Computers & Graphics 104 (2022) 116–128

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on MIG 2021

Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans

Darragh Higgins [a], Katja Zibrek [b], Joao Cabral [a], Donal Egan [a], Rachel McDonnell [a,*]

[a] Trinity College Dublin, Ireland
[b] INRIA Rennes, France

ARTICLE INFO

Article history:
Received 18 October 2021
Received in revised form 11 February 2022
Accepted 27 March 2022
Available online 31 March 2022

Keywords:
Virtual humans
Perception
Synthetic voice
Virtual reality

ABSTRACT

In this paper, we investigate the effect of a realism mismatch in the voice and appearance of a photorealistic virtual character in both immersive and screen-mediated virtual contexts. 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 the first 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. 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. The second experiment was conducted with a view towards heightening the appearance realism of the same character for the same scenarios, with an additional lower level of voice realism employed to strengthen the mismatch of perceptual cues. While voice type did not appear to impact reports of empathic responses towards the character, there was an observed effect of voice realism on reported social presence, which was not detected in the first study. There were also significant results on affinity and voice trait measurements that provide evidence in support of perceptual mismatch theories of the Uncanny Valley.

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.

Although this aspect of human-like virtual character composition is important for verisimilitude, other aspects of character creation can play a role in perceived realism, such as the naturalness of a characters vocal expressions. In sound analysis, there has been a lot of progress in creating artificial voices using speech synthesis, in order to generate voices that can be modified to include specific features, to sound more female or male, have a particular age, etc. These sound manipulations can create artefacts, and such a modified voice may be perceived as unnatural.

While synthesised voices have been explored in relation to the effect they have on human perception, many aspects of social behaviours directed at highly realistic virtual agents with unnatural voices have not been fully explicated. It has been shown that a mismatch in the realism of human face and voice can create a feeling of unease and discomfort in experiences of quasi-realistic virtual characters [1]. This effect has been associated with the Uncanny Valley [2], wherein objects that are increasingly human-like are perceived as more familiar and pleasant. However, once they reach a specific level of near accurate human resemblance, remaining aspects of the object's inanimate nature create a mismatch with its apparent human-likeness, resulting 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 either screen-based interactions or immersive virtual environments, could trigger a negative reaction from the observer. In addition, the particularly immersive nature of VR has seen interactions with , virtual agents who behave realistically, which can create the sense of "being there with another" (social presence, see [3]). An unnatural voice may disrupt this illusion, especially if the appearance of an agent is photorealistic and causes the already mentioned uncanny valley effect. Furthermore,

* Corresponding author.
E-mail addresses: HIGGIND3@tcd.ie (D. Higgins), katja.zibrek@inria.fr (K. Zibrek), CABRALJ@tcd.ie (J. Cabral), doegan@tcd.ie (D. Egan), ramcdonn@tcd.ie (R. McDonnell).

https://doi.org/10.1016/j.cag.2022.03.009
0097-8493/

with the advancement of animation technology, it remains unclear whether greater anthromporphisation of digital humans can overcome perceptions of uncanniness, or whether they can create even greater mismatches when combined with low levels of voice realism.

In the first study outlined below, we investigated the effects of synthetic voice on social presence, comfort with the character and emotional response to the character in an immersive VR environment. 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.

An additional experiment was designed with the same scenarios, to expand the range of mismatch conditions that participants were exposed to in the first experiment [4]). The format was that of a screen-based human–character interaction study, designed to test the effect of higher appearance realism of the virtual character, with an extra level of voice realism chosen for its more salient synthetic composition. Many of the aspects of the first study, the character's actions and expressions, the questionnaire items and the original voice synthesiser, were retained. The highly realistic character for our second experiment was presented on a screen instead of 3D immersive VR, as 2D screens allow for higher quality render detail than 3D environments. The contrasting low realism voice clips were generated using a Text-To-Speech system, which is the current industry standard voice for interactive agents or chatbots. We added extra questions to assess the effect of voice in this second experiment. We also included additional questions designed to assess whether subjects expressed feelings empathy for the realistic characters. The motivation for our second experiment stemmed from studies which describe an influence of mismatched modes of realism in virtual characters on perceived appeal and discomfort in human–agent interactions [1,5]. We intended to add greater depth to the mismatch of conditions from those used in the first study, to evaluate any potential effect of less consistent realism in the character's features.

2. Background

In VR studies, autonomous virtual characters can induce a very strong sensation of being actually present and alive with the user, commonly referred to as 'social presence', which is apparent by users' response to them. The term refers to a sense of 'being together' with another, involving the modelling of mental states and theories of other minds. This is an aspect of presence distinct from the positional notion of telepresence which mediates modelling of spatial environments in virtual contexts [6], and which has links to illusory sensations of embodied placement [7].

2.1. Social presence and virtual characters

Particularly, maintaining personal distance from a character, similar to real life encounters with people, reveals user's comfort with the character and can be observed and measured in VR to influence social presence (proximity measure, see [8]). 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 [9,10], and can also be increased when the appearance of the character matches its behaviour [11]. There is some evidence showing the importance of haptics [12], while a study investigating audio in VR found a positive relationship between audio quality and the sense of social presence [13]. While there are many other determinants of social presence [14], not much is known about the importance of quality and naturalness of the character's voice, although there is evidence that the close resemblance of an avatar's voice to one's own voice can lead to greater senses of immersion and stronger task performance in virtual gaming environments [15].

Moreover, it has been proposed that mismatching human-like conversation qualities can negatively sway perceived social presence [16]. Social presence has been measured to increase in experiments using synthetic voices where verbal utterances are matched deliberately with corresponding personality cues [17] suggesting that social presence measurements could be altered by a misalignment of visual and vocal realism.

2.2. The uncanny valley

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 [18] 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 [19–21]. For example, disproportionately large eyes will appear more disturbing in realistic photographs than in images of an artificial character [20] and a realistic skin texture will appear less appealing on a character with exaggerated, unrealistic proportions [21]. Similarly, realistic human operated avatars in both screen-based environments and immersive VR settings have been shown to benefit from improved behavioural realism by means of accurate representations of non-verbal behaviour [22] implicating the necessity for aligned behavioural and visual affordances.

Following this premise, a mismatch in fidelity between the voice and appearance could be proposed to produce an uncanny effect. This was shown in the study of Mitchell et al. [1]. 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. [23] 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 maximising voice naturalness is beneficial, even when it produces perceptual mismatches.

2.3. Synthetic voices

In addition to the uncanny valley, some voices may be more suitable to specific visual features of the characters. For instance, it has been proposed that a useable vocal agent should adopt matching cues, suggesting that a visually robotic character should have a robotic voice for the purpose of interactive functionality [24], which is supported by uncanny valley theories of realism inconsistency [25]. Similarly a study which investigated voice attribution to robots of different appearance [26] found that people assign voices according to social constructs, e.g., a male voice would 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. [27,28]. There is a need for more research in this topic. Nevertheless, Cabral et al. [27] 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.

Text-To-Speech (TTS) systems, while generally lacking in elements of human expressivity, have advanced in strength since their inception. TTS has been employed in similar studies of virtual characters to assess the influence of synthetic voices on persuasiveness and credibility ratings [29], as well as investigations into the perception of McGurk effects in animated digital humans [30]. These studies, provided us with the scope to run a second study using Text-To-Speech voice generation as a low level of voice realism.

2.4. Photorealism, affect and multimodal cues

The secondary experiment described here was also influenced by work which has sought to measure the influence of photorealism on reports of perceived uncanniness [31] as well as affective responses [32] and the role of empathy in these types of affective situations [33]. While affected scenarios have been observed to influence social presence and emotional responses in screen-based avatar–avatar studies [34], we identified a niche related to our first experiment where photorealism could be used to assess interactions with virtual humans under these kinds of conditions.

Developments in animation technology for virtual humans have allowed for strong mismatches in appearance fidelity. High levels of human-like realism have been recently measured to be considered more appealing and less uncanny than lower quality ones [35], while mismatches between human and human-like voice and appearance combinations have been demonstrated to impact human judgement and comprehension in human–agent interactions [36]. Indeed, since the anthropomorphisation of agents have been shown to compensate for low-quality communication in chatbot interactivity [37], a gap in the literature exists for further investigation on the effects of high fidelity digital humans when combined with unrealistic voices.

This higher level of photorealism gave us motivation to include a low level of voice realism for our second experiment, in order to deliberately invoke the effect of violated expectations which arise from misaligned perceptual cues. This effect has been linked to the uncanny valley [38] and has been supported by brain imaging studies on incongruent appearance and movement [39].

There are indications that multimodal cues are important for speech intelligibility, when comparing audio-visual to audio only conditions [40], which demonstrates the extent to which matching cues may be significant for successful communication. Indeed, in user interactions with agents, there is evidence to suggest that auditory, speech based agent feedback can be considered preferable to visual or behavioural feedback when each modality is tested in isolation, although such evidence also suggests a preference for multimodal feedback [41]. Screen-based studies have previously investigated multimodal cue mismatches for the purpose of studying the uncanny valley [42], under the framework of Bayesian modelling. Bayesian models are used to quantify human cognitive prediction processes, which can give credence to studies which link prediction errors to the occurrence of the uncanny valley effect. This has been particularly successful under conditions of mismatched voice and appearance [43] such as ours.

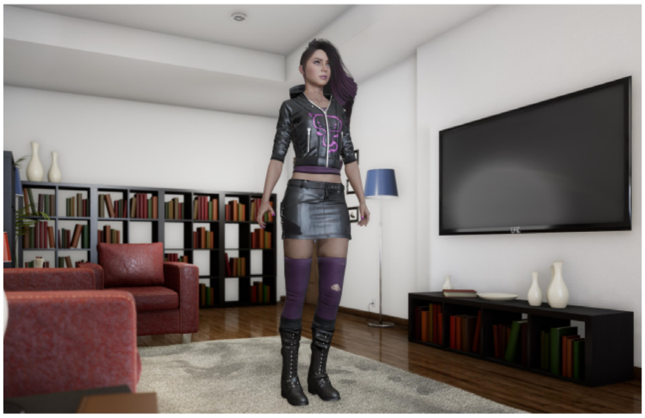

Fig. 1. Virtual character in a realistic living room VR environment used in Experiment 1.

3. Experiment 1

Our aim is to investigate peoples' responses towards a photorealistic virtual character with a synthetic voice in an immersive 3D virtual environment. 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 synthesised 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.

3.1. Stimuli creation

We chose the same photorealistic character and environment as the work by Zibrek et al. [44] which was obtained from Epic Games freely accessible Paragon character assets[1] and Unreal Marketplace[2] (Fig. 1).

We also used the same 3 scenario recordings as in [44]: friendly, unfriendly and sad. 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 was intended to create an uncomfortable situation and a negative emotional response, which we believed would affect the proximity comfort.

The performances of the scenarios were captured from a single female actor, whose body, face, and voice were recorded using state of the art motion capture and audio hardware. The actor was instructed to act out the pre-scripted scenarios, and the recorded body and facial motion was then applied to the virtual agent's rig. See supplemental movie for stimuli examples.

[1] https://www.unrealengine.com/marketplace/paragon-phase

[2] https://docs.unrealengine.com/en-us/Resources/Showcases/RealisticRendering

3.2. Voice synthesis

The synthetic speech stimuli were generated by using an high-quality speech manipulation system/vocoder called TANDEM–STRAIGHT [45,46]. This system has been widely used in the research community and it has been maintained and improved further by the authors. We have used the Matlab GUI from 2014, for Windows OS, which incorporates more recent work which has developed frameworks for improving the system's voice morphing capabilities [47]. 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.

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. [44], 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 [48]. Spatialized audio was also used to ensure that the sound recording came from the exact location of the agent's mouth.

3.3. Measures

We used the same questionnaire as in Zibrek et al. [44] 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 [8]). The measure originates from anthropology [49], 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" [50], "Excited", "Afraid" [51] and "Calm" [51] after observing the character (See Table 1, Emotional Response). The next three questions asked about Affinity and Realism, based on measures previously used by McDonnell et al. [52] ("Eeriness", "Appeal", "Familiarity" for Affinity, and "Overall", "Movement", "Appearance" and "Behaviour" Realism). We also tested for social presence based on the questionnaire by Bailenson et al. [8] which consists of 5 questions related to the social presence with the character in VR (See Table 1, Social Presence). Finally, we asked about the place illusion [7], 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 [44,53].

3.4. 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. [44] 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.

3.5. 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.

Table 1
Questions arranged by group and variable name. Each statement could be answered on a scale from 1 – "Not at all" to 7 – "Extremely".

Group	Variable name	Statement
Emotional Response	Concerned	*"The girl I just observed made me feel concerned."*
	Excited	*"The girl I just observed made me feel excited."*
	Afraid	*"The girl I just observed made me feel afraid."*
	Calm	*"The girl I just observed made me feel calm, at ease:"*
Affinity	Appeal	*"I found the girl appealing, likable."*
	Eerie	*"I found the girl eerie, creepy."*
	Familiar	*"I found the girl familiar, I have seen a similar person before."*
Realism	Overall Realism	*"I found the girl realistic overall."*
	Appearance Realism	*"I found the girl's appearance realistic."*
	Movement Realism	*"I found the girl's movements realistic."*
	Behaviour Realism	*"I found the girl's behavior realistic."*
Social Presence	Item 1	*"It feels as if I am in the presence of another person in the room with me."*
	Item 2	*"It feels as if the girl is watching me and is aware of my presence."*
	Item 3	*"The thought that the girl isn't real crossed my mind often."*
	Item 4	*"The girl appears to be alive."*
	Item 5	*"The girl is only a computerized image, not a real person."*
Empathy	Item 1	*"The character's emotions are genuine."*
	Item 2	*"I can feel the character's emotions."*
	Item 3	*"I can see the character's point of view."*
	Item 4	*"The character's reaction to the situation is understandable"*
	Item 5	*"When watching the video, I was fully absorbed"*
	Item 6	*"I can identify with the character in this message"*
Voice Traits	Voice Likeability	*"How much did you like the character's voice?"*
	Voice Consistency	*"How well did you think the voice matched the character's appearance?"*
	Voice Expressiveness	*"How expressive did you find the character's voice?"*
	Voice Understanding	*"How easy did you find it to understand what the character was saying?"*

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).

3.6. Results

To explore the effects of voice type (*Real, Synthetic*) and scenario (*Sad, Friendly, Unfriendly*) on people's subjective responses, we analysed 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 analysed 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 Fig. 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.

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.

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.

More importantly, the effect was further influenced by Voice type ($F(2, 223) = 4.651, p = 0.011$) - participants were least

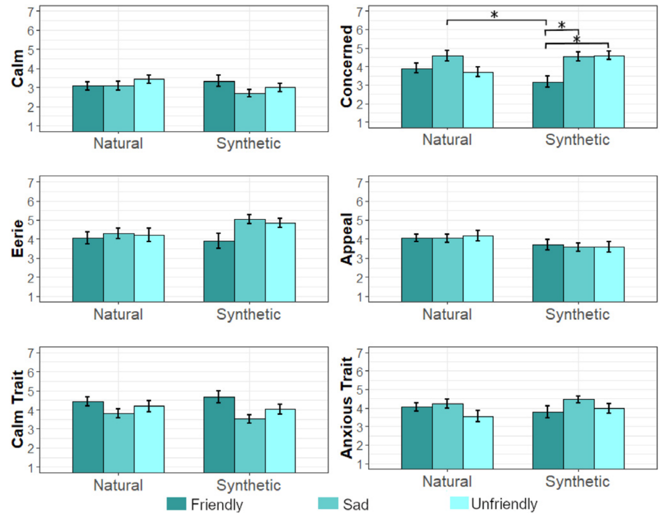

Fig. 2. Interaction between Voice type and Scenario for selected dependent variables. Lines above bars denote significant differences, $* = p < 0.05$.

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 Fig. 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.

H3: Voice type and discomfort

We did not find any effects of Voice or Scenario on the variables Appeal, Familiar. Fig. 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.

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.

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

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

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

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

3.7. 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 synthesised voices that lack emotion. The experiment design posited in Section 4 sought to investigate this possibility with different levels of voice synthesis.

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 [44]. 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. Our work in Section 4 explores this relationship in the opposite direction, using a more realistic representation of the character. 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.

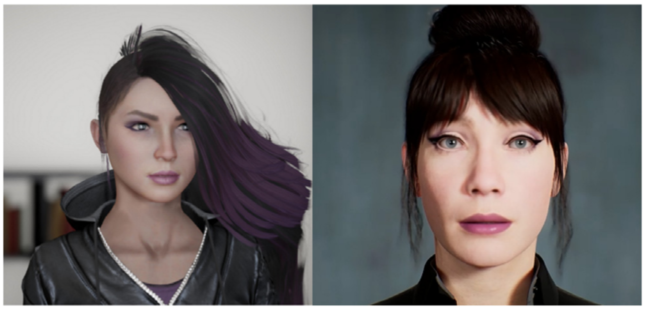

Fig. 3. Comparison of realistic virtual character used in Experiment 1 (left) and the higher-quality Metahuman that was created to match her appearance in Experiment 2 (right).

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 for Experiment 1 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,[3] which have a much higher quality facial rig and textures (Fig. 3). We believe that the added realism could induce different responses, and therefore designed a screen-based experiment based on the above work to explore the effects of heightened photorealism.

4. Experiment 2

Given that the aforementioned Metahuman characters represent the foremost quality of photorealistic virtual humans in circulation, our second experiment was designed to combine less realistic synthetic voices, with higher photorealism that these characters provide. We expected that the greater salience in perceptual mismatch cues would magnify previously outlined results on affinity and emotional response. In terms of our previously outlined hypotheses, we sought evidence that this stronger mismatch in voice and appearance realism could give us further confirmation of H2 *That the synthetic voice conditions would impact perception of character's traits and affect the emotional response of participants to the character expressing different types of emotions*, as well as some evidence for H3, *That synthetic voice would increase discomfort with the character*. Previous methodologies in this domain offer support for H3 under heightened mismatch conditions, as outlined by Mitchell et al. [1].

We also expected low scores for the social presence and realism measures used in the above design, given that the experiment was conducted on screen instead of in a virtual environment. This decision was forced by our goal of increased appearance realism, as the highest level of animation detail for the Metahuman characters (LOD 0) has been observed to induce reports of greater affinity than the highest available level for VR applications (LOD4). Previous work has identified differences in reported affinity between these two levels of detail [35], which provided us with motivation to test this realism under mismatch conditions (see Fig. 4). As well as these considerations, a further feature of our second study was an extra dimension of voice realism. Using a commercial Text-to-Speech (TTS) system allowed us to expand the range of realism conditions for voice in the opposite direction to our appearance realism manipulation. In doing so, we also aimed to replicate and develop findings in similar experiments that have investigated the role of synthetic voices in virtual characters [27], but without the less realistic TTS condition.

Finally, there have been recent studies which have expanded the work on empathy towards virtual characters, in virtually mediated environments. There is a role to be played by empathetic responses in immersion and perspective taking of virtual characters [54]. Perspective-taking, also termed 'mentalising', is considered to be an important process in both cognitive and affective empathy, and describes the the act of attributing mental states to others, and considering the relationship of these states to the causal properties of the environmental scenario or context [55]. Moreover, an aspect commonly measured in research which studies the occurrence of the uncanny valley is that of 'familiarity', which has been observed to influence empathy towards digital humans expressing pain [56]. With the heightened photorealism for our second experiment, we aimed to explore whether the characters could induce empathy in participants.

4.1. Stimuli creation

The virtual human stimuli for our second experiment were developed in Unreal Engine 4.26,[4] combined with the Metahuman Creator tool.[5] and tracked at the highest available level of detail (LOD 0) from video of the actor performing the three scenarios also used in the experiment described above. The higher animation quality was defined by a greater number of blendshapes (669) and joints (713) in the facial rig, as well as being highly geometrically complex (24000 facial vertices). This motivated our move towards maximal realism which necessitated our conducting a screen-based experiment. Our aim was to further investigate the effects on emotion and affinity at high levels of photorealism. The scenarios from Experiment 1 were re-recorded by our actor via a front-facing phone camera. We then used Faceware tracking software to retarget the performance to generate our Metahuman stimuli. The character used for this experiment was designed to closely resemble the Paragon character from the first experiment for the sake of consistency (Fig. 3 Since our viewpoint was fixed in this experiment, We chose a closer camera viewpoint than in Experiment 1 to increase intimacy and allow the facial performance to be viewed in more detail.

4.2. Voice synthesis

There were three levels of voice realism used in this experiment. Firstly, we used the original natural voice of the actor whose motion was captured for generating the character animations. The second level was the TANDEM–STRAIGHT [45,46] vocoder system outlined above, which was included for congruence with the previous work. However, in this experiment we only transformed the VTL parameter, by the same factor of 1/5. The reason for not applying the transformation of the other speech parameters, as in the first experiment, was that for this stimuli it introduced excessive audible speech distortion, which could have an undesirable effect in the consistency of the results in regard of speech quality. From our own listening tests we found that the synthetic speech sounded similarly robotic/artificial when compare with the stimuli of experiment 1, which was our goal. By avoiding the transformation of the other parameters, particularly F0 and speech rate, we also aimed to preserve as much as possible the expressive aspects of the recorded voice, in this experiment. We also created voice clips from commercial Text-To-Speech software IBM Watson[6] which

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

[4] https://www.unrealengine.com

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

[6] https://www.ibm.com/watson

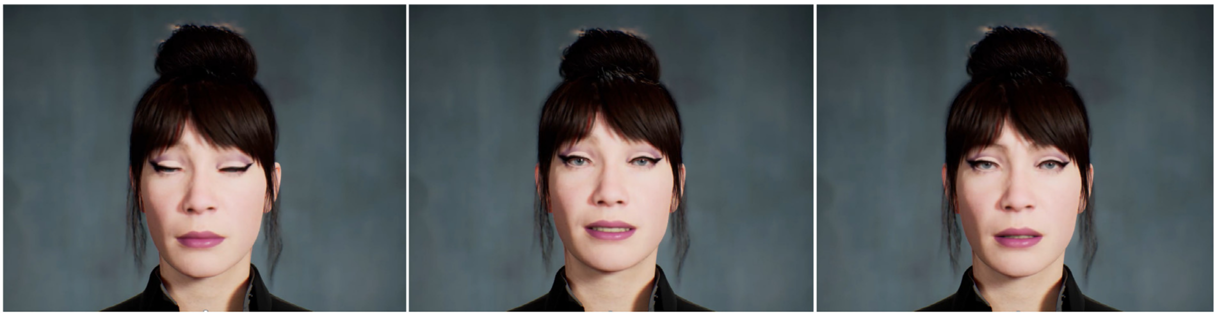

Fig. 4. Screenshots from scenarios (left) Sad, (middle) Friendly, and (right) Unfriendly for character in Experiment 2. Participants viewed the character's performance with either a natural, synthetic or text-to-speech voice.

is the current industry standard for virtual agents and chatbots. The voice clips were created with Python at prosody calculated from the duration of the natural voice videos, and then manually mapped to the face movements. The motivation for inclusion of TTS vocals was twofold as it is more widely used in the industry than our previous synthetic voice condition and it also provided a lower level of vocal realism, which allowed us to expand the mismatch of perceptual cues when combined with greater appearance fidelity.

4.3. Measures

In keeping with the work outlined above, we included the same questionnaire validated by Zibrek et al. [57] for use in studies which focus on interactions with virtual characters, through a combination of established metrics for measuring social presence [8], emotional responses [50,51,58] and perceived realism [52] minus the place illusion item included in the work outlined above, given that we considered screen-based virtual character interactions insufficient for inducing the illusion of place. We also included a measure of empathic concern derived from related studies [54,59] to assess whether the photorealism component of the stimuli could induce affective perspective-taking, in addition to cognitive and associative state-based empathy, in participants outside of an immersive virtual environment (See Table 1, Empathy). Since the experimental design incorporated three levels of voice realism, we also used measures from Cabral et al. [27] designed to acquire data on subjective impressions of synthetic voices in virtual character interactions (See Table 1, Voice Traits).

4.4. Participants and procedure

Participants were recruited using Prolific,[7] where we were able to attain an even split of participants on gender. The participants were provided with an information sheet, and consent form, a demographic survey and instructions before the experiment began. We provided a scenario of context for the virtual character interactions in order to mimic the virtual environment in the study outlined above. The mixed factorial design that was employed meant that each participant was only exposed to one voice condition for three different scenarios. Each participant viewed the three scenario videos twice, for two blocks of questions. The first block consisted of our emotional response and empathy measures. The second block contained our items for social presence, realism and affinity ratings. The order of the videos was randomised between blocks, with three attention check questions per participant.

4.5. Analysis

Analysis was conducted on data acquired from the 60 participants recruited on Prolific, with a further 8 participants rejected on the grounds of failed attention checks. The 60 included responses were analysed with a mixed factorial ANOVA on between subjects factor of 'Voice' and within subjects factor of 'Scenario', using IBM's statistical analysis software SPSS. Since several of our between-subjects measures violated homogeneity of variance on a Levene's test, we conducted a non-parametric Games–Howell test on the Bonferonni corrected data adjusted for multiple comparisons, to establish whether the effects of between subjects factor Voice were influenced by within subjects factor Scenario.

4.6. Results

We aimed to assess the influence of three levels of voice realism (Real, Synthetic, TTS) as our between subjects factor, with Scenario (Friendly, Unfriendly, Sad) as our within subjects factor. For a full account of the significant effects for this study, see Table 2. All unreported scales or effects were not significant.

H1: Voice type and social presence

As before, the 5 measured items of the Social Presence scale were tested for reliability for each Scenario, and due to sufficient correlation (Cronbach's alpha $\alpha = 0.82$ for Friendly, $\alpha = 0.80$ for Sad, $\alpha = 0.78$ for Unfriendly), we used a cumulative score of all 5 items (where items 3 and 5 were reversed) as the final result and treated it as a continuous scale.

A main effect for Voice was observed on the combined Social Presence variables for this experiment (See Table 2), in a deviation from the results of the first experiment. Particularly, the Text-To-Speech voice condition was measured as the lowest scoring on total means of our Social Presence metrics ($\bar{x} = 12.65, SD = 1.344$), see Fig. 5 implicating a lower level of voice realism evident in the TTS condition as a factor in subjective perceptions of social presence.

H2: Voice type, traits and emotions

A significant main effect of Voice type was detected on variable Excited (see Fig. 6) for all scenarios with the lowest scores recorded for the TTS voice condition for all scenarios ($\bar{x} = 1.5, SD = 0.171$) and the highest overall for the Real Voice ($\bar{x} = 2.4, SD = 0.171$). Observed interaction effects indicated that excitement for TTS condition was lowest for the Unfriendly Scenario ($\bar{x} = 1.20, SD = 0.250$) and the highest for the Synthetic Voice in the Friendly Scenario ($\bar{x} = 3.25, SD = 0.313$). There were no other effects of Voice observed in the data.

Scenario was observed to significantly effect all four emotional response variables; Concern, Excited, Afraid, and Calm, in stark contrast to Scenario results for the first experiment, but also

[7] https://www.prolific.co/

Table 2
Summary of all main effects with corresponding post-hoc analysis for Experiment 2.

Realism		
Effect	Analysis of variance	Post-hoc
MOVEMENT: Scenario	$F(2, 54) = 6.29, p = 0.003$	Sad was rated to have the most realistic movement compared to Unfriendly ($p = 0.008$)
BEHAVIOUR: Scenario	$F(2, 54) = 6.86, p = 0.002$	Behavioural realism was measured to be significantly higher in Sad compared to Unfriendly ($p = 0.005$).
BEHAVIOUR: Voice	$F(2, 54) = 3.382, p = 0.041$	Natural Voice more realistic than TTS Voice ($p = 0.04$).
Emotional Response		
CONCERN: Scenario	$F(2, 54) = 26.224, p < 0.001$	Concern for both Sad and Unfriendly measured to be greater than Friendly ($p < 0.001$).
EXCITED: Scenario	$F(2, 54) = 32.25, p < 0.001$	Friendly had higher ratings of excitement than both Sad and Unfriendly ($p < 0.001$).
EXCITED: Voice	$F(2, 54) = 8.19, p < 0.001$	Both Natural Voice ($p = 0.002$) and Synthetic Voice ($p = 0.003$) considered more exciting compared to TTS Voice .
AFRAID: Scenario	$F(2, 54) = 9.785, p < 0.001$	Afraid variable was reported greater in both Unfriendly and Sad compared to Friendly ($p < 0.001$).
CALM: Scenario	$F(2, 54) = 13.266, p < 0.001$	Friendly received more ratings on Calm than both Sad and Unfriendly ($p < 0.001$).
Voice Traits		
LIKEABILITY: Voice	$F(2, 54) = 12.642,\ p < 0.001$	Real voice was observed to be significantly more Likeable than either Synthetic or TTS voices($p < 0.001$) .
LIKEABILITY: Scenario	$F(2, 54) = 5.749, p = 0.004$	Unfriendly scenario was considered the least likeable compared to Friendly ($p < 0.007$) .
CONSISTENCY: Voice	$F(2, 54) = 4.969, p = 0.010$	Consistency ratings were lowest for Synthetic Voice compared to the Real voice ($p = 0.002$).
EXPRESSIVENESS: Voice	$F(2, 54) = 5.464, p = 0.007$	TTS was recorded as the least expressive compared to Real condition ($p = 0.010$).
Social Presence		
SOCIAL PRESENCE: Voice	$F(2, 54) = 4.58, p = 0.014$	TTS voice prompted the lowest social presence scores, significantly less than Synthetic Voice ($p = 0.017$).
Empathy		
ITEM 1: Scenario	$F(2, 54) = 5.38, p = 0.006$	Character's emotions perceived as most genuine for Sad scenario over Unfriendly ($p = 0.010$).
ITEM 2: Scenario	$F(2, 54) = 11.034, p < 0.001$	Sad scenario most induced participants to feel the character's emotions, in contrast to the Unfriendly condition ($p < 0.001$) and the Friendly condition ($p = 0.08$).
ITEM 5: Scenario	$F(2, 54) = 7.663, p < 0.001$	Subjects reported that they were most absorbed for Sad condition compared to bottom ranked Unfriendly scenario ($p = 0.014$) as well as Friendly scenario ($p = 0.011$).
Affinity		
APPEAL: Scenario	$F(2, 54) = 14.965, p < 0.001$	Friendly scenario rated significantly more appealing than Unfriendly ($p < 0.001$).
APPEAL: Voice	$F(2, 54) = 3.309, p = 0.044$	TTS condition was given significantly lower rating for appeal than Real Voice ($p = 0.043$).
FAMILIAR: Scenario	$F(2, 54) = 4.544, p = 0.013$	Friendly scenario was more familiar when compared to Unfriendly scenario ($p = 0.029$).
EERIE: Scenario	$F(2, 54) = 8.679, p =< 0.001$	Scores on Eeriness were strongest for the Unfriendly scenario compared to both Friendly and Sad ($p < 0.001$)

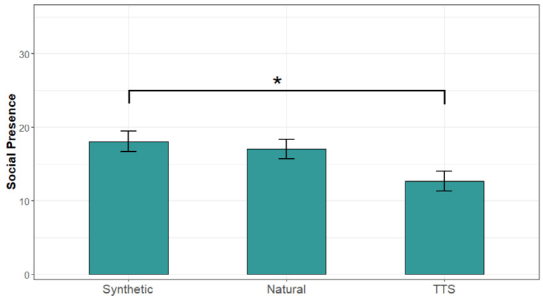

Fig. 5. Main effect of Voice on Social Presence in Experiment 2. Lines above bars denote significant differences, $* = p < 0.05$.

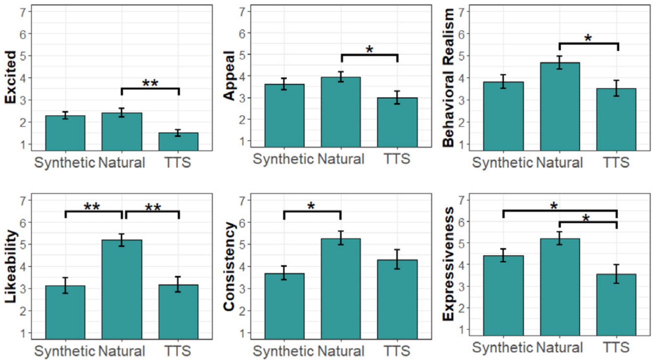

Fig. 6. Main effect of Voice on various scales in Experiment 2. Lines above bars denote significant differences, $* = p < 0.05$, $** = p < 0.001$.

predictably given the expressed valence of the respective scenarios. For instance, the Concern variable scored highest in the Sad scenario ($\bar{x} = 5.13, SD = 0.230$), while Afraid variable was

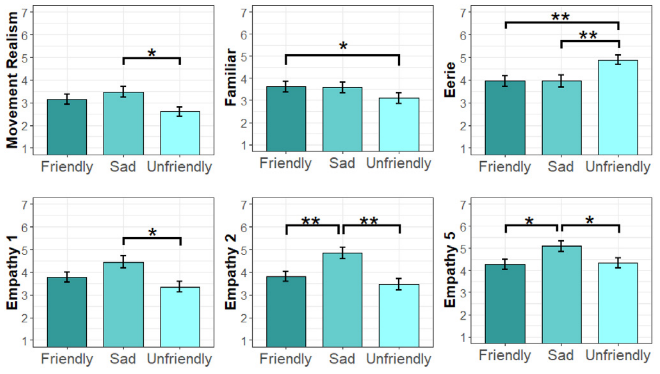

Fig. 7. Main effect of Scenario on various scales in Experiment 2. Lines above bars denote significant differences, $* = p < 0.05$, $** = p < 0.001$.

most prominently rated in the Unfriendly scenario ($\bar{x} = 4.0, SD = 0.251$). Similarly, the Friendly scenario evoked stronger results on both Excited ($\bar{x} = 2.93, SD = 1.81$) and Calm ($\bar{x} = 3.68, SD = 0.241$).

Conversely, the lowest observed means for Concerned ($\bar{x} = 3.15, SD = 0.247$) and Afraid ($\bar{x} = 2.65, SD = 0.230$) variables were recorded in the Friendly scenario, while Calm ($\bar{x} = 2.167, SD = 0.164$) was lowest for Unfriendly and Excited ($\bar{x} = 1.53, SD = 0.107$) was reported as the lowermost for the Sad scenario.

H3: Voice type and discomfort

Voice type was not observed to influence variables Eeriness or Familiarity, although a main effect was detected on Appeal ratings (see Fig. 6), with TTS voice condition evoking the least appeal ($\bar{x} = 2.983, SD = 0.270$) and Real Voice being the most appealing ($\bar{x} = 3.95, SD = 0.270$).

The virtual character was considered most appealing for the Real Voice condition in the Friendly Scenario ($\bar{x} = 4.550, SD = 0.400$) and least appealing for the TTS Voice in the Unfriendly Scenario ($\bar{x} = 2.1, SD = 0.329$).

Other results

Appeal variable was also significantly impacted by Scenario, with the character deemed to be least appealing in the Unfriendly Scenario overall ($\bar{x} = 2.750, SD = 0.190$). Similarly, an effect of Scenario was detected on both Eeriness and Familiarity, see Fig. 7. Eeriness variable was scored lowest overall for the Unfriendly Scenario ($\bar{x} = 4.883, SD = 0.205$) which mirrored the results for Familiarity ($\bar{x} = 3.1, SD = 0.227$).

Surprisingly, there were no effects observed of Voice type on our Empathy variable. However, Scenario did significantly effect empathy levels, particularly Item 1, Item 2 and Item 5, see Fig. 7. The Empathy variable was reported strongest on these measurements for the Sad condition, at its highest rating for this latter empathy measure ($\bar{x} = 5.317, SD = 0.231$).

For our Realism variables (See Table 1) we observed a main effect of Scenario on Movement Realism at its strongest for the Sad Scenario ($\bar{x} = 3.467, SD = 0.219$) and Behaviour Realism also particularly high for Sad ($\bar{x} = 4.467, SD = 0.241$), with Voice only influencing Behaviour Realism (Fig. 6). The highest recorded Behaviour Realism mean was from the Natural Voice condition ($\bar{x} = 4.667, SD = 0.324$). There was no observed effect of Voice on Appearance or Overall Realism in the data analysis. Average means for Overall Realism ($\bar{x} = 3.733, SD = 0.400$) and Appearance Realism ($\bar{x} = 4.155, SD = 0.399$) across all Voice and Scenario Conditions indicate a general absence of observable effect.

Finally, for our subjective Voice Trait variables, we observed a main effect of Voice type on the first three measures Likeability, Consistency and Expressiveness (see Fig. 6). Interestingly, the Synthetic voice was reported to be less likeable ($\bar{x} = 3.117, SD = 0.332$) and less consistent with appearance ($\bar{x} = 3.683, SD = 0.358$) than either the TTS or Real Voice. However, TTS Voice was considered to be less expressive overall ($\bar{x} = 3.550, SD = 0.353$). There was also a significant effect observed for Scenario on Likeability, with the character in the Unfriendly Scenario considered to have the least likeable voice ($\bar{x} = 3.417, SD = 0.225$).

5. General discussion

These two experiments, their resulting datasets and the analyses in tandem provide some useful insight into the experience of social presence and emotional responses during interactions with photorealistic virtual humans. It was expected that the Synthetic Voice in Experiment 1 would be sufficient to reduce reported Social Presence (H1), but this effect was not observed in the VR study. However, with the addition of a lower level of Voice realism, with the TTS Voice, there was a significant result of Voice on our Social Presence variables. It appears that this variable was adversely influenced by the TTS Voice, which allows the possibility of future confirmation of H1 if similar conditions of voice and appearance can be achieved in VR. Regardless, it is some evidence towards the hypothesis that realistic voices are necessary for generating subjective senses of social presence, in non-interactive contexts.

The results for the effect of Scenario changed across both of the studies described above, with emotional response measurements following different paths. The first study found the Sad scenario to be the lowest scoring on our Calm variable, with Unfriendly scenario as the least calm for the second experiment. The second experiment also saw novel findings on Afraid, Excited and Concern, with the latter being most strongly expressed for the Sad scenario. This feeling of concern was evident in the data from our Empathy variables, with empathic responses recorded highest for the Sad condition, despite an effect of Voice being absent from the data. However, the second experiment measured the TTS Voice condition as the lowest for the Excited variable, which adds somewhat to evidence for H2, although it remains only partial confirmation. There was a distinct difference between results of the effect of Scenario on emotional response metrics between the two conducted experiments, with participants responding more noticeably to the character's expressions of friendliness, unfriendliness and sadness. Given the higher quality face detail and animation for the second study, it is possible that important nonverbal cues of emotion were preserved and inducted by subjects. However, it is also conceivable that the absence of distractions entailed by a VR experience, combined with the higher intensity of a face only emotion expression, lead to a more accessible interpretation of emotion information. Furthermore, the singular presence of the virtual character's face presented on screen in our second experiment positioned it over a greater proportion of participants' visual fields. This may have increased the perceptual salience of the expressed emotions, in a further potential explanation of these results.

In terms of our voice mismatching methods, the TTS Voice was considered to be the least expressive voice condition, as well as the least appealing. While future work might be required to explore whether voice expressiveness impacts Social Presence with virtual characters, our finding that the photorealistic character with the TTS voice was the least appealing can be considered some evidence for our hypothesis on the effect of voice on uncanny ratings (H3), as well as a contribution to the growing number of perceptual mismatch theories which seek to explain the proposed Uncanny Valley effect. Additionally, the finding that the Real Voice condition was rated the most appealing could indicate that the photorealism of the virtual character was sufficiently perceptually aligned with a natural human voice. Our results mirror previous findings on preference for natural voices over synthetic voices, as described by Cabral et al. [60]. However, there was a noticeable novelty to our voice consistency results, as the synthetic voice was considered to be much less consistent with character appearance than the natural voice.

The Voice Trait variables introduced in the latter study could shed some light on the first person experience of the Uncanny Valley effect. While the TTS condition was the least appealing on measures for uncanniness, the Synthetic voice was considered the least likeable and also least consistent with the appearance of the character, despite being more expressive than the TTS voice while there were no significant effects on Voice Understanding. The confusion may perhaps be shown in future work to arise at a level of habitability [24], wherein vocal agents are proposed to be most successful at speech when aligned with a users perceptually conforming interactional affordances. In other words, as applied to our study, with Synthetic Voice resembling the Real Voice much more closely than the TTS Voice, it was nonetheless belonging to a virtual human, albeit a photorealisic one. It nevertheless seems unclear to active perceivers what kind of interaction is occurring for the near natural voice, compared to the human or clearly non-human voices, potentially explaining our results for Voice Likeability and Voice Consistency, given that the TTS voice was specifically targeted at increasing inconsistency between voice and appearance realism in this study.

Our finding that the Real Voice condition was significantly both more appealing and more consistent with the appearance of the character is supported by previous work which has used human versus computer generated humanoid experiments to study voice and appearance combinations, and uncovered an overarching preference for consistency [36]. However, this is not mirrored in the lower voice realism condition, for the reasons described above. Our work found that our Synthetic voice was rated least likeable and least consistent with character appearance in our second experiment. The preservation of speech rate in the synthetic voice for Experiment 2, compared to the morphed speech rate from Experiment 1 may have had an impact on these results for our consistency measure, and would be a useful further direction for research. Indeed, such work could be of use to both studies examining uncanny virtual humans of habitable vocal agents, as well as the combination of their respective traits in interactive contexts. Future work in this area could elucidate the extent of an influence of speech rate. The results for the first study are consistent with previous findings which have not determined differences in social presence or uncanny valley measurements with photorealistic characters in VR [57]. While there were observed effects on these measures and other measurements that we adopted from Zibrek et al. [57] in the results of the second experiment, this was unlikely to be simply a consequence of increased appearance realism alone. The data suggest the heightened photorealism and mismatching with lower quality voice was the primary factor which dictated these differences, especially given the general lack of interaction effects in the results of both studies.

6. Limitations and future work

To this end, it is necessary to caveat that the work described above has not achieved clear cut and indisputable results dictating that highly mismatched appearance and voice realism influence social presence or induce the uncanny valley effect. Further studies should incorporate differing levels of appearance realism employed in conjunction with at least the three levels of voice realism used in our second study, in order to achieve further clarity in this domain.

Future work should address different levels of voice realism in dialogue with a virtual character, as opposed to just constrained video-question formats. It would also be a useful direction for addressing the domain of empathy for virtual characters in greater detail. The main limitations on this study are the constraints placed upon the interactions with the characters. Real-time lip sync and gesture generation animations in tandem with high functioning dialogue systems provide an interesting future avenue for this kind of perceptual mismatch research.

In human–avatar interaction research, perceptually salient modulations of voice, such as the ones described in this research, have previously been proposed to influence ratings of attractiveness and immersion and task performance [15] for gaming environments. Emotionally valenced game-based scenarios between human controlled avatars and agents could be a fruitful future direction for research.

Another limitation to this study is an absence of proximity measures for conditions of higher photorealism and TTS voice. Research in this field could help to clarify our results if there are differences in preferred closeness to more photorealistic or less vocally realistic virtual humans. Future studies may also employ highly stylised versions of the presented character, in order to better understand the relationship between voice an appearance.

A further limitation stems from the difference in screen-based and immersive stimuli between the first and second experiment. Future studies should seek to replicate this kind of work using TTS voices in 3D VR contexts. This work would expand upon previous findings in experiments that have analysed perceptions of feedback from virtual agents in 3D VR. Such work has suggested that auditory feedback is preferable above other forms, namely behavioural and visual feedback, when partnered with virtual agents in a task-based environment [41], which could serve to elucidate to the effect of speech feedback from VR agents with both natural, morphed or clearly synthetic voices, in both task-based and emotionally valenced scenarios such as ours.

Similarly, experimental work might seek to employ lower levels of realism when aligned with multiple levels of voice realism. This may help to process perceptual mismatch data for the purpose of visualisation in terms of the uncanny valley.

7. Conclusion

This paper investigated the effect of synthetic voice on agents with photorealistic appearance, which are becoming more commonplace in the industry (e.g., Soul Machines,[8] etc.). Our main contribution is our new insights which show that synthetic voices are considered unappealing when displayed with a photorealistic virtual human – a natural voice is a better choice for these types of characters, where possible. This may have implications for developers of agent systems that aim for photorealism, since text-to-speech voices have not yet reached the level of realism of virtual faces, and the mismatch reduces likeability. However, we also found that synthetic voices do not affect empathy or concern levels towards the character. This implies that synthetic voices can be used with photorealistic characters for applications requiring an empathetic response, such as in therapy or doctor training, where photorealism is considered a more appropriate choice than a cartoon style [32].

We also found that characters with more expressive synthetic voices can have equal social presence and appeal as characters with natural voices implying that higher levels of expressiveness should be an aim for future synthetic voice synthesis. However, it may also be argued that for ethical purposes, an obviously synthetic voice could be advantageous in gaining agent rapport,

[8] https://www.soulmachines.com/

to retain trust with the user who may in the future find it difficult to determine if a virtual human is real, based on its appearance.

Our work is one of the first to investigate photorealism with a character with impressively high levels of visual fidelity. We hope our work will lead to future investigation of this important topic, since the advancement in realism is much more rapid in virtual characters than the number of perceptual studies being conducted to increase our understanding of interactions with these digital creations.

CRediT authorship contribution statement

Darragh Higgins: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Katja Zibrek:** Conceptualization, Methodology, Investigation, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Joao Cabral:** Resources, Software, Writing – original draft, Writing – review & editing. **Donal Egan:** Resources, Software. **Rachel McDonnell:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and under the ADAPT Centre for Digital Content Technology (Grant No. 13/RC/2106_P2) and RADICal (Grant No. 19/FFP/6409).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2022.03.009.

References

[1] Mitchell WJ, Szerszen Sr KA, Lu AS, Schermerhorn PW, Scheutz M, MacDorman KF. A mismatch in the human realism of face and voice produces an uncanny valley. I-Perception 2011;2(1):10–2.

[2] Mori M. The uncanny valley. Energy 1970;7(4):33–5.

[3] Biocca F, Harms C, Gregg J. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In: 4th Annual international workshop on presence. 2001, p. 1–9.

[4] Zibrek K, Cabral J, McDonnell R. Does synthetic voice alter social response to a photorealistic character in virtual reality? In: Proceedings of motion interaction and games. 2021, [in press].

[5] Macdorman KF, Green RD, Ho CC, Koch CT. Too real for comfort? Uncanny responses to computer generated faces. 2008, URL: http://www.macdorman.com.

[6] Biocca F, Burgoon JK, Stoner GM. Criteria and scope conditions for a theory and measure of social presence SEE PROFILE. 2001, URL: https://www.researchgate.net/publication/239665882.

[7] Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc B 2009;364(1535):3549–57.

[8] Bailenson JN, Blascovich J, Beall AC, Loomis JM. Interpersonal distance in immersive virtual environments. Pers Soc Psychol Bull 2003;29(7):819–33.

[9] Zibrek K, McDonnell R. Social presence and place illusion are affected by photorealism in embodied VR. In: Motion, interaction and games. 2019, p. 1–7.

[10] Zibrek K, Kokkinara E, McDonnell R. 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; 2017, p. 3.

[11] Zibrek K, Kokkinara E, McDonnell R. The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? IEEE Trans Vis Comput Graphics 2018;24(4):1681–90.

[12] Sallnäs EL. Haptic feedback increases perceived social presence. In: International conference on human haptic sensing and touch enabled computer applications. Springer; 2010, p. 178–85.

[13] Skalski P, Whitbred R. Image versus sound: A comparison of formal feature effects on presence and video game enjoyment. PsychNology J 2010;8(1).

[14] Oh CS, Bailenson JN, Welch GF. A systematic review of social presence: Definition, antecedents, and implications. Front Robot AI 2018;5:114. http://dx.doi.org/10.3389/Frobt.

[15] Kao D, Ratan R, Mousas C, Magana AJ. The effects of a self-similar avatar voice in educational games. In: Proceedings of the ACM on human-computer interaction. vol. 5, Association for Computing Machinery; 2021, http://dx.doi.org/10.1145/3474665.

[16] Hu P, Wang K, Liu J. Speaking and listening: Mismatched human-like conversation qualities undermine social perception and trust in AI-based voice assistants. In: PACIS. 2019.

[17] Lee KM, Nass C. Designing social presence of social actors in human computer interaction. 2003.

[18] Masahiro M. The uncanny valley. Energy 1970;7(4):33–5.

[19] Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc Cogn Affect Neurosci 2012;7(4):413–22.

[20] Seyama J, Nagayama RS. The uncanny valley: Effect of realism on the impression of artificial human faces. Presence Teleoperators Virtual Environ 2007;16(4):337–51.

[21] Zell E, Aliaga C, Jarabo A, Zibrek K, Gutierrez D, McDonnell R, et al. To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. ACM Trans Graph 2015;34(6):1–12.

[22] Maloney D, Freeman G, Wohn DY. "Talking without a voice": Understanding non-verbal communication in social virtual reality. Proc ACM Hum-Comput Interact 2020;4. http://dx.doi.org/10.1145/3415246.

[23] Ferstl Y, Thomas S, Guiard C, Ennis C, McDonnell R. 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. 2021, p. 76–83.

[24] Moore RK. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. 2016, URL: http://arxiv.org/abs/1607.05174.

[25] MacDorman KF, Chattopadhyay D. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 2016;146:190–205. http://dx.doi.org/10.1016/j.cognition.2015.09.019.

[26] Torre I, Dogan FI, Kontogiorgos D. Voice, embodiment, and autonomy as identity affordances. In: HRI 2021-robo-identity: exploring artificial identity and multi-embodiment march 2021. 2021.

[27] Cabral J, Cowan B, Zibrek K, McDonnell R. The influence of synthetic voice on the evaluation of a virtual character. 2017, p. 229–33. http://dx.doi.org/10.21437/Interspeech.2017-325.

[28] Potard B, Aylett M, Braude D. Cross modal evaluation of high quality emotional speech synthesis with the virtual human toolkit. 10011, 2016, p. 190–7. http://dx.doi.org/10.1007/978-3-319-47665-0_17.

[29] Parmar D, Ólafsson S, Utami D, Murali P, Bickmore T. Navigating the combinatorics of virtual agent design space to maximize persuasion. 2020, URL: www.ifaamas.org.

[30] Thézé R, Gadiri MA, Albert L, Provost A, Giraud AL, Mégevand P. Animated virtual characters to explore audio-visual speech in controlled and naturalistic environments. Sci Rep 2020;10. http://dx.doi.org/10.1038/s41598-020-72375-y.

[31] Ciechanowski L, Przegalinska A, Magnuski M, Gloor P. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. Future Gener Comput Syst 2019;92:539–48. http://dx.doi.org/10.1016/j.future.2018.01.055.

[32] Volante M, Babu SV, Chaturvedi H, Newsome N, Ebrahimi E, Roy T, et al. Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. IEEE Trans Vis Comput Graphics 2016;22:1326–35. http://dx.doi.org/10.1109/TVCG.2016.2518158.

[33] Hartmann T, Toz E, Brandon M. Just a game? unjustified virtual violence produces guilt in empathetic players. Media Psychol 2010;13:339–63. http://dx.doi.org/10.1080/15213269.2010.524912.

[34] Higgins D, Fribourg R, Mcdonnell R. Remotely perceived: Investigating the influence of valence on self-perception and social experience for dyadic video-conferencing with personalized avatars. http://dx.doi.org/10.3389/frvir.2021.668499. URL: www.frontiersin.org.

[35] Higgins D, Egan D, Fribourg R, Cowan B, McDonnell R. Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny? In: ACM symposium on applied perception 2021. New York, NY, USA: ACM; 2021, p. 1–5. http://dx.doi.org/10.1145/3474451.3476242, URL: https://dl.acm.org/doi/10.1145/3474451.3476242.

[36] Gong L, Nass C. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference.. Hum Commun Res 2007;33:163–93.

[37] Go E, Sundar SS. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Comput Hum Behav 2019;97:304–16.

[38] Wang S, Lilienfeld SO, Rochat P. The uncanny valley: Existence and explanations. Rev Gen Psychol 2015;19:393–407. http://dx.doi.org/10.1037/gpr0000056.

[39] Urgen BA, Kutas M, Saygin AP. Uncanny valley as a window into predictive processing in the social brain. Neuropsychologia 2018;114:181–5. http://dx.doi.org/10.1016/j.neuropsychologia.2018.04.027.

[40] Devesse A, Dudek A, van Wieringen A, Wouters J. Speech intelligibility of virtual humans. Int J Audiol 2018;57:908–16. http://dx.doi.org/10.1080/14992027.2018.1511922.

[41] Baxter M, Bleakley A, Edwards J, Clark L, Cowan BR, Williamson JR. You, move there!: Investigating the impact of feedback on voice control in virtual environments. In: ACM international conference proceeding series. Association for Computing Machinery; 2021, http://dx.doi.org/10.1145/3469595.3469609.

[42] Meah LF, Moore RK. The uncanny valley: A focus on misaligned cues. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 8755, Springer Verlag; 2014, p. 256–65. http://dx.doi.org/10.1007/978-3-319-11973-1_26.

[43] Sarigul B, Saltik I, Hokelek B, Urgen BA. Does the appearance of an agent affect how we perceive his/her voice? audio-visual predictive processes in human-robot interaction. In: ACM/IEEE international conference on human-robot interaction. IEEE Computer Society; 2020, p. 430–2. http://dx.doi.org/10.1145/3371382.3378302.

[44] Zibrek K, Martin S, McDonnell R. Is photorealism important for perception of expressive virtual humans in virtual reality? ACM Trans Appl Percept 2019;16(3). http://dx.doi.org/10.1145/3349609.

[45] Kawahara H, Morise M, Takahashi T, Nisimura R, Irino T, Banno H. 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. 2008, p. 3933–6.

[46] Kawahara H, Morise M, Takahashi T, Banno H, Nisimura R, Irino T. Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems. 2010, p. 38–41.

[47] Kawahara H, Morise M, Banno H, Skuk VG. Temporally variable multi-aspect N-way morphing based on interference-free speech representations. In: 2013 Asia-Pacific signal and information processing association annual summit and conference. 2013, p. 1–10.

[48] Bailenson JN, Blascovich J, Beall AC, Loomis JM. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence 2001;10(6):583–98.

[49] Hall ET. The hidden dimension. Doubleday & Co; 1966.

[50] Davis MH. Measuring individual differences in empathy: Evidence for a multidimensional approach. J Personal Soc Psychol 1983;44(1):113.

[51] Golan Ofer BCS, Hill J. The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without asperger syndrome. J Autism Dev Disord 2006;36(2):169–83.

[52] McDonnell R, Breidt M, Buelthoff H. Render me real investigating the effect of render style on the perception of animated virtual humans. ACM Trans Graph 2012;31(4):91:1–91:11.

[53] Skarbez R, Neyret S, Brooks FP, Slater M, Whitton MC. A psychophysical experiment regarding components of the plausibility illusion. IEEE Trans Vis Comput Graphics 2017;23(4):1369–78.

[54] Ho JC, Ng R. Perspective-taking of non-player characters in prosocial virtual reality games: Effects on closeness, empathy, and game immersion. Behav Inf Technol 2020. http://dx.doi.org/10.1080/0144929X.2020.1864018.

[55] van Loon A, Bailenson J, Zaki J, Bostick J, Willer R. Virtual reality perspective-taking increases cognitive empathy for specific others. PLoS One 2018;13. http://dx.doi.org/10.1371/journal.pone.0202442.

[56] Bouchard S, Bernier F, Boivin E, Dumoulin S, Laforest M, Guitard T, et al. Empathy toward virtual humans depicting a known or unknown person expressing pain. Cyberpsychology Behav Soc Netw 2013;16(1):61–71.

[57] Zibrek K, Martin S, McDonnell R. Is photorealism important for perception of expressive virtual humans in virtual reality? ACM Trans Appl Percept 2019;16(3):1–19.

[58] Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Personal Soc Psychol 1988;54(6):1063.

[59] Shen L. Mitigating psychological reactance: The role of message-induced empathy in persuasion. Hum Commun Res 2010;36:397–422. http://dx.doi.org/10.1111/j.1468-2958.2010.01381.x.

[60] Cabral JP, Cowan BR, Zibrek K, McDonnell R. The influence of synthetic voice on the evaluation of a virtual character. In: Proc. interspeech 2017. 2017, p. 229–33. http://dx.doi.org/10.21437/Interspeech.2017-325.