A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker

Tiffany D. Do
University of Central Florida
Orlando, FL, USA
tiffanydo@knights.ucf.edu

Ryan P. McMahan
University of Central Florida
Orlando, FL, USA
rpm@ucf.edu

Pamela J. Wisniewski
University of Central Florida
Orlando, FL, USA
pamwis@ucf.edu

ABSTRACT

Virtual humans can be used to deliver persuasive arguments; yet, those with synthetic text-to-speech (TTS) have been perceived less favorably than those with recorded human speech. In this paper, we investigate standard concatenative TTS and more advanced neural TTS. We conducted a 3x2 between-subjects experiment (n=79) to evaluate the effect of a virtual human's speech fidelity at three levels (Standard TTS, Neural TTS, and Human speech) and the listener's gender (male or female) on perceptions and persuasion. We found that the virtual human was perceived as significantly less trustworthy by both genders, if they used neural TTS compared to human speech, while male listeners (but not females) also perceived standard TTS as less trustworthy than human speech. Our findings indicate that neural TTS may not be an effective choice for persuasive virtual humans and that gender of the listener plays a role in how virtual humans are perceived.

CCS CONCEPTS

• **Human-centered computing → User studies**; **User centered design**.

KEYWORDS

virtual humans, social perception, speech fidelity, text-to-speech

ACM Reference Format:
Tiffany D. Do, Ryan P. McMahan, and Pamela J. Wisniewski. 2022. A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491102.3517564

1 INTRODUCTION

Virtual humans are in-scene characters that are human-like in appearances [15]. Virtual humans can be effective as an interface between systems and users as they can facilitate more natural social interactions. For example, users perceived computers as more useful, reliable, engaging, and interactive when a human face was present in the user interface [32]. This ability to enhance social interactions can be particularly effective for persuasion, which has a variety of applications in social computing. For instance, while persuasive virtual agents can advocate for medical counselling [45] and energy conservation [53], they can also deceive users and propagate false information [43]. However, the evaluation of persuasive agents has been a long-standing problem, and recent research indicates that further study on their perception is necessary due to their increasingly widespread use and potential impact on users [17]. As both persuasive virtual agents and synthetic speech become more prevalent in commercial products, it is important to analyze their technologies and impacts on society [52].

Several factors can influence the perception of virtual humans, including appearance [31, 58] and animation [45]. We are particularly interested in whether the *speech fidelity* (i.e., how similar synthetic speech is to natural human speech [33]) of a persuasive virtual human has a role in perception. Specifically, we are interested in *social perception*, which is the cognitive process of perceiving people, including the content, information, and functions they provide [69]. Social perception can be important to a user's experience and is key to building successful relationships between virtual agents and humans [19]. For instance, Edwards et al. [18] argue that perceptions of trust is crucial for conversational agents and plays a role in the adoption of the technology [67].

We use the term "fidelity" to refer to "the objective degree of exactness with which real-world experiences and effects are reproduced by a computing system" [35]. Early studies [39, 60] found that speech fidelity had an effect on social perceptions of virtual agents. Participants perceived persuasive speakers with human voices as more knowledgeable and involved compared to speakers with classic synthetic voices. More-recent studies found that virtual humans with modern text-to-speech (TTS) voices received lower ratings of engagement [15], trustworthiness [11], and credibility [45] compared to virtual humans with recorded human voices. Building upon these studies, we are interested in examining neural TTS. Neural TTS is a relatively recent advancement that mimics human speech more accurately than standard concatenative TTS by utilizing long-short term (LSTM) neural networks that are conditioned on previous utterances [13, 14]. Neural TTS has the potential to improve social perception of virtual humans compared to standard TTS. For example, Microsoft suggested that their Neural TTS voices can make interactions with virtual assistants more natural and engaging [28]. Since neural TTS reproduces human speech with a higher degree of realism than concatenative TTS [13], we

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517564

consider it as a higher fidelity option for speech synthesis. However, it is important to consider whether improved fidelity has an effect on social perceptions. In a 2019 review on the state of speech in human-computing interaction (HCI), Clark et al. [12] noted that is necessary to investigate newer technologies, including neural TTS, to determine the extent to which earlier research can be applied to current interactions.

We also examine the effects of the listener's gender as recent work indicates that it may have an effect on perception of a persuasive agent [30]. Although early work found that there was no interaction between listener gender and speech fidelity [39], it is important to determine if older speech research applies to newer speech technologies [12, 56]. Furthermore, recent research indicates that the gender of a user can influence perceptions such as engagement, acceptance, and trust in technologies, but has been insufficiently studied [48]. Hence, we decided to investigate whether listener gender has any significant interactions with the speech fidelity of persuasive virtual humans. Through this research, we answer the following research questions regarding the speech of virtual humans:

- **RQ1**: How does higher fidelity TTS affect social perceptions of a persuasive virtual human speaker?
- **RQ2**: Does the listener's gender interact with speech fidelity to affect social perceptions of a persuasive virtual human speaker?

In order to investigate these questions, we conducted a 3 x 2 between-subjects experiment. We compared a modern concatenative TTS voice (Standard TTS), a neural version of the same synthetic voice (Neural TTS), and a recorded human voice (Human speech). This paper makes two primary contributions:

(1) We show that higher fidelity TTS does not improve perceptions of a persuasive virtual human, and may even negatively affect perceptions across listener genders. We hypothesize that this may be due to the "uncanny valley" effect [38].
(2) We show that listener gender interacts with speech fidelity, which did not occur with older TTS. Male listeners rated both synthetic speech conditions as significantly less trustworthy than the human speech condition, while female listeners had no significant differences across all speech conditions.

2 RELATED WORK

2.1 Interacting with Persuasive Agents

The ability to change a listener's attitude (i.e., persuasion) can be a powerful effect in human-centered computing. Fogg [20] discussed the uses of persuasive computing and described how persuasive technology can be leveraged across multiple domains. For example, previous research has found that a persuasive virtual human has potential in medical applications [2, 45] and environmental applications [53]. Petty and Cacioppo's model of persuasion [47] has been particularly influential on theories of social influence [65] and has been used in many persuasive studies (c.f., [55, 61, 68]). Petty and Cacioppo described the path to persuasion as the thoughtful consideration of arguments central to the issue. Dubiel et al. [17] note that the study of persuasive agents has been a long-standing problem, and has mostly focused on text rather than speech. However, they argue that more research is necessary on the use of synthetic voices in persuasive agents due to their increasingly widespread use.

We focus on embodied virtual agents due to their potential to enhance social interactions and perceptions and increasingly widespread use [62]. Roubroeks et al. [53] found that a persuasive virtual agent's social agency (i.e., social presence) increases psychological reactance. Prior work found that the visual presence of human-like faces and animated agents elicit stronger social responses from users and guide social perceptions [10, 62]. Social perceptions, such as perceptions of trust and credibility, can be important for persuasive applications in particular. For example, although users can be persuaded by the logic of an argument, they may have unfavorable perceptions towards the message [39] or the speaker [60]. These social perceptions can influence a user's experience. Edwards and Sanoubari [18] note that trust is especially important for persuasive conversational agents, and ultimately plays a role in the adoption or extinction of the technology [67]. Additionally, social perceptions can enhance general experiences. For instance, a more positive social perception of a virtual counsellor's intimacy can enhance patient outcomes [49].

Persuasive virtual humans can be affected by different design factors such as appearance [31, 58, 72], gender [71], and behavior. For instance, multiple studies have indicated that the attractiveness of a virtual human can affect persuasion [31, 58]. Additionally, Guadagno et al. [22] noted that virtual humans with high behavioral realism (i.e., natural animations) were more influential than those with low behavioral realism. Parmar et al. [45] found that too much animation may be distracting from a persuasive message. We build upon these prior works by replicating their robust methodological designs to study social perceptions in the context of persuasive virtual humans. We extend beyond these works by examining the influence of speech fidelity, as discussed in the next section.

2.2 Virtual Humans and Speech Fidelity

Speech quality is an important aspect when designing persuasive virtual humans. For example, a virtual human's speech quality can have implications on social perceptions of the speaker. Studies investigating the effects of speech fidelity on social perceptions have shown mixed results. Early studies [21, 42] indicated that users disliked inconsistencies between face and voice (e.g., users disliked a virtual avatar with a human voice or a human with a synthetic voice). On the other hand, a more recent study by Cabral et al. [9] revealed that participants considered both a synthetic voice and human voice to be consistent with a virtual character. In early studies using early synthetic voice engines, Mullennix et al. [39] and Stern et al. [60] found that speech fidelity did not have a significant effect on persuasion or perceptions of an argument. However, both studies found that speech fidelity had a significant effect on perceptions of the speaker, where speakers with human voices were perceived more favorably (e.g., knowledgeable, truthful, involved) than speakers with synthetic voices. TTS technology has considerably improved in the last few decades since these seminal studies were conducted. Our work builds upon Mullennix et al.'s [39] and Stern et al.'s [60] early work, which compared early concatenative synthetic speech and human speech. A notable difference, however, is that we compare modern concatenative speech technology as well as state-of-the-art in neural TTS.

Recent studies have investigated the perception of virtual humans with modern TTS voice engines. For instance, Craig and Schroeder's [15] suggested that the negative effect of synthetic speech in educational virtual humans may need to be re-evaluated due to improvements in software. Chiou et al. However, [11] found that even modern concatenative TTS negatively affected trust and other perceptions of an educational virtual human. Similarly, Parmar et al. [45] investigated the role of a persuasive virtual human's speech using a modern TTS voice engine. Although they found no difference in attitude change (i.e., persuasion), they found that there was a significant main effect of voice for social perceptions such as trust, general satisfaction, and ratings of credibility. Much like these studies, we also compared a modern synthetic voice to a human voice. However, unlike these studies, we also evaluated neural synthetic TTS, in addition to standard concatenative TTS.

There are several reasons why virtual humans with synthetic speech may be perceived less positively than those with human speech. For example, Ardnt et al. [5] found that low-quality TTS causes higher mental load compared to high-quality TTS and suggested that synthetic speech quality may affect listening comprehension. Additionally, although current commercial TTS models (e.g., Azure, Cerence) can emulate speaking styles (i.e., neutral, newscaster), they are unable to synthesize high-fidelity emotion or prosody [8], which can cause participants to feel uneasy. For example, Abdulrahman et al. [1] found that a virtual human with TTS was perceived as more eerie than one with human speech. Furthermore, there may be influences of popular culture, such as films and television shows. Humphry and Chesher [29] argue that the portrayal of characters with "robotic" voices in popular culture may have caused people to perceive artificial voices as "dangerous" or "menacing".

2.3 Gender Effects of the Listener and Agent

While early research found that both the gender of a persuasive agent and the listener significantly affected perceptions [39, 68], recent research argues that gender stereotypes of virtual agents and synthetic voices are no longer as effective as previously assumed [41, 50]. Furthermore, more recent research indicates that while a persuasive virtual agent's gender has no significant effects on persuasion or perception, the gender of the listener affects perceptions of the persuasive agent [30]. Thus, we were motivated to investigate listener gender to determine if it interacts with higher fidelity speech technologies. Although Mullennix et al. [39] did not find any interaction effects between speech fidelity (human or synthetic) and listener gender, they investigated speech engines developed in the 1990s, and early results may not apply to current interactions with improved speech technology [52].

2.4 Neural Text to Speech

Neural TTS has recently gained attention as a method of creating realistic sounding synthetic voices in contrast with standard concatenative synthetic voices. Coto-Jimenez and Goddard-Close [14] noted that using deep neural networks as a postfiltering step can create models that more accurately mimic human voices. In 2017, Arik et al. [4] released a breakthrough system that utilized deep neural networks to create real-time neural TTS. Although standard concatenative TTS voices have been commercially available for several decades, Microsoft only released their commercially available neural TTS voices in 2018 [28], while Amazon introduced neural TTS voices in AWS Polly in 2019 [57]. Neural TTS has the potential to improve user experience. For example, Microsoft claimed that their neural TTS voices reduce listening fatigue and sound more realistic than standard synthetic voices. Additionally, they suggested that neural TTS can make interactions with virtual assistants more natural and engaging [28].

The perception of neural TTS has seen recent interest in research. Cohn and Zellou [13] compared neural TTS voices to standard concatenative TTS voices. They found that the TTS voices were perceived as more human-like, natural, and familiar than standard TTS. These results reveal that neural TTS is perceived as closer to human speech than standard TTS and provide compelling reasons to further investigate the perception of neural TTS. We are particularly interested in the use of neural TTS in virtual humans to determine if it can improve social perceptions of virtual humans with synthetic speech. We build upon this work by using the same TTS voices that Cohn and Zellou investigated within the context of persuasive virtual humans. Since previous work has shown that higher fidelity speech causes a virtual human to be perceived more favorably than lower fidelity speech, we were motivated to investigate neural TTS as a higher fidelity TTS option.

3 METHODS

We conducted a 3 x 2 between-subjects experiment to evaluate the effects of a female virtual human's speech fidelity at three levels (Standard TTS, Neural TTS, and Human speech) and the human listener's gender at two levels (male or female) on persuasion and perception of the speaker, message, argument, and voice. We had one participant who self-identified as non-binary; therefore, we we did not have enough power to include this person in our statistical analyses regarding gender differences but followed best practices of presenting disaggregate data by gender (i.e., descriptive statistics) [63] and to be inclusive of all genders [59]. We intended to replicate the methods of previous experiments that focused on persuasion [39, 60, 68] by using the same source material for our stimulus argument, questionnaires to measure persuasion and social perception ratings, and methodologies for analysis. Our study was conducted between-subjects since we are investigating persuasive agents using the same argument. Each participant listens to the argument once and then provides opinions on their attitude towards the topic and the speaker.

3.1 Research Hypotheses

We proposed the following hypotheses regarding the speech of a female virtual human:

- **H1 (Social Perceptions of the Speaker):** The Neural TTS condition will have more favorable ratings of the speaker (i.e., competence, trustworthiness, and assertiveness) than the Standard TTS condition, considering prior results indicating that Neural TTS voices are perceived as more human-like, natural, and familiar [13].

- **H2 (Listener gender):** Male listeners will report less favorable ratings of the speaker, considering prior results indicating that men may perceive persuasive virtual humans more negatively [30] than women. Gender of the listener will not interact with speech fidelity to affect social perception measures, considering early results indicating that listener gender does not interact with speech fidelity [39].

3.2 Dependent Variables

3.2.1 Degree of Persuasion. We measured the degree of persuasion (i.e. attitude change) using pre- and post-tests, which assess attitudes on four topics: tuition raises, senior comprehensive exams, animal rights, and the environment. The target topic was senior comprehensive exams, while the other three were distractor topics. The pre- and post-tests consisted of 12 statements that measured attitude on these four topics using a 7-point Likert scale (Strongly Disagree to Strongly Agree). The mean score of each topic was combined for each participant to measure their overall attitude toward each topic. These statements can be found in Appendix A.

3.2.2 Social Perceptions. We also replicated the social perception measures of prior studies investigating persuasive speakers [39, 60, 68]. We measured social perceptions of the speaker, message, voice, and argument using numerical scale questionnaires. These questionnaires consisted of numerical scale items that placed favorable perceptions (e.g., competent, honest, intelligent) on one anchor and unfavorable perceptions (e.g., incompetent, dishonest, unintelligent) on the opposite anchor. Unlike the other perception questionnaires, the questionnaire for perception of the voice did not have favorable and unfavorable anchors and is meant to provide insights on noticeable auditory differences between voices that can affect perception [39, 60]. The anchors for all social perception questionnaires are displayed in Appendix B.

3.3 Stimulus Materials

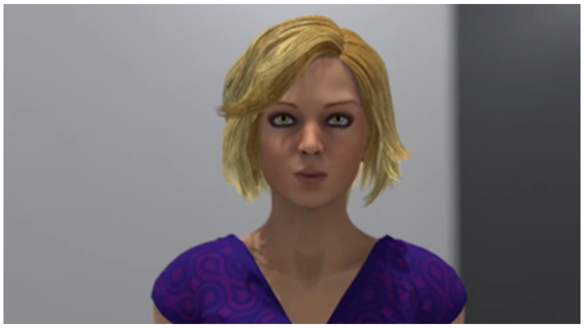

Figure 1: An image of the female virtual human speaker used in our study of speech fidelity and listener gender.

Participants listened to a persuasive message, sourced from Petty and Cacioppo [46], that was designed to change their attitude in favor of university-wide senior comprehensive exams. The message was approximately 2 minutes and 20 seconds long for all conditions. For our speech fidelity levels, we used a modern concatenative TTS voice (Standard TTS), a neural version of the same TTS voice (Neural TTS), and a recorded human voice (Human). Since our participants were recruited from universities within the United States, our synthetic voices were the Joanna voices from Amazon Web Services (AWS) Polly [3], which use a standard American accent. We chose this voice because the majority of virtual assistants use female voices that are localized to the region [29]. Additionally AWS Polly is one of the most well-known TTS services for developers [51] and has seen widespread use in recent research (e.g., [13, 16, 70]). Samples of all voices used can be found at: youtu.be/m6RSxaCYmpk.

The recorded human voice had a standard American accent, which was recorded using a high-definition condenser microphone. Our recorded human speech had faint white noise due to non-studio recording conditions. We added the same white noise generated by the microphone to both synthetic speech recordings to avoid the potential confound. Participants were instructed to use headphones for the entirety of the experiment. The speaker was a female virtual human that was animated using Holotech's FaceRig software [27]. We used a female virtual human due to our investigation of female speech. Our study specifically focuses on female speech because the majority of virtual assistants use female personas [34] We used the model "Jane", which is a pre-set avatar modelled after an average Caucasian human female. To ensure behavioural realism, we used the default settings for idle movement, which included animations, such as blinking and breathing, that were designed to mimic natural idle movement. All three message conditions (Standard TTS, Neural TTS, and Human) were lip-synced using FaceRig's audio based lip-sync, which tracks phenomes from sound input and translates them into atomic animations [26]. Our model and environment can be seen in Figure 1.

3.4 Procedure

The following procedure was reviewed and approved by our university Institutional Review Board (IRB). The study consisted of one online Qualtrics survey that lasted approximately 15 minutes. Each participant completed a background survey that captured their demographics, gender, education, and technology experience. Afterwards, they completed the pre-test that measured their attitude on four topics. Participants were then randomly assigned to one of three conditions of speech fidelity and watched a video of the virtual speaker that delivered the persuasive message. Condition assignment was controlled by gender such that the different genders were evenly distributed across conditions. Finally, participants completed the post-test and questionnaires regarding their perception of the speaker, argument, message, and voice.

3.5 Participants

Participants were recruited through listservs from universities in the United States. An a priori power analysis based on the work by Mullenix et al. [39] indicated that at least 52 participants would be required to detect differences between synthetic female and human female speech. Due to the online nature of our procedure, our survey was completed by more male participants early on than female participants, but we kept recruitment open until we were able to balance male and female participants for all three conditions.

As a result, a total of 79 participants (39 male, 39 female, 1 non-binary) were recruited to take part in the study. Due to the low number of non-binary participants, we were unable include their data in our statistical analysis of gender differences (H2). However, we provide descriptive statistics for comparison purposes. Overall, participants had a mean age of 27.68 , within a range from 19 to 68. All participants reported no visual, audio, or neurological disabilities. Additionally, all participants reported proficiency in the English language.

3.6 Data Analysis Approach

To examine persuasion, we computed the mean score across participants for each message topic (animal rights, environment, tuition, and exams). We then computed the difference between the post-test and the pre-test and used a two-way ANOVA to test for interaction and main effects of speech fidelity and the listener's gender. We analyzed social perceptions similarly to studies that used the same perception questionnaires [44, 60, 68]. Like these studies, we conducted principal component analysis (PCA) [66] on item ratings of perceptions of the speaker, message, argument, and voice. We combined items into factor scores for each respective social perception measure, which were produced by averaging the item weights. We then retained items that had factor loadings above 0.5 after varimax rotation [68] and were at least 0.2 greater than any of their cross-loadings [25]. All factors had satisfactory composite reliability (CR > 0.7) and average variance extracted (AVE > 0.5) [24]. Our factors and their CR, AVE, and percent of variance explained values can be found in Appendix C.

Perception of the speaker had three factors, which were *Competent* (combines competent, informed, qualified, and intelligent), *Trustworthy* (combines trustworthy, sincere, honest, and active), and *Assertive* (combines assertive, bold, and forceful). Perception of the message had two factors, which were *Interesting* (combines stimulating and interesting) and *Supported* (combines supported and specific). Perception of the argument had only one factor, *Effective*, which all items were loaded onto. Finally, perception of the voice had three factors, which were *Nondistinctive* (combines faint accent and faint nasality), *Lively* (combines lively and didn't talk enough), and *Unconfident* (combines soft-spoken and slow speaking).

4 RESULTS

4.1 Persuasion

We first examined the persuasiveness of the virtual human to ensure that our findings replicate those of past studies (i.e., ensure that we are examining the perception of a persuasive virtual human). We examined the pre-test/post-test difference of the target topic (Exams) against all control topics. See Table 1 for the mean and standard deviations regarding the pre-test/post-test difference of all topics. The score difference shows the change in attitude from pre-test to post-test, where a significantly positive score indicates an increased agreement with the topic. We found that the virtual human was statistically persuasive across all conditions, suggesting that persuasion occurred.

Across all genders, we found that the Exams topic was significantly more persuasive than the Animal Rights topic, $t(156) = 4.55, p < 0.01$, the Environment topic, $t(156) = 5.12, p < 0.01$, and the Tuition topic, $t(156) = 2.33, p = 0.01$. We also found that the Tuition topic was significantly more persuasive than the Environment topic, $t(156) = 4.38, p < 0.01$. However, previous studies found that the stimulus message also affected attitude on campus related topics, such as tuition [60, 68]. Finally, we did not find a significant interaction effect between speech fidelity and listener's gender on persuasion, $F(2, 72) = 0.42, p = 0.66$. Furthermore, we did not find a significant main effect of speech fidelity, $F(2, 76) = 0.01, p = 0.99$ or gender, $F(1, 76) = 3.08, p = 0.08$ on persuasion.

Table 1: Mean and standard deviation of attitude score differences (persuasion) from pre-test to post-test by topic across all genders.

	Standard **M (SD)**	**Neural** **M (SD)**	**Human** **M (SD)**
Exams	1.96 (3.67)	2.00 (3.42)	1.88 (3.57)
Animal rights	0.04 (2.27)	-0.48 (1.87)	0.23 (1.14)
Environment	-0.27 (1.51)	-0.37 (1.47)	0.00 (1.06)
Tuition	1.12 (1.75)	0.41 (1.99)	1.23 (1.73)

4.2 Effects of Speech Fidelity on Social Perceptions (H1)

We examined mean factor score ratings to analyze social perceptions of the speaker, message, argument, and voice. We conducted a series of one-way ANOVAs to test for main effects of speech fidelity on all factors across all genders. We used a series of Tukey's HSD (Honestly Significant Difference) [54] tests for post hoc analysis. The summaries of our tests are displayed in Table 3. See Table 2 for the means and standard deviations of all factors.

To investigate the effect of speech fidelity on perception of the speaker, we examined the mean factor scores on three dimensions of perception of the speaker: Competent, Trustworthy, and

Table 2: Mean and standard deviation of factor scores for social perceptions of the Speaker, Message, Argument, and Voice across all listener genders.

	Standard ***M (SD)***	**Neural** ***M (SD)***	**Human** ***M (SD)***
Perception of the Speaker			
Competent	4.85 (1.16)	4.62 (0.86)	5.08 (1.11)
Trustworthy	3.95 (1.24)	3.56 (0.94)	4.47 (1.18)
Assertive	4.00 (1.10)	4.32 (1.32)	4.38 (1.13)
Perception of the Message			
Interesting	3.40 (1.39)	3.83 (1.41)	3.75 (1.62)
Supported	5.04 (1.36)	4.98 (1.120	5.31 (1.01)
Perception of the Argument			
Effective	5.95 (1.44)	5.76 (1.48)	5.92 (1.38)
Perception of the Voice			
Nondistinctive	4.60 (1.30)	4.74 (1.20)	5.15 (1.09)
Lively	2.57 (1.02)	2.56 (1.07)	3.12 (1.01)
Unconfident	3.77 (1.30)	3.94 (0.98)	4.17 (0.79)

Table 3: Summary of two-way ANOVA tests for main and interaction effects of speech fidelity and listener's gender. An asterisk denotes a significant difference at p < 0.05.

	Speech Fidelity		Listener's Gender		Speech x Gender	
	F	*p*	F	*p*	F	*p*
Perception of the Speaker						
Competent	1.26	0.29	0.03	0.87	0.93	0.40
Trustworthy	4.32	0.02*	3.53	0.06	3.76	0.03*
Assertive	0.79	0.46	0.02	0.89	3.07	0.05
Perception of the Message						
Interesting	0.62	0.54	0.05	0.82	0.97	0.38
Supported	0.58	0.56	0.28	0.60	0.10	0.98
Perception of the Argument						
Effective	0.13	0.88	0.52	0.47	0.09	0.91
Perception of the Voice						
Nondistinctive	1.52	0.23	0.06	0.82	1.09	0.34
Lively	2.47	0.09	5.36	0.02*	1.50	0.23
Unconfident	0.98	0.38	0.75	0.39	0.38	0.68

Assertive. See Table 2 for the mean and standard deviations of these factor scores. We found a significant main effect of speech fidelity on the *Trustworthy* dimension, $F(2, 76) = 4.32, p = 0.017$. A post hoc (Tukey's HSD) test revealed that Human speech ($M = 4.47, SD = 1.18$) was rated as significantly more *Trustworthy* than Neural TTS ($M = 3.56, SD = 0.94$), $p = 0.01$. There were no significant differences between Human speech and Standard TTS ($M = 3.95, SD = 1.24$), or between Neural TTS and Standard TTS. H1 (Social Perceptions of the Speaker) was not supported because Neural TTS did not have significantly more favorable ratings of the speaker when compared to Standard TTS. We found no significant main effects of speech fidelity (see Table 3) on perceptions of the message, argument, or voice.

4.3 Effects of Listener Gender (H2)

4.3.1 Effects on Social Perceptions. We examined the mean factor scores on three dimensions of perception of the voice: Nondistinctive, Lively, and Unconfident. We found a significant main effect of gender, $F(2, 76) = 5.36, p = 0.02$. A post hoc (Tukey's HSD) test revealed that female listeners ($M = 3.03, SD = 1.08$) rated the virtual human's voice as more *Lively* than the male listeners ($M = 2.49, SD = 0.97$) did as a main effect across all conditions. We did not find a significant main effects of gender on any other social perception measures. For descriptive and comparison purposes, the non-binary participant, who was assigned to the Neural TTS condition, reported the rating of 2.0 for *Lively*.

4.3.2 Interaction Effects of Listener Gender. We found a significant interaction effect between speech fidelity and gender on the Trustworthy dimension (see Table 3), $F(2, 72) = 3.76, p = 0.03$. We show the interaction plot in Figure 2. A one-way ANOVA revealed there was a significant effect of speech fidelity on the *Trustworthy* dimension for male listeners, $F(2, 36) = 7.06, p = 0.003$. A post hoc (Tukey's HSD) test on the male listener data revealed that Human speech was rated as significantly more *Trustworthy* than both the Neural TTS ($p = 0.003$) and the Standard TTS ($p = 0.018$). There was no significant difference between the synthetic speech conditions. In contrast to male listeners, a one-way ANOVA revealed that there was no significant effect of speech fidelity on the Trustworthy dimension for female listeners, $F(2, 36) = 0.725, p = 0.491$. H2 (Listener gender) was partly supported because male listeners rated the speaker more negatively than female listeners did (i.e., less *Lively*, but unlike our predictions, there was a significant interaction effect between listener gender and speech fidelity.

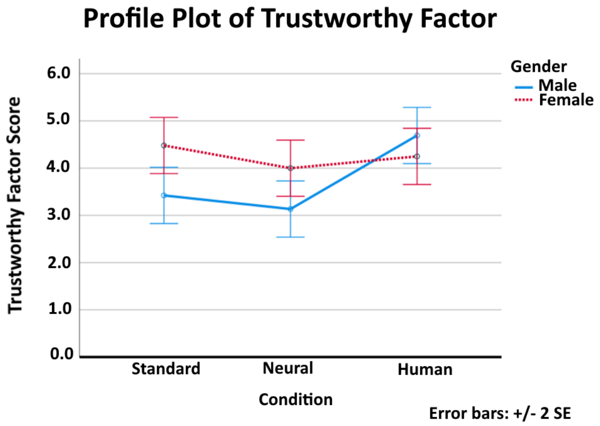

Figure 2: Interaction plot of mean Trustworthy Factor ratings between male and female listeners.

5 DISCUSSION

In this section, we first discuss implications for further research based on our results, and then implications for virtual human design and development. We conclude the section by discussing the limitations of our work and paths for the future.

5.1 Neural TTS No Better than Standard TTS

Our results did not reflect our predictions regarding the effects of speech fidelity. Since we compared a Neural and Standard version of the same TTS voice, we anticipated that the higher fidelity Neural TTS would produce more positive social perception ratings. Interestingly, we found that the virtual human with Human speech was rated as significantly more *Trustworthy* than one with Neural TTS, but there was no significant difference between Human speech and Standard TTS. Furthermore, there were no significant differences between Neural TTS and Standard TTS with respect to any perception ratings. Although the Neural TTS voice used in this study was rated as more human-like and natural than its Standard TTS counterpart [13], it did not produce more positive perceptions. Thus, our results indicate that more realistic TTS technology may not necessarily improve social perceptions of persuasive virtual humans, and may even be less optimal.

These results may be due to a phenomenon similar to the "uncanny valley" effect for aesthetics of robots and agents [38], which

Table 4: Mean and standard deviation of factor scores for social perceptions of the Speaker, Message, Argument, and Voice between male and female listeners.

	Standard		**Neural**		**Human**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
	M (SD)	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***
Perception of the Speaker						
Competent	4.73 (1.34)	4.96 (0.96)	4.54 (0.87)	4.69 (0.91)	5.33 (0.98)	4.82 (1.22)
Trustworthy	3.42 (1.36)	4.48 (0.87)	3.13 (0.75)	4.00 (0.97)	4.69 (1.18)	4.25 (1.19)
Assertive	3.59 (1.25)	4.41 (0.78)	4.33 (1.51)	4.13 (0.99)	4.74 (0.93)	4.03 (1.22)
Perception of the Message						
Interesting	3.15 (1.42)	3.65 (1.37)	3.92 (1.51)	3.85 (1.38)	4.08 (1.58)	3.42 (1.66)
Supported	5.19 (1.27)	4.89 (1.49)	5.04 (1.07)	5.00 (1.22)	5.35 (1.01)	5.27 (1.55)
Perception of the Argument						
Effective	5.76 (0.97)	6.15 (1.82)	5.82 (1.46)	5.87 (1.49)	5.79 (1.38)	6.04 (1.44)
Perception of the Voice						
Nondistinctive	4.77 (1.38)	4.42 (1.24)	4.77 (1.27)	4.69 (1.22)	4.85 (1.05)	5.46 (1.09)
Lively	2.23 (0.97)	2.92 (0.98)	2.12 (0.85)	3.04 (1.13)	3.12 (0.82)	3.12 (1.21)
Unconfident	3.73 (1.54)	3.81 (1.07)	3.96 (1.02)	4.00 (0.98)	3.92 (1.12)	4.42 (0.96)

describes an effect where increasing realism increases appeal only up to a point. At this point, increasing realism leads to unacceptable perceptions until realism essentially matches the real world. This may explain our results because the "uncanny valley" phenomenon can apply to non-visual aspects of a simulated experience [36], such as the auditory aspects of our current study. The Neural TTS condition may have produced experiences similar to human speech, except for key inflections and pauses within the speech synthesis. On the other hand, the Standard TTS condition likely produced less similar experiences, so an unnatural point within the synthesis is less likely to stand out. Since the virtual agent has a human-like appearance, these unnatural points in the synthesis may have caused participants to feel the "eeriness" that is commonly described by the uncanny valley effect.

This effect may also be attributed to mental load. Since mental load is inversely related to TTS quality, the Neural TTS condition may have caused a lower mental load compared to the Standard TTS condition [5]. The decreased mental load of Neural TTS may have allowed users to analyze specific qualities of speech (e.g., key inflections and pauses), thus causing the uncanny valley effect. On the other hand, the higher mental load of Standard TTS may have prohibited quality assessment.

5.2 User Gender May Affect Perceptions of Trust of Virtual Humans with TTS

We anticipated that listener gender would not interact with speech fidelity to affect social perceptions because an early study investigating older synthetic voices did not find an interaction effect [39]. However, we found a significant interaction effect between the listener's gender and the speaker's speech fidelity (see Table 3). In our study, male listeners rated the speaker with the Human speech condition as significantly more *Trustworthy* than both Neural TTS and Standard TTS conditions. On the other hand, we found that female listeners did not significantly vary in judgments of the speaker's *Trustworthiness* across conditions. These results reveal that listener gender may interact with modern synthetic speech, unlike older synthetic speech.

Recent reviews on speech in HCI [52, 56] argue that early speech research may no longer apply to current interactions and may need to be investigated. Notably, our results differed from those found by Mullennix et al. [39], who investigated early concatenative synthetic speech engines that were developed several decades ago. While Mullenix et al. reported that both male and female listeners rated human speech as more trustworthy than synthetic speech, we only found a significant difference for male listeners. Our findings indicate that TTS voices may have improved to the point where both Standard TTS and Neural TTS may now be suitable for female listeners. However, this improvement may not suffice for male listeners, who still gave lower ratings of trust even for higher quality TTS. These findings have important implications for virtual human research. Notably, male listeners may perceive trust of virtual humans with TTS differently than female listeners. Because the gender of the listener plays a role in how a virtual human's speech is perceived, future studies that investigate the perception of virtual human speech should aim for gender balancing of participants.

These findings can possibly be explained by the perception of the voice ratings. Although there were no significant interaction effects between listener gender and speech fidelity, female listeners found the voice of the virtual human to be more *Lively* than the male listeners did. Since male listeners rated the virtual human's voice as less *Lively* overall, they may have scrutinized the *Trustworthiness* of the virtual human more than female listeners and therefore gave lower ratings for the synthetic speech conditions. These results are in line with those of Kantharaju et al.'s, [30] who recently reported that male listeners perceived persuasive virtual agents more negatively (i.e., distant, arrogant, and forceful) than female listeners did, regardless of the virtual agent's gender.

Our results suggest that there may be growing evidence that men perceive persuasive virtual agents more negatively than women do. Furthermore, this negative perception could help explain our novel interaction effects between listener gender and speech fidelity (i.e., male listeners required more realistic speech for the virtual human to be acceptable). However, more work is necessary to determine the extent of these gender differences. Our results may also differ from early studies due to the changing landscape of HCI. Within the last two decades since Mullennix et al. [39] conducted their study, there has been a rapid increase in persuasive computing and virtual assistants, which mainly use female TTS personas [34]. There has been some concern that the gendering of virtual personal assistants may pose a societal harm by promoting harmful stereotypes of women [23, 34]. These interactions may have influenced a male bias against female TTS that was not present in early studies. Finally, we consider gender differences regarding trust as a possible reason for our results. Murphy and Tocher [40] indicate that women are more reliant on building trust through communication in comparison to men. Similarly, Awad et al. [6] argue that women are more network oriented, and react more to websites with humanistic elements in comparison to men. Since the virtual agent has a humanistic appearance and directly persuades the user through verbal communication, women may have been influenced to trust all conditions, regardless of speech quality.

5.3 Implications for Virtual Human Design

Our results have important implications for the design of persuasive virtual humans. Our findings indicate that Neural TTS may not be a favorable choice for a virtual human's speech. With respect to perception ratings, Neural TTS may actually be more unfavorable than Standard TTS, if we consider the Human speech condition as a "gold standard" [17] of quality. Furthermore, there were no significant differences between the Neural TTS and Standard TTS conditions. These results indicate that Standard TTS may be sufficient for applications. In this case, developers would not have to pay four times as much for neural TTS [3, 37] or hire voice actors, which may slow down development or incur more costs.

We also found that listener gender plays an important role in social perception of virtual humans. Our results suggest that developers should keep their target audience in mind when designing the speech of virtual humans. Since male listeners perceived virtual humans with both types of synthetic speech as less *Trustworthy* than one with recorded human speech, it would be favorable to choose recorded human speech over synthetic speech, if the application is expected to have a substantial number of male users. On the other hand, female listeners did not have a significant difference between synthetic speech and recorded human speech. The use of either types of synthetic speech or recorded human speech may be suitable for an application intended for a female audience.

5.4 Limitations and Future Work

It is important to note that our study is limited to our specific domain. We focused on a Caucasian female character and used a pre-written argument. Repeating the experiment with a virtual human of a different gender or ethnicity may yield different results. For example, users usually prefer agents with localized accents [56], and using a non-localized accent may negatively affect perceptions. We investigated a female virtual human as an initial study due to the prevalence of female virtual agents. Although recent research indicates that gender stereotypes of virtual agents and synthetic voices are no longer as effective as previously assumed [41, 50], further work is required to determine if these results hold for male and androgynous virtual humans, especially as the HCI community moves away from using female agents as a default [7]. We also used a standard American accent for all voice conditions. In addition, most of our participants were recruited from American university email lists and may not reflect the general population.

In the future, we plan to further investigate the effects of speech fidelity using a mixed-method study. Our study was limited to Likert scale questions and did not provide the opportunity for open-ended discussion, which may provide further insights on our results. It would also be useful to examine different types of avatar styles (i.e., rendering styles), which may interact with voice [64].

6 CONCLUSION

Our study focused on the effects of a persuasive virtual human's speech fidelity on social perception. We compared a standard concatenative TTS synthetic voice, a neural version of the same synthetic voice, and a recorded human voice using an established methodology employed in prior studies investigating persuasion [39, 60, 68]. While speech fidelity did not play a role in persuasion in previous studies [39, 60], we were interested in social perception due to its role in user experience and general satisfaction.

We were motivated to investigate Neural TTS because prior research indicated that persuasive virtual humans with modern concatenative TTS synthetic voices were perceived as less trustworthy and less credible than virtual humans with human voices [45]. We predicted that Neural TTS would produce more favorable perceptions of the speaker compared to Standard TTS because it more accurately mimics human speech. However, we found that there were no significant differences between the two synthetic speech conditions. Our results indicate that Standard TTS may be a more effective choice than Neural TTS in regards to perception of the speaker's *Trustworthiness*. We also found that asides from *Trustworthiness*, there were little differences between all three speech conditions. Contrary to our predictions, we found that listener gender interacts with speech fidelity to affect social perception, where male listeners were more distrusting of virtual humans with TTS, unlike female listeners. Our results suggest that gender plays an important role in the perception of virtual humans with modern synthetic speech. Finally, we recommend that developers keep the gender of their target audience in mind when designing virtual humans in their applications.

REFERENCES

[1] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2019. A Comparison of Human and Machine-Generated Voice. In *25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 41, 2 pages. https://doi.org/10.1145/3359996.3364754

[2] Sun Joo (Grace) Ahn. 2016. *Using Avatars and Agents to Promote Real-World Health Behavior Changes.* John Wiley & Sons, Ltd, Chapter 12, 171–180. https://doi.org/10.1002/9781118952788.ch12

[3] Amazon Inc. 2021. Amazon Polly Pricing. https://aws.amazon.com/polly/pricing/

[4] Sercan Ö. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. 2017. Deep Voice: Real-time Neural Text-to-Speech. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 195–204. https://proceedings.mlr.press/v70/arik17a.html

[5] Sebastian Arndt, Jan-Niklas Antons, Rishabh Gupta, Khalil ur Rehman Laghari, Robert Schleicher, Sebastian Möller, and Tiago H. Falk. 2013. Subjective quality ratings and physiological correlates of synthesized speech. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. 152–157. https://doi.org/10.1109/QoMEX.2013.6603229

[6] Neveen F. Awad and Arik Ragowsky. 2008. Establishing trust in electronic commerce through online word of mouth: An examination across genders. *Journal of Management Information Systems* 24, 4 (2008), 101–121. https://doi.org/10.2753/MIS0742-1222240404

[7] Dania Bilal and Jessica K. Barfield. 2021. Hey There! What Do You Look Like? User Voice Switching and Interface Mirroring in Voice-Enabled Digital Assistants (VDAs). *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 1–12. https://doi.org/10.1002/pra2.431

[8] Sung-Woo Byun and Seok-Pil Lee. 2021. Design of a Multi-Condition Emotional Speech Synthesizer. *Applied Sciences* 11, 3 (Jan 2021), 1144. https://doi.org/10.3390/app11031144

[9] João Paulo Cabral, Benjamin R. Cowan, Katja Zibrek, and Rachel McDonnell. 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In *Proc. Interspeech 2017*. 229–233. https://doi.org/10.21437/Interspeech.2017-325

[10] Veena Chattaraman, Wi Suk Kwon, Juan E. Gilbert, and Yishuang Li. 2014. Virtual shopping agents Persona effects for older users. *Journal of Research in Interactive Marketing* 8, 2 (2014), 144–162. https://doi.org/10.1108/JRIM-08-2013-0054

[11] Erin K. Chiou, Noah L. Schroeder, and Scotty D. Craig. 2020. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers and Education* 146 (2020). https://doi.org/10.1016/j.compedu.2019.103756

[12] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[13] Michelle Cohn and Georgia Zellou. 2020. Perception of Concatenative vs. Neural Text-To-Speech (TTS): Differences in Intelligibility in Noise and Language Attitudes. In *Proc. Interspeech 2020*. 1733–1737. https://doi.org/10.21437/Interspeech.2020-1336

[14] Marvin Coto-Jiménez and John Goddard-Close. 2018. LSTM Deep Neural Networks Postfiltering for Enhancing Synthetic Voices. *International Journal of Pattern Recognition and Artificial Intelligence* 32, 1 (2018). https://doi.org/10.1142/S021800141860008X

[15] Scotty D. Craig and Noah L. Schroeder. 2017. Reconsidering the voice effect when learning from a virtual human. *Computers and Education* 114 (2017), 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

[16] Brody Downs, Aprajita Shukla, Mikey Krentz, Maria Soledad Pera, Katherine Landau Wright, Casey Kennington, and Jerry Fails. 2020. Guiding the Selection of Child Spellchecker Suggestions Using Audio and Visual Cues. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 398–408. https://doi.org/10.1145/3392063.3394390

[17] Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive Synthetic Speech: Voice Perception and User Behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 6, 9 pages. https://doi.org/10.1145/3405755.3406120

[18] Justin Edwards and Elaheh Sanoubari. 2019. A Need for Trust in Conversational Interface Research. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 21, 3 pages. https://doi.org/10.1145/3342775.3342809

[19] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

[20] B.J. Fogg. 2002. *Persuasive Technology: Using Computers to Change What We Think and Do.*

[21] Li Gong and Clifford Nass. 2007. When a Talking-Face Computer Agent is Half-Human and Half-Humanoid: Human Identity and Consistency Preference. *Human Communication Research* 33, 2 (2007), 163–193. https://doi.org/10.1111/j.1468-2958.2007.00295.x

[22] Rosanna E. Guadagno, Jim Blascovich, Jeremy N. Bailenson, and Cade Mccall. 2007. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology* 10, 1 (2007), 1–22. https://doi.org/10.108/15213260701300865

[23] Florian Habler, Valentin Schwind, and Niels Henze. 2019. Effects of Smart Virtual Assistants' Gender and Language. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 469–473. https://doi.org/10.1145/3340764.3344441

[24] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2009. *Multivariate Data Analysis (7th Edition).*

[25] Tilo Hartmann, Werner Wirth, Holger Schramm, Christoph Klimmt, Peter Vorderer, André Gysbers, Saskia Böcking, Niklas Ravaja, Jari Laarni, Timo Saari, Feliz Gouveia, and Ana Maria Sacau. 2016. The Spatial Presence Experience Scale (SPES). *Journal of Media Psychology* 28, 1 (2016), 1–15. https://doi.org/10.1027/1864-1105/a000137

[26] Holotech Studios. 2017. Audio based Lipsync. https://facerig.com/docs/facerig-studio-docs/advanced-user-interface/audio-based-lipsync/

[27] Holotech Studios. 2021. FaceRig. https://facerig.com/

[28] Xuedong Huang. 2018. Microsoft's new neural text-to-speech service helps machines speak like people. https://azure.microsoft.com/en-us/blog/microsoft-s-new-neural-text-to-speech-service-helps-machines-speak-like-people/

[29] Justine Humphry and Chris Chesher. 2021. Preparing for smart voice assistants: Cultural histories and media innovations. *New Media & Society* 23, 7 (2021), 1971–1988. https://doi.org/10.1177/1461444820923679

[30] Reshmashree B. Kantharaju, Dominic De Franco, Alison Pease, and Catherine Pelachaud. 2018. Is Two Better than One? Effects of Multiple Agents on User Persuasion. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 255–262. https://doi.org/10.1145/3267851.3267890

[31] Rabia Fatima Khan and Alistair Sutcliffe. 2014. Attractive Agents Are More Persuasive. *International Journal of Human-Computer Interaction* 30, 2 (2014), 142–150. https://doi.org/10.1080/10447318.2013.839904

[32] Yanghee Kim, Amy L Baylor, and Gabreille Reed. 2004. Pedagogical agents' personas: Which affects more, image or voice? *The Annual Conference of Association for Educational Communications and Technology (AECT)* October (2004). https://doi.org/10.13140/2.1.1802.4327

[33] Chang Liu and Diane Kewley-Port. 2004. Vowel formant discrimination for high-fidelity speech. *The Journal of the Acoustical Society of America* 116, 2 (2004), 1224–1233. https://doi.org/10.1121/1.1768958

[34] Nóra Ni Loideain and Rachel Adams. 2020. From Alexa to Siri and the GDPR: The gendering of Virtual Personal Assistants and the role of Data Protection Impact Assessments. *Computer Law and Security Review* 36 (2020), 1–14. https://doi.org/10.1016/j.clsr.2019.105366

[35] Ryan P. McMahan, Doug A. Bowman, David J. Zielinski, and Rachael B. Brady. 2012. Evaluating Display Fidelity and Interaction Fidelity in a Virtual Reality Game. *IEEE Transactions on Visualization and Computer Graphics* 18, 4 (2012), 626–633. https://doi.org/10.1109/TVCG.2012.43

[36] Ryan P. McMahan, Chengyuan Lai, and Swaroop K. Pal. 2016. Interaction fidelity: The uncanny valley of virtual reality interactions. *International Conference on Virtual, Augmented and Mixed Reality* 9740 (2016), 59–70. https://doi.org/10.1007/978-3-319-39907-2_6

[37] Microsoft Corporation. 2021. Cognitive Services pricing. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/speech-services/

[38] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[39] John W. Mullennix, Steven E. Stern, Stephen J. Wilson, and Corrie Lynn Dyson. 2003. Social perception of male and female computer synthesized speech. *Computers in Human Behavior* 19, 4 (2003), 407–424. https://doi.org/10.1016/S0747-5632(02)00081-X

[40] Gregory B. Murphy and Neil Tocher. 2011. Gender differences in the effectiveness of online trust building information cues: An empirical examination. *Journal of High Technology Management Research* 22, 1 (2011), 26–35. https://doi.org/10.1016/j.hitech.2011.03.004

[41] Procheta Nag and Özge Nilay Yalçın. 2020. Gender Stereotypes in Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 41, 8 pages. https://doi.org/10.1145/3383652.3423876

[42] Clliford Nass and Scott Brave. 2005. *Wired for Speech.*

[43] Greg Nyilasy. 2019. Fake news: When the dark side of persuasion takes over. *International Journal of Advertising* 38, 2 (2019), 336–342. https://doi.org/10.1080/02650487.2019.1586210

[44] Kohei Ogawa, Christoph Bartneck, Daisuke Sakamoto, Takayuki Kanda, Tetsuo Ono, and Hiroshi Ishiguro. 2018. Can an android persuade you? In *Geminoid Studies*. Springer Singapore, 235–247. https://doi.org/10.1109/ROMAN.2009.5326352

[45] Dhaval Parmar, Stefán Ólafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2020. Navigating the Combinatorics of Virtual Agent Design Space to Maximize Persuasion. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems,

Richland, SC, 1010–1018.

[46] Richard E. Petty and John T. Cacioppo. 1986. *Communication and persuasion: Central and peripheral routes to attitude change.* Springer-Verlag, New York.

[47] Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* 19, C (1986), 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

[48] Pierre Philip, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, and Jean Arthur Micoulaud-Franchi. 2020. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *npj Digital Medicine* 3, 2 (2020), 1–7. https://doi.org/10.1038/s41746-019-0213-y

[49] Delphine Potdevin, Nicolas Sabouret, and Céline Clavel. 2020. An intimate virtual counselor for a better user experience. In *ACM International Conference on Intelligent Virtual Agents 2020 (IVA 2020).* 1–3. https://doi.org/10.1145/3383652.3423859

[50] Nisha Raghunath, Paris Myers, Christopher A. Sanchez, and Naomi T. Fitter. 2021. Women Are Funny: Influence of Apparent Gender and Embodiment in Robot Comedy. In *Social Robotics*, Haizhou Li, Shuzhi Sam Ge, Yan Wu, Agnieszka Wykowska, Hongsheng He, Xiaorui Liu, Dongyu Li, and Jairo Perez-Osorio (Eds.). Springer International Publishing, Cham, 3–13.

[51] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD).* 1–10. https://doi.org/10.1109/SPED.2019.8906599

[52] Jon Rogers, Loraine Clarke, Martin Skelly, Nick Taylor, Pete Thomas, Michelle Thorne, Solana Larsen, Katarzyna Odrozek, Julia Kloiber, Peter Bihr, Anab Jain, Jon Arden, and Max von Grafenstein. 2019. Our Friends Electric: Reflections on Advocacy and Design Research for the Voice Enabled Internet. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300344

[53] Maaike Roubroeks, Jaap Ham, and Cees Midden. 2011. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics* 3 (2011), 155–165. https://doi.org/10.1007/s12369-010-0088-1

[54] Neil Salkind. 2010. *Tukey's Honestly Significant Difference (HSD) Test.* Sage, Thousand Oaks, CA. 1–5 pages. https://doi.org/10.4135/9781412961288.n181

[55] Daniel Schulman and Timothy Bickmore. 2009. Persuading Users through Counseling Dialogue with a Conversational Agent. In *Proceedings of the 4th International Conference on Persuasive Technology* (Claremont, California, USA) *(Persuasive '09).* Association for Computing Machinery, New York, NY, USA, Article 25, 8 pages. https://doi.org/10.1145/1541948.1541983

[56] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (2021), 43 pages. https://doi.org/10.1145/3386867

[57] Julien Simon. 2019. Amazon Polly Introduces Neural Text-To-Speech and Newscaster Style. https://aws.amazon.com/blogs/aws/amazon-polly-introduces-neural-text-to-speech-and-newscaster-style/

[58] Paul Skalski and Ron Tamborini. 2007. The role of social presence in interactive agent-based persuasion. *Media Psychology* 10, 3 (2007), 385–413. https://doi.org/10.1080/15213260701533102

[59] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26, 4 (jun 2019), 62–65. https://doi.org/10.1145/3338283

[60] Steven E. Stern, John W. Mullennix, Corrie Lynn Dyson, and Stephen J. Wilson. 1999. The persuasiveness of synthetic speech versus human speech. *Human Factors* 41, 4 (1999), 588–595. https://doi.org/10.1518/001872099779656680

[61] Steven E. Stern, John W. Mullennix, and Ilya Yaroslavsky. 2006. Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human Computer Studies* 64, 1 (2006), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

[62] Su-Mae Tan and Tze Wei Liew. 2020. Designing Embodied Virtual Agents as Product Specialists in a Multi-Product Category E-Commerce: The Roles of Source Credibility and Social Presence. *International Journal of Human–Computer Interaction* 36, 12 (2020), 1136–1149. https://doi.org/10.1080/10447318.2020.1722399

[63] Cara Tannenbaum, Robert P. Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. 2019. Sex and gender analysis improves science and engineering. *Nature* 575 (2019), 137–146. https://doi.org/10.1038/s41586-019-1657-6

[64] Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, and Naomi Harte. 2018. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. *2018 International Conference on Multimodal Interaction (ICMI '18)* (2018), 423–427. https://doi.org/10.1145/3242969.3242984

[65] Renske Van Enschot-Van Dijk, Lettica Hustinx, and Hans Hoeken. 2003. *The Concept of Argument Quality in the Elaboration Likelihood Model.* Springer Netherlands, Dordrecht, 319–335. https://doi.org/10.1007/978-94-007-1078-8_25

[66] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal Component Analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[67] James E. Young, Richard Hawkins, Ehud Sharlin, and Takeo Igarashi. 2009. Toward acceptable domestic robots: Applying insights from social psychology. *International Journal of Social Robotics* 1, 1 (2009), 95–108. https://doi.org/10.1007/s12369-008-0006-y

[68] Catherine Zanbaka, Paula Goolkasian, and Larry Hodges. 2006. Can a Virtual Cat Persuade You? The Role of Gender and Realism in Speaker Persuasiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06).* Association for Computing Machinery, New York, NY, USA, 1153–1162. https://doi.org/10.1145/1124772.1124945

[69] Leslie A Zebrowitz. 1990. *Social perception.* Thomson Brooks/Cole Publishing Co.

[70] Sebastian Zepf, Neska El Haouij, Wolfgang Minker, Javier Hernandez, and Rosalind W. Picard. 2020. EmpathicGPS: Exploring the Role of Voice Tonality in Navigation Systems during Simulated Driving. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20).* Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382935

[71] Katja Zibrek, Ludovic Hoyet, Kerstin Ruhland, and Rachel McDonnell. 2013. Evaluating the Effect of Emotion on Gender Recognition in Virtual Humans. In *Proceedings of the ACM Symposium on Applied Perception* (Dublin, Ireland) *(SAP '13).* Association for Computing Machinery, New York, NY, USA, 45–49. https://doi.org/10.1145/2492494.2492510

[72] Katja Zibrek and Rachel McDonnell. 2014. Does Render Style Affect Perception of Personality in Virtual Humans?. In *Proceedings of the ACM Symposium on Applied Perception* (Vancouver, British Columbia, Canada) *(SAP '14).* Association for Computing Machinery, New York, NY, USA, 111–115. https://doi.org/10.1145/2628257.2628270

A ATTITUDE ASSESSMENT QUESTIONNAIRE

The following statements, shown grouped by topic, were administered to the participant [46]. Statements marked with an asterisk are reverse scored. This asterisk is not shown to the participant.

Environment

- The proper disposal of industrial toxic waste is one of the most serious problems facing our country.
- **The "greenhouse effect" is not as serious as the media would have us believe.
- Oil drilling off the coast of California should not be allowed under any circumstances.

Comprehensive Exams

- Required comprehensive exams before college graduation, in a student's major, can benefit both the student and the university through increased corporate and individual donations.
- **Required comprehensive exams before college graduation, in a student's major, are a waste of time and money for the student and the university.
- Students attending universities that require comprehensive exams have higher chances of getting better paying jobs.

Animal Rights

- The use of animals for research purposes is inhumane and morally unjustified.
- **Animal experimentation is an essential tool for scientific and medical research.
- **Research involving animal subjects may some day be instrumental in saving the life of your child or the child of someone close to you.

Tuition Raises

- **A 5 percent raise in tuition would be an unfair burden on the students who are attending the university.
- A 5 percent raise in tuition could substantially raise the quality of the education at a university.
- The income raised by a 5 percent tuition hike could raise the quality of life for the students who are there.

B SOCIAL PERCEPTION ANCHORS

- **Perception of the Speaker:** Incompetent-Competent, Dishonest-Honest, Unassertive-Assertive, Uninformed-Informed, Untrustworthy-Trustworthy, Timid-Bold, Unintelligent-Intelligent, Evasive-Straightforward, Inactive-Active, Unqualified-Qualified, Insincere-Sincere, Meek-Forceful
- **Perception of the Message:** Boring-Stimulating, Vague-Specific, Unsupported-Supported, Complex-Simple, Unconvincing-Convincing, Uninteresting-Interesting
- **Perception of the Argument:** Bad-Good, Foolish-Wise, Negative-Positive, Beneficial-Harmful, Effective-Ineffective, Convincing-Unconvincing
- **Perception of the Voice:** Loud-Soft spoken, Deep-Squeaky, Fast speaking-Slow speaking, Heavy accent-Faint accent, Talked too long-Didn't talk long enough, Heavy nasality-Faint nasality

C PERCEPTION FACTORS

Factor Eigenvalues, Composite Reliability (CR), and Average Variance Extracted (AVE).

Factor	Eigenvalue	%Variance	CR	AVE
Perception of the Speaker				
Competent	4.15	34.61	0.82	0.79
Trustworthy	2.16	17.96	0.81	0.52
Assertive	1.25	10.41	0.85	0.65
Perception of the Message				
Interesting	2.01	40.24	0.87	0.76
Supported	1.25	25.06	0.74	0.59
Perception of the Argument				
Effective	4.47	74.52	0.95	0.75
Perception of the Voice				
Nondistinctive	1.67	23.83	0.78	0.64
Lively	1.50	21.43	0.75	0.60
Unconfident	1.18	16.86	0.76	0.61