, especially when compared to the Friendly Scenario ($p < 0.01$). This could be due to particular nonverbal expressions of emotions in this scenario.

Participant in the Synthetic voice condition did not perceive the voice to be particularly natural ($\overline{x} = 2.3, SD = 1.2$), which was our aim.

Overall, participants felt a relatively high level of Place Illusion ($\overline{x} = 5.96$, $SD = 1.09$) and there was no effect of Voice Type or Scenario on Place Illusion.

Social Presence was in the medium range across all conditions ($\overline{x} = 19.9, SD = 4.9$). While we attempted to create a photorealistic character, the mean rating for Appearance Realism was medium ($\overline{x} = 3.76$, $SD = 1.57$), similarly for Overall realism ($\overline{x} = 3.67$, $SD = 1.57$), while Behaviour realism was slightly higher compared to other realism assessments ($\overline{x} = 4.17$, $SD = 1.51$).

Overall, the characters received medium ratings on Appeal, were perceived as slightly more eerie and less familiar (Appeal: $\overline{x} = 3.9, SD = 1.5$; Eerie: $\overline{x} = 4.4, SD = 1.8$; Familiar: $\overline{x} = 2.8, SD = 1.5$).

6 GENERAL DISCUSSION

In this study, we investigated the effect of an unnatural, synthetic voice on the perception of an emotional virtual character. We expected the synthetic voice to: reduce social presence (H1), impact perception of traits and emotional response to the character (H2) and increase discomfort (H3). H1 and H3 were rejected, and we found only a partial confirmation for H2. This result is interesting as it would be expected that the obvious distortion we made to the naturalness of voice would have a greater impact on the perception of the character. It is possible that the voice transformation method retained many of the characteristics of a natural voice, which may not be the case with more robotic transformations or synthetic voices with low expressiveness. Therefore, even though participants did not rate the voice as natural, it might have carried enough relevant information about the character's emotion, gender, etc. Another possibility is that the voice we choose had a cartoon-like quality which may have made it less eerie and more appealing than other types of synthesized voices that lack emotion. Different levels of voice distortion should be used in future work to understand their effect on character perception. Also, testing voices that more closely match current text-to-speech would be interesting.

We did find an effect of voice on the level of concern participants felt after viewing the character in different scenarios. The friendly character with the synthetic voice was the condition where concern was the lowest, which is an expected response. Perhaps the transformation method we used removed some nuance from the actor's voice and therefore participants could focus on the content of what was said. A less obvious distortion to the voice could probably reduce or completely remove this effect, hence a wider range of voice synthesis approaches should be explored within emotional scenarios.

We also did not find strong reactions to the scenario, e.g., the unfriendly character did not evoke fear and friendly character was not particularly exciting. We did find that the sad character was perceived as least calm and least comfortable to stand close to. It appears that rather than evoking empathy, this character was perceived as more unsettling, which is not in line with previous study investigating proximity with the empathetic character [Zibrek et al. 2019]. However, this could be due to the fact we used only a character of realistic appearance, while the previous study included three levels of stylisation. In future work, we could explore different voice transformation approaches on various stylisation levels of characters in order to better understand the relationship between voice and character appearance. Additional scenarios could also be explored, preferably pre-tested for the emotional effect they have on the users.

The chosen measure of discomfort, which we implemented by placing the participant in close proximity with the character, may have had its drawbacks as well. Participants could have moved away from the character prior to hearing her speak, therefore the voice might have not affected the proximity in the same way across all participants. Also, proximity differs according to the cultural background of participants. While the visitors to the Science Gallery typically come from various countries, we did not specifically ask for this information in our questionnaire. This would have been a valuable addition to the experiment.

Our character received medium ratings of appeal and realism, and participants also perceived it as slightly more eerie and less familiar, regardless of the voice condition. It is possible that we did not achieve a sufficient level of appearance realism in order to create a mismatch with the synthetic voice, which could explain the lack of more noticeable differences between natural and synthetic voice conditions. Our chosen character from Paragon [UE4 2018a] was recently considered state-of-the-art in photorealism for virtual characters in AAA-games, but the past few months have seen the introduction of characters such as Metahumans by Epic Games[1], which have a much higher quality facial rig and textures. We believe that the added realism could induce different responses, which will be investigated in future work.

Apart from these limitations, our results indicate that even obvious modifications to the naturalness of voice do not interfere considerably with our social presence, likeability, and emotional response to a virtual character in VR. Some subtleties of emotional communication could be affected though, so the strive to improve the naturalness of synthesised voice does have its merit. We believe our study is a small part of a larger scale research investigating synthesised voices for realistic virtual humans in immersive environments.

ACKNOWLEDGMENTS

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at

[1] https://metahuman.unrealengine.com

the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. Also, funded by SFI project RADICal (Grant No. 19/FFP/6409).

REFERENCES

Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence* 10, 6 (2001), 583–598.

Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2003. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin* 29, 7 (2003), 819–833.

Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.

Joao Cabral, Benjamin Cowan, Katja Zibrek, and Rachel McDonnell. 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. 229–233. https://doi.org/10.21437/Interspeech.2017-325

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113.

Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*. 76–83.

Baron-Cohen Simon Golan, Ofer and Jacqueline Hill. 2006. The Cambridge Mindreading (CAM) Face-Voice Battery: Testing Complex Emotion Recognition in Adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders* 36, 2 (2006), 169–183.

Edward Twitchell Hall. 1966. The hidden dimension. (1966).

Hideki Kawahara, Masanori Morise, Toru Takahashi, Hideki Banno, Ryuichi Nisimura, and Toshio Irino. 2010. Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems. 38–41.

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 3933–3936.

Mori Masahiro. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.

Rachel McDonnell, Martin Breidt, and Heinrich Buelthoff. 2012. Render me Real Investigating the Effect of Render Style on the Perception of Animated Virtual Humans. *ACM Transactions on Graphics* 31, 4 (2012), 91:1–91:11.

Wade J Mitchell, Kevin A Szerszen Sr, Amy Shirong Lu, Paul W Schermerhorn, Matthias Scheutz, and Karl F MacDorman. 2011. A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2, 1 (2011), 10–12.

M. Mori. 1970. The Uncanny Valley. *Energy* 7, 4 (1970), 33–35.

Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Front. Robot. AI 5: 114. doi: 10.3389/frobt* (2018).

Blaise Potard, Matthew Aylett, and David Braude. 2016. Cross Modal Evaluation of High Quality Emotional Speech Synthesis with the Virtual Human Toolkit, Vol. 10011. 190–197. https://doi.org/10.1007/978-3-319-47665-0_17

Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. 2018. DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods* 50, 1 (Feb. 2018), 323–343. https://doi.org/10.3758/s13428-017-0873-y

Eva-Lotta Sallnäs. 2010. Haptic feedback increases perceived social presence. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 178–185.

Ayse Pinar Saygin, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. 2012. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience* 7, 4 (2012), 413–422.

Jun'ichiro Seyama and Ruth S Nagayama. 2007. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments* 16, 4 (2007), 337–351.

Paul Skalski and Robert Whitbred. 2010. Image versus sound: A comparison of formal feature effects on presence and video game enjoyment. *PsychNology Journal* 8, 1 (2010).

Richard Skarbez, Solene Neyret, Frederick P Brooks, Mel Slater, and Mary C Whitton. 2017. A psychophysical experiment regarding components of the plausibility illusion. *IEEE transactions on visualization and computer graphics* 23, 4 (2017), 1369–1378.

Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364, 1535 (2009), 3549–3557.

Ilaria Torre, Fethiye Irmak Dogan, and Dimosthenis Kontogiorgos. 2021. Voice, Embodiment, and Autonomy as Identity Affordances. In *HRI 2021-Robo-Identity: Exploring Artificial Identity and Multi-Embodiment March 2021*.

UE4. 2018a. Paragon Phase. Unreal{E}ngine4,{P}aragon{P}hase.https://www.unrealengine.com/marketplace/paragon-phase.

UE4. 2018b. Unreal Engine 4, Realistic Rendering. https://docs.unrealengine.com/en-us/Resources/Showcases/RealisticRendering.

Eduard Zell, Carlos Aliaga, Adrian Jarabo, Katja Zibrek, Diego Gutierrez, Rachel McDonnell, and Mario Botsch. 2015. To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–12.

Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. 2017. Don't stand so close to me: investigating the effect of control on the appeal of virtual humans using immersion and a proximity-based behavioral task. In *Proceedings of the ACM Symposium on Applied Perception*. ACM, 3.

Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. 2018. The Effect of Realistic Appearance of Virtual Characters in Immersive Environments-Does the Character's Personality Play a Role? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1681–1690.

Katja Zibrek, Sean Martin, and Rachel McDonnell. 2019. Is Photorealism Important for Perception of Expressive Virtual Humans in Virtual Reality? *ACM Trans. Appl. Percept.* 16, 3, Article 14 (Sept. 2019), 19 pages. https://doi.org/10.1145/3349609

Katja Zibrek and Rachel McDonnell. 2019. Social presence and place illusion are affected by photorealism in embodied VR. In *Motion, interaction and games*. 1–7.