2021-09-14

Human or Robot?: Investigating voice, appearance and gesture motion realism of conversational social agents

Ylva Ferstl
Trinity College Dublin, yferstl@tcd.ie

Sean Thomas
Technological University Dublin, seanhtho@gmail.com

Cédric Guiard
Eisko, cedric.guiard@eisko.com

See next page for additional authors

Recommended Citation

Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. Association for Computing Machinery, New York, NY, USA, 76–83. DOI:10.1145/3472306.3478338

Funder: D-REAL, ADAPT, RADICal

Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents

Ylva Ferstl
yferstl@tcd.ie
Trinity College Dublin
Dublin, Ireland

Sean Thomas
sean.a.thomas@mytudublin.ie
Technological University Dublin
Dublin, Ireland

Cédric Guiard
cedric.guiard@eisko.com
Eisko
Paris, France

Cathy Ennis
cathy.ennis@tudublin.ie
Technological University Dublin
Dublin, Ireland

Rachel McDonnell
ramcdonn@tcd.ie
Trinity College Dublin
Dublin, Ireland

Figure 1: We explore the impact of voice, motion, and appearance realism configuration on an agent's perceived likability, human-likenes, and speech-gesture match. Shown are two sample frames of the *Human* and the *Robot* character in our *Full Natural* motion condition.

ABSTRACT

Research on creation of virtual humans enables increasing automatization of their behavior, including synthesis of verbal and non-verbal behavior. As the achievable realism of different aspects of agent design evolves asynchronously, it is important to understand if and how divergence in realism between behavioral channels can elicit negative user responses. Specifically, in this work, we investigate the question of whether autonomous virtual agents relying on synthetic text-to-speech voices should portray a corresponding level of realism in the non-verbal channels of motion and visual appearance, or if, alternatively, the best available realism of each channel should be used. In two perceptual studies, we assess how realism of voice, motion, and appearance influence the perceived match of speech and gesture motion, as well as the agent's likability and human-likeness. Our results suggest that maximizing realism of voice and motion is preferable even when this leads to realism mismatches, but for visual appearance, lower realism may be preferable. (A video abstract can be found at https://youtu.be/arfZZ-hxD1Y.)

CCS CONCEPTS

• **Computing methodologies** → **Animation**.

KEYWORDS

gesture motion, text-to-speech, animation style, perception, conversational agents, agent design, human-computer interfaces, anthropomorphism

ACM Reference Format:
Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478338

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09.
https://doi.org/10.1145/3472306.3478338

1 INTRODUCTION

Virtual agents are now widely used in human-computer interfaces, such as virtual assistants or chatbots. Many of these agents rely on text-to-speech (TTS) technology to produce spoken utterances and allow more natural interaction. However, commercial TTS output is

commonly easily recognizable as synthetic speech through a clear lack of voice naturalness. Similarly automatically generated nonverbal behavior for the virtual agents, specifically co-speech gesture, still falls noticeably behind motion-captured or hand-animated performances. While one might aim for the highest possible realism for each channel of behavior, this may lead to mismatches across modalities, with, for example, natural full-body animation paired with a synthetic TTS voice. Indeed, some findings suggest that congruence in realism may be preferred even when this decreases the realism of one of the behavioral channels [39]. In this work, we examine if realism should be matched across modalities.

In two user studies, we investigated perceived speech-gesture match, as well as agent likability and anthropomorphism under variations of realism of voice, motion, and appearance. We examine whether congruence in realism between channels is preferred over using the highest available realism each. For voice, we compare real speech recordings and text-to-speech audio; for motion, we design 4 levels of animation realism from full motion capture to motion only preserving the gesture strokes; and for appearance, we compare a human and a *Robot* character. In the first study, users rated how well the gestures match the speech under the different combinations of channel realism. With this, we aimed to examine the question of optimal motion style when relying on synthetic speech production and how agent appearance may affect this. In the second study, users judged the agent's likability and anthropomorphism, key factors in the design of conversing social agents.

Our results show that higher motion and voice realism independently improve speech-gesture match, as well as perceived likability and human-likeness. These results speak for the use of the highest available realism each for voice and motion, even if this leads to realism mismatches. Our results also suggest that more creative freedom may be used in character choice for gesture generation systems, with character realism not significantly impacting gesture match or human-likeness. However, participants perceived our non-human character to be more likeable. Our results provide guidelines for developers of social agents and gesture generation systems.

2 RELATED WORK

Gesture generation from speech: Numerous works have explored methods of automatically generating gesture motion for speech audio (e.g. [13, 29, 39]) or text input (e.g. [1, 35, 58]). Evaluations of gesture generation systems have focused on how close generated motion is to natural gesture behavior, however, on open question remains if natural human motion is the best choice for a character with limited realism in other modalities. In a study by Ondras et al. [39], lower realism, "machine-like" motion was rated as matching synthetic text-to-speech audio better than more natural motion on the Pepper robot. Determining appropriate motion style is therefore an important open question for the design of embodied conversational agents. Regarding visual realism, previous works have used everything from simplistic stick figures [19], to 3D models of a lay figure [29], to low- [43] and higher realism humanoid models [55], robots both as 3D models [59] and video-taped [26, 39] as well as "live" physical robots [23, 24]. However, performances of individual gesture generation systems have not been compared across levels of character realism. A mismatch in realism of generated voice and appearance was found to be perceived as eerie [34], particularly for a visually realistic character, and appropriate choice of character may be therefore be important for systems relying on text-to-speech audio.

Motion perception: Perceptual research has shown that body motion is a key factor in the perceived realism and likability of virtual humans. Ondřej et al. [40] found body motion to be as important for attractiveness as physical appearance in scripted multi-modal performances. Body motion was even found to be more important than facial motion for the recognition of a number of emotions [11]. For interactions between conversational characters, body motion is an important cue in the plausibility of unscripted group conversations. Ennis et al. [12] found that observers are more sensitive to temporal mis-alignments of body motions between agents within one group conversation, compared to mismatches between the agents' voices and gestures. For conversing agents, gesture motion is an important part of the overall animated performance. In a study by Salem et al. [44], when a robot used gestures, it was rated as more likeable and human-like. Neff [37] showed that virtual characters are perceived as more natural, friendly and trustworthy when producing more complex gesture sequences rather than singleton gestures. Ferstl et al. [14] showed that even small alterations, such as wider hand opening, that do not change the gesture content impact the perceived congruence of speech and gesture.

Speech synthesis and perception: Research into the development of virtual agents is interested in creating fully automatic behavior, including not only the generation of motion, but also the verbal utterances of the agents. Text-to-speech systems for conversational dialog are often trained on scripted utterances [48], but this can fail to capture the natural variation of prosody in spontaneous, unscripted utterances [20]. Natural conversational language contains frequent repetitions, hesitations, and false starts [18], backchanneling, short utterances, limited vocabulary and many colloquialisms [7], and may even lack any sentence-delimiting mark [4], all of which can sound off-putting when played back with TTS lacking the original prosodic variations. Real human voices are consistently preferred over TTS voices, being rated as more expressive and likeable [6], and the more human-like the voice, the less eerie it appears [2, 30]. Voice qualities such as pronunciation and emotion were furthermore found to be one of the eight key dimensions most frequently used to assess humanness of an agent in a speech interface [10], along constructs such as interpersonal connection and conversational interactivity. TTS engines mimicking natural conversational speech are an active area of research [47, 49]. For embodied virtual agents, combined synthesis of speech and appropriate body motion has also been proposed [1].

Uncanny Valley: We have discussed that mismatches in realism of an agent's voice and appearance can elicit negative reactions [34]. This and similar findings have been attributed to the Uncanny Valley effect, originally hypothesized by Mori [36], stating that the more human a robot's appearance becomes, the more it evokes positive and empathetic responses, until a tipping point,

where subtle shortcomings in human-likeness cause feelings of eeriness, or even disgust and fear, similar to an animated corpse.

Much of the support for the Uncanny Valley hypothesis has come from games and movies rather than targeted research. Creators of Princess Fiona in "Shrek" stated they deliberately made Fiona less human-like when she had become "too real, and the effect was getting distinctly unpleasant" [54], and the lack of success of the movie "Polar Express" has often been attributed to Uncanny Valley effects.

A number of researchers have contested the standard explanation of the Uncanny Valley phenomenon. Bartneck et al. [3] found no difference in the likability of a human and the highly humanoid physical robot Geminoid HI-1, despite the robot's visual closeness to the real human. The authors also varied realism of motion, with a minimal movement condition of only eye blinks and lip synchronisation, and an increased realism condition also including eye gaze variation and subtle random movements of the body. Interestingly, the authors did not find more realistic movement to significantly increase the robot's human-likeness or likability, though the human was perceived as less human-like when producing limited movement compared to full movement. Hanson et al. [21] found no dip in users' acceptance for faces ranging from cartoon to realistic, and MacDorman et al. [31] reported that CG faces did not follow the expectation of appearing most eerie when falling just short of human. McDonnell et al. [33] note that contrary to the Uncanny Valley theory, the most realistic CG faces in their study were rated similarly on appeal and likability as cartoon faces, but the authors did observe a dip in ratings of appeal for faces midway between abstract and realistic. Thompson et al. [50] assessed perceptions of human-likeness across increasing levels of motion realism for a CG lay figure and a (relatively low-realism) human character and found no difference between characters: Human-likeness ratings increased monotonically with higher realism motion, and eeriness decreasing in the same pattern for both characters. Piwek et al. [42] found natural motion to improve acceptability of CG characters, but Urgen et al. [53] found that a mismatch in realism of appearance and motion causes Uncanny Valley effect for a visually highly realistic physical robot.

The evidence so far suggest a potential Uncanny Valley danger for visually highly realistic characters falling short on other aspects, whereas we are not aware of any studies directly investigating if highly realistic motion requires matching realism in the voice, or vice-versa.

3 PERCEPTUAL STUDIES

We designed two perceptual studies investigating the effects of voice, motion, and character appearance realism on perceived speech-gesture match (study I), and agent likability and human-likeness (study II). There were 2 voice and appearance conditions, and 4 motion conditions, described below.

We obtained all data for creating the experiment stimuli from the Trinity Speech-Gesture II dataset [15]. This dataset contains 6 hours of audio recordings and full-body motion capture of spontaneous, conversational speech of one male native English speaker. For 4 hours of this dataset, gesture phases have previously been hand-annotated [13], segmenting the motion data into individual gestures. Specifically, we utilize the hand-annotation of the gestures' stroke phases, the core expressive phase of a gesture [28].

3.1 Character appearance conditions

Our *Human* character is a custom high-end 3D-scanned model, created by Eisko, a leading Digital Double company (see Figure 1 (left)). The character has over 200 scanned blendshapes, including phonemes for speech. The character was rendered in Unity's High Definition Render Pipeline using a HDRI background[1] and state-of-the-art shaders and advanced lighting and post-processing effects to reach as high a level of realism as possible.

For the *Robot* character, we used Unity3D Kyle model. The two characters and our render view are illustrated in Fig. 1.

3.2 Speech conditions

For voice, the *Real* condition represented the original speech audio recording. To create the *TTS* condition, the real speech was transcribed and passed through IBM's Watson Text-To-Speech API using the US Henry-V3 voice. TTS speech was pre-aligned by passing the transcription's timing as additional attributes into the API. Using the generated TTS audio, we then manually refined alignment of the TTS with the real voice in Audacity. This alignment was necessary to preserve the original speech-motion relationship. The IBM's Watson service was chosen for its frequent use in commercial systems, and the Henry-V3 voice was determined as most closely resembling the real speech voice.

3.3 Motion conditions

For body motion, the 4 conditions were:

(1) **Full Natural:** Full-body motion capture to animate the virtual character.
(2) **Natural Idle:** Motion capture for the arms and hands, and idle body motion for the remainder of the body.
(3) **Robotic Gestures:** Motion capture of the gesture stroke phase plus synthesized transitions for the arms and hands, and idle motion for the remainder of the body.
(4) **Robotic Gestures Reduced:** Same as previous but every other gesture stroke is removed (gesture rate reduced by 50%).

The motion conditions were designed to mimic the most common types of animation procedures used in current systems. The *Natural Idle* condition was designed to represent the common choice of gesture generation systems to synthesize only upper body motion [5, 29, 58], and the *Robotic* conditions to represent stroke-based, gesture-by-gesture systems [16, 32, 38].

The robotic styles were generated with software based on the open-source animation environment DANCE [45], which takes input motion data and a list of stroke timings and produces output motion with synthesized gesture transitions (preparations, retractions, and direct transitions) using splines. This processing retains the core gesture information, contained in the stroke phase, while creating an overall more synthetic or robotic appearance.

[1] https://hdrihaven.com/

3.4 Lip synchronization and eye blinks

Lip synchronization for the *Human* character was produced with the Oculus Lipsync Unity plugin, taking input audio and producing viseme blendshape activations. We animated natural eye gaze with the character maintaining eye-contact with the virtual camera most of the time but sometimes glancing elsewhere. For the *Robot* character, which did not have a mouth, the eyes were animated to slightly flash with the rhythm of the speech.

For the *Human* character, we procedurally animated eye blink behavior. Average conversational Spontaneous Eyeblink Rate (SEBR) values are between 10.5 and 32.5 blinks per minute for adults during conversation [9]. Based on this, we set a minimum (X_1) and maximum (X_T) blink interval threshold of X_1 = (60/32.5) = *1.84 seconds* and X_T = (60/10.5) = *5.71 seconds*. A single set of intervals was randomly generated with values ranging from X_1 to X_T to ensure that there was no variance in blink times between different motion stimuli. Following the lead of other authors in the field [56], the animation of each individual blink lasts *0.17 seconds*, with *0.07 seconds* to close the eyes and *0.1 seconds* to reopen [56].

3.5 Procedure

We had a 2x4x2 within-subjects factorial design: 2 voice conditions (*Real* and *TTS*), 4 motion conditions (*Full Natural*, *Natural Idle*, *Robotic*, and *Robotic Reduced*), and 2 appearance conditions (*Human* and *Robot*).

Participants first read the study instructions and completed a short training showing representative samples of the experiment stimuli to familiarize themselves with the task and to establish an expectation of the range of animations and characters in the experiment. Participants were then informed about attention checks and completed an example.

There were 2 trials for each condition, a total of 32 trials. Each trial consisted of watching a 15 second video clip followed by a 7-point Likert scale rating of either speech-gesture match (study I) or likability and human-likeness (study II). The chosen 15 second duration was similar to previous works on evaluation gesture generation models (10s in Kucherenko et al. [29], 5-10s in Yoon et al. [57], <15s in Ondras et al. [39]), as well as based on the finding of Ennis et al. [12] that 10 seconds are sufficient for participants to judge conversational behavior of virtual agents.

There were 16 different speech segments; per participant, each speech segment was randomly assigned to one of the 8 condition combinations of voice and motion to avoid bias between conditions, and the participant then saw this speech-condition configuration on each character during the experiment. All clips were presented in randomized order.

Study completion time was approximately 20 minutes. We recruited participants via Prolific and required fluency in English. For quality control of participants' responses, four attention checks were placed randomly within each quartile of experiment trials: Instead of the expected rating question, attention trials asked a multiple choice question about what the speaker said. Participants were reimbursed at an above Prolific-average rate of 10.20 Euro per hour. We considered this appropriate since online crowdworkers have been shown to yield comparable results to in-lab participants, particularly so with fair reimbursement [27].

Please see the supplemental movie[2] for additional details on the stimuli creation, and the full set of stimuli can be accessed here[3]. Note that all links are anonymous.

4 STUDY I - SPEECH-GESTURE MATCH

Our first study focused on the perceived match between two of the three manipulated behavioral channels, speech and motion. We hypothesized that for an agent with a TTS voice, lacking the rich variation of prosody and emphasis of a real voice, a robotic style of motion may be deemed more appropriate. Due to the lack of speech emphasis, we also hypothesized that less frequent gesturing may appear more appropriate.

Participants rated the statement, "How appropriate were the gestures for the speech?", on a 7-point Likert scale ranging from "Very bad match" (1) to "Very good match" (7). The phrasing of the question prompt mirrors that of the GENEA gesture generation challenge [29]. 21 participants completed the study (10 females, 1 other, aged 18-35 years, M = 24.9, SD = 5.2), of which 1 participant failed more than 1 attention check and was rejected.

4.1 Results

We conducted ANOVA analysis when the normality assumption was not violated (Shapiro–Wilk's normality test). Specifically, we computed a three-way ANOVA with repeated measures and the within-group factors of voice, motion, and character. When the sphericity assumption was violated (Mauchly's sphericity test), the degrees of freedom were corrected using the Greenhouse–Geisser correction. Estimated Marginal Means was used to check significances for pairwise comparisons. When the normality assumption was violated, Aligned Rank Transform was used.

The results for the perceived speech-gesture match are visualized in Fig. 2 (left). There was a significant main effect of voice and motion, and a significant three-way interaction of voice, motion, and character appearance (see Table 1 for all effect sizes). *Real Voice* significantly increased perceived speech-gesture match.

Full Natural motion significantly increased perceived speech-gesture match compared to all other motion conditions ($p < 0.05$ for *Natural Idle*, $p < 0.001$ for *Robotic* and *Robotic Reduced*). *Natural Idle* significantly increased speech-gesture match compared to *Robotic* ($p < 0.01$) and *Robotic Reduced* ($p < 0.001$). *Robotic* motion significantly improved speech-gesture match compared to *Robotic Reduced* ($p < 0.001$).

4.2 Discussion

Voice and motion realism significantly affected speech-gesture match, with higher realism eliciting more positive ratings. This effect was independent for the two channels: For a low-realism TTS voice, high-realism motion was preferred over "matched" lower realism motion. We therefore reject our hypothesis that TTS voices lacking the colorfulness of natural speech are better matched with more robotic motion. We also reject our hypothesis that a reduced rate of gesturing is more appropriate for the emphasis-lacking TTS voice.

[2] https://youtu.be/GVWzVs45FAs
[3] https://youtube.com/playlist?list=PL04OYHqbFqt01IvcvboZPBDAQP1F9BaaT

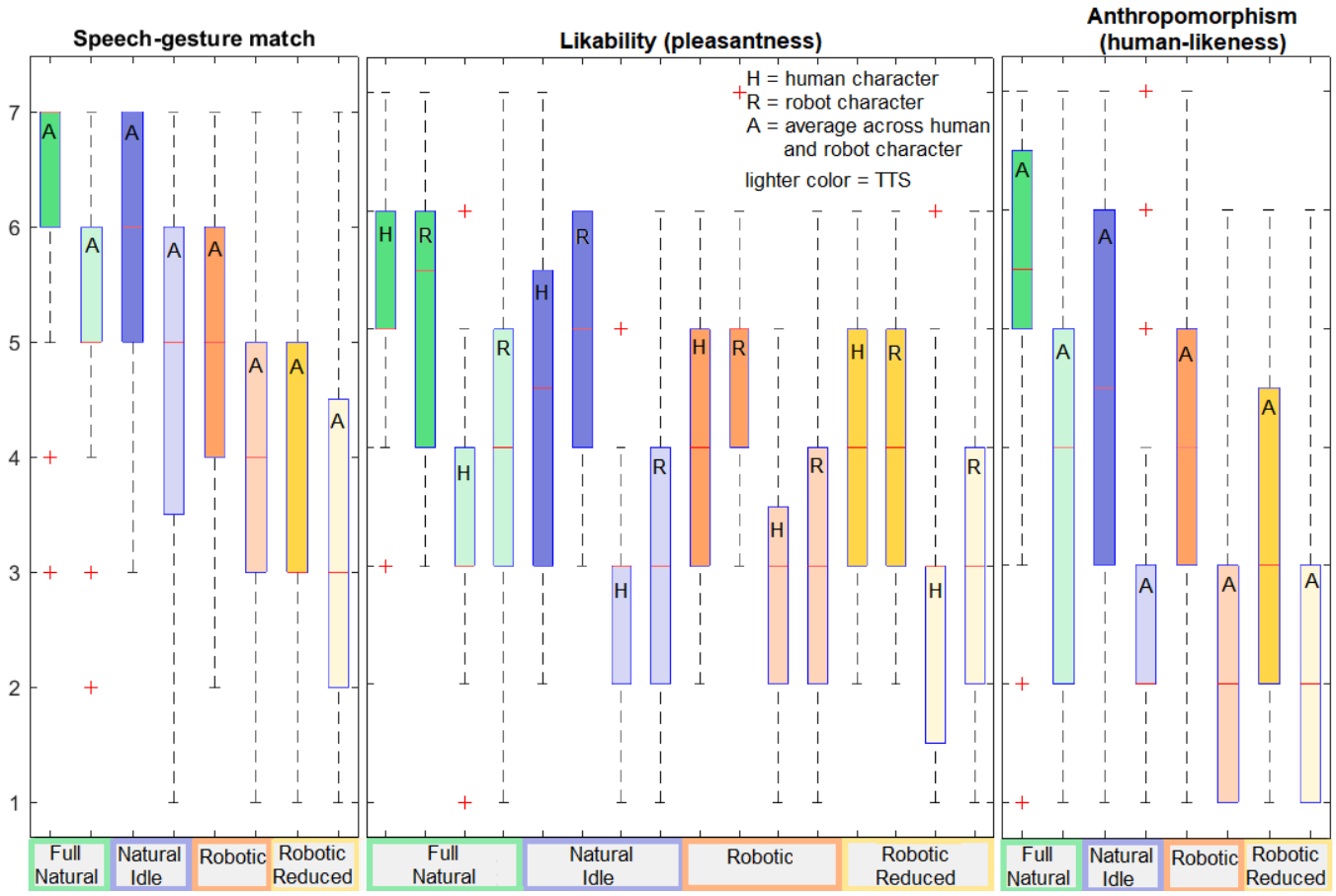

Figure 2: Boxplots for all main effects of study I (left) and II (middle & right). Lighter coloring indicates the *TTS* voice condition, darker colorings represent the *Real* voice condition. Left: For speech-gesture match (study I), motion and voice, had a significant effect, but not appearance. Speech-gesture match results are averaged over the *Human* and the *Robot* character. Middle: Likability (study II) was significantly influenced by motion, voice, and character appearance. Right: Anthropomorphism (study II) was significantly influenced by motion and voice. Anthropomorphism results are averaged over characters.

5 STUDY II - LIKABILITY AND ANTHROPOMORPHISM

In our second study, we assessed the effects of voice, motion, and appearance on user perceived likability and anthropomorphism. Likability and anthropomorphism are of key interest in the design of virtual agents and together can reveal potential Uncanny Valley effects, with a dip in likability under increasing anthropomorphism. Based on the previous research of Sec. 2, we hypothesized that higher voice and motion realism would be perceived as more likeable and human-like, and that lower voice and motion quality would impact likability of the *Human* character more strongly. We also hypothesized that the *Human* character would be rated higher on anthropomorphism. A separate group of participants saw the same video clips as participants in study I, but instead rated the character on 7-point Likert scales ranging from "very unpleasant" to "very pleasant", and "very machinelike" to "very humanlike", items from the Godspeed questionnaire [22], for likability and anthropomorphism, respectively. A single question prompt for each dimension was chosen for task simplicity and the phrasing followed that of Godspeed: "Please rate your impression of the character". 21 participants completed the study (10 females, 1 other, aged 18-54 years, M = 27.2, SD = 8.9), of which 1 participant failed more than 1 attention check and was rejected.

5.1 Results

Statistical analysis was conducted as described in Sec. 4.1. Results are visualized in Fig. 2 (middle & right) and effect sizes are reported in Table 1.

Likability: There was a significant main effect of voice, motion, and character, with no significant interactions (see Table 1). *Real Voice* significantly increased likability. *Full Natural* motion significantly increased likability compared to all other motion conditions, and *Natural Idle* significantly increased likability compared to *Robotic Reduced*. The *Robot* character was rated significantly more likeable.

Anthropomorphism: There was a significant main effect of voice and motion, and no significant interactions. *Real Voice* was rated significantly more human-like. *Full Natural* motion was rated significantly more human-like than all other motion conditions, and *Natural Idle* significantly increased human-likeness compared to *Robotic Reduced*.

5.2 Discussion

Voice and motion realism significantly again independently affected both likability and human-likeness, with higher realism rated more positively. For likability, appearance also had a significant effect, with, interestingly, the *Robot* character being preferred over the *Human* character.

Based on our results, we accept our hypothesis that higher realism voice and motion are perceived as more pleasant, and that (in)congruence of voice and motion has no effect. We found no interaction of character appearance with motion or voice, and so we reject the hypothesis that the *Human* character's likability is impacted more strongly by decreases in motion or voice realism. However, the higher likability of the *Robot* character could in part be due to most of the human conditions containing reduced realism or voice and/or motion (7 out of the 8 configurations). This is also visible in Fig. 2, where the human's likability ratings appear comparable to the robot's for *Full Natural* motion and *Real* voice, and less so for the degraded conditions.

We also reject our hypothesis that the *Human* character is more human-like. Instead, human-likeness appears to be driven by voice and motion realism only. Alternatively, the *Human* character's realistic visual appearance could have elicited high expectations regarding human-likeness that were not met sufficiently for his motion and voice in most conditions, lowering users' ratings of human-likeness. For the *Robot* character, users may have lower expectations and hence be more forgiving of such shortcomings. In line with our findings, Chaminade et al. [8] also suggest imperfect motion capture on a human character could induce negative feelings; the authors find motions portrayed on anthropomorphic characters to be perceived as less natural.

The higher likability ratings for the *Robot* character could be due to the specific characters used, or due to a novelty effect for the *Robot* character. Another potential reason is the imperfect lipsync, used only for the *Human* character, since the robot did not have a mouth. Any other imperfect animation effects may also have been more perceivable on the detailed *Human* character. Finally, the combination of imperfect animation with high-realism appearance may have had an uncanny valley effect for the *Human* character.

6 GENERAL DISCUSSION

We present the first investigation into the importance of the three agent design aspects of voice, motion, and appearance on perceived speech-gesture match as well as likability and anthropomorphism in two perceptual studies. For voice, we compared text-to-speech generated audio, as commonly used for virtual agents with verbal behavior, and original audio recordings. For motion, we designed 4 levels of realism, ranging from full-body motion capture to arm gestures with reduced frequency and synthetic gesture transitions. For appearance, we compared a high-realism human character and a robot character.

In our first study, we measured the perceived match between the speech and motion of the animated character. We found no effect for character, but a significant impact of both motion and voice, with higher realism always eliciting more positive ratings. Our findings imply that even for characters producing lower realism, "robotic" TTS verbal utterances, one should aim for high-realism motion rather than matching the more robotic voice style. Similarly, for the verbal utterances, the highest available voice realism should be used independently of available motion realism. We found no effect of character appearance on speech-gesture match, again emphasizing the preference of higher realism voice and motion irrespective of character choice, as well as pointing to more creative freedom for character appearance. This also indicates that evaluation of gesture generation systems may be relatively robust across the variations of 3D characters of previous studies (see Sec. 2).

In our second study, we assessed likability and anthropomorphism of the portrayed character. For both dimensions, we again found higher realism voice and motion to be consistently preferred. Voice realism had an especially large effect on likability, with the real voice much preferred. Human-likeness was also strongly affected by voice realism, followed by motion realism. In comparing effect strengths of voice and motion, however, it is important to note that there were only 2 voice conditions with greatly different level of realism, whereas there were 4 motion conditions with subtler differences between levels of realism. Contrary to the finding of Bartneck et al. [3], higher realism motion did increase likability and anthropomorphism of the *Robot* character. Bartneck et al. [3]'s robot is however distinctly dissimilar to our *Robot* character, with the former being much closer to a human in appearance, as well as being a physical robot. In line with the findings of Thompson et al. [50] and Piwek et al. [42], we found that increasing motion realism consistently increased ratings of anthropomorphism and likability.

Character appearance only had a significant effect on likability in favor of the *Robot* character. This may be pointing to a Uncanny Valley effect for the high-realism *Human* character paired with imperfect animation. Alternatively, stylized characters may be preferred irrespective of animation realism. McDonnell et al. [33] previously also found stylized, lower realism (cartoon) characters to be perceived as more friendly and pleasant than human faces, with similar ratings of appeal and trustworthiness, and in Torre et al. [51] such a lower realism character was rated as more appealing, attractive, and happier than a human face. Interestingly, in our study, character appearance did not affect human-likeness ratings: The *Robot* character was rated equally human-like than the *Human* character, again pointing to the driving forces of voice and motion realism.

Overall, our results show that optimal voice and motion realism should be preferred, even when this leads to realism mismatches between channels, speaking against the existence of an Uncanny Valley for diverging realism of motion and voice and in line with the findings of Parmar et al. [41]. For character appearance, visual realism may only impact likability, with less realistic characters being preferred.

For evaluating gesture generation systems, based on our results, we find the use of lower realism characters appropriate; highly

Table 1: Summary of all main effects and interactions with corresponding post–hoc analysis.

Effect	Test	Post-hoc
Speech-gesture match	ANOVA	
Voice	$F_{1,19} = 32.02$, $p < 0.001$, $\eta^2 = 0.63$	*Real Voice* higher ($p < 0.001$).
Motion	$F_{1.7,32.6} = 53.98$, $p < 0.001$, $\eta^2 = 0.74$	*Full Natural* motion higher compared to all others ($p < 0.05$ for *Natural Idle*, $p < 0.001$ for *Robotic* and *Robotic Reduced*). *Natural Idle* higher than *Robotic* ($p < 0.01$) and *Robotic Reduced* ($p < 0.001$). *Robotic* higher than *Robotic Reduced* ($p < 0.001$).
Voice * Motion * Character	$F_{2.4,45.3} = 3.43$, $p < 0.05$, $\eta^2 = 0.15$	No interesting significant interactions found in post-hoc.
Likability	Aligned Rank Transform	
Voice	$F_{1,605} = 257.69$, $p < 0.001$, $\eta^2 = 0.70$	*Real Voice* higher ($p < 0.001$).
Motion	$F_{3,605} = 22.25$, $p < 0.001$, $\eta^2 = 0.54$	*Full Natural* higher compared to all others (all $p < 0.001$). *Natural Idle* higher than *Robotic Reduced* ($p < 0.01$)
Character	$F_{1,605} = 12.19$, $p < 0.001$, $\eta^2 = 0.22$	*Robot* character more likeable ($p < 0.05$).
Anthropomorphism	ANOVA	
Voice	$F_{1,19} = 52.59$, $p < 0.001$, $\eta^2 = 0.73$	*Real Voice* more human-like ($p < 0.001$).
Motion	$F_{1.7,31.5} = 30.06$, $p < 0.001$, $\eta^2 = 0.61$	*Full Natural* motion more human-like than all others (all $p < 0.001$). *Natural Idle* more human-like than *Robotic Reduced* ($p < 0.001$).

realistic characters may require additional resources, such as lip synchronization and other facial motion, while providing no benefits for perceived speech-gesture match, agent likability, or human-likeness. Furthermore, for evaluating generated gesture motion of systems using TTS, a fair comparison to ground truth motion should display the latter with aligned TTS also. This avoids confounding effects of voice quality with gesture motion quality, since voice quality independently affects how well gesture motion is perceived to match the speech.

7 LIMITATIONS AND FUTURE WORK

In this work, we only compared two distinct character appearances, one human and one robot. In future work, we would like to explore more character variations, including levels of realism on a continuum for the same character, such as in McDonnell et al. [33]. Similarly, we only investigated two levels of voice realism, real and TTS, but it would be interesting to examine levels of realism between the two in order to assess whether ratings of likability and anthropomorphism develop monotonically. For example, it would be interesting to compare the original audio recordings to a real human voice without emphasis or emotional markers, retaining the voice quality but removing other conversational information carried through the speech. We found strong negative effects for *TTS* voice, which may have been emphasized by the conversational nature of the translated speech recordings. Since the original speech was produced spontaneously in an unscripted manner, it was ripe with hesitations and repetitions, common in free-flowing conversation. While this often goes unnoticed in real speech recordings, it can stand out in TTS audio. It was, however, necessary to keep these speech parts for the *TTS* voice in order to preserve the original speech-motion relationship. For scripted speech without hesitations and repetitions, the negative effect of a TTS voice may be mitigated. Here, we only assessed a single state-of-the-art TTS voice, chosen to match the original recordings as best as possible, but the specific choice of TTS voice (and implied gender) can influence perceptions of likability and human-likeness [30].

Our results are limited to one male speaker, and using other speakers with differing gesture style or gender could impact perceptions. In future work, we plan to include a variety of speakers.

In future work, we want to explore how personality perceptions are impacted by agent design. Personality design is a critical factor for virtual agents [25], and has been show to be influenced by the agent's verbal behavior [17, 52], gesture [46], and appearance [33].

ACKNOWLEDGMENTS

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).

REFERENCES

[1] Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[2] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. 2018. The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. In *Proc. Interspeech 2018*. 2863–2867. https://doi.org/10.21437/Interspeech.2018-1093

[3] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger – a critical look at the Uncanny Valley. In *Proc. of Robot and Human Interactive Communication*. 269–276.

[4] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2000. Longman grammar of spoken and written English.

[5] Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2020. Affective synthesis and animation of arm gestures from speech prosody. *Speech Communication* 119 (2020), 1–11.

[6] 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

[7] Wallace Chafe and Jane Danielwicz. 1987. Properties of Spoken and Written Language. Technical Report No. 5. (1987).

[8] Thierry Chaminade, Jessica Hodgins, and Mitsuo Kawato. 2007. Anthropomorphism influences perception of computer-animated characters' actions. *Social cognitive and affective neuroscience* 2, 3 (2007), 206–216.

[9] Michael J Doughty. 2001. Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry and Vision Science* 78, 10 (2001), 712–725.

[10] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[11] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion Capture: Emotionally Expressive Characters for Games. In *Proc. of Motion on Games*. ACM, 53–60.

[12] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is Believing: Body Motion Dominates in Multisensory Conversations. *ACM Transactions on Graphics (SIGGRAPH)* 29, 4, Article 91 (July 2010), 9 pages. https://doi.org/10.1145/1778765.1778828

[13] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* 89 (2020), 117–130.

[14] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[15] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive Gesture Generation from Speech through Database Matching. *Computer Animation and Virtual Worlds* 32 (2021).

[16] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. It's A Match! Gesture Generation Using Expressive Parameter Matching. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*. 151–158.

[17] Alastair Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*. 40–48.

[18] Alex Gilmore. 2004. A comparison of textbook and authentic interactions. *ELT journal* 58, 4 (2004), 363–374.

[19] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[20] Haohan Guo, Shaofei Zhang, Frank K Soong, Lei He, and Lei Xie. 2021. Conversational End-to-End TTS for Voice Agents. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 403–409.

[21] David Hanson, Andrew Olney, Steve Prilliman, Eric Mathews, Marge Zielke, Derek Hammons, Raul Fernandez, and Harry Stephanou. 2005. Upending the uncanny valley. In *AAAI*, Vol. 5. 1728–1729.

[22] Chin-Chang Ho and Karl F. MacDorman. 2010. Revisiting the Uncanny Valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518.

[23] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*. 57–64.

[24] Chien-Ming Huang and Bilge Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–64.

[25] Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies* 53, 2 (2000), 251–267.

[26] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[27] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can We Trust Online Crowdworkers? Comparing Online and Offline Participants in a Preference Test of 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 30, 8 pages.

[28] Adam Kendon. 1972. Some relationships between body motion and speech. *Studies in dyadic communication* 7, 177 (1972), 90.

[29] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. (2020).

[30] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics* 14 (2020), 105. https://doi.org/10.3389/fnbot.2020.593732

[31] Karl F. MacDorman, Robert D. Green, Chin-Chang Ho, and Clinton T. Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior* 25, 3 (2009), 695–710.

[32] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35.

[33] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. 2012. Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–11.

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

[35] Izidor Mlakar, Zdravko Kačič, and Matej Rojc. 2013. TTS-driven synthetic behaviour-generation model for artificial bodies. *International Journal of Advanced Robotic Systems* 10, 10 (2013), 344.

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

[37] Michael Neff. 2016. Hand gesture synthesis for conversational characters. *Handbook of Human Motion* (2016), 1–12.

[38] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[39] Jan Ondras, Oya Celiktutan, Paul Bremner, and Hatice Gunes. 2020. Audio-driven robot upper-body motion synthesis. *IEEE transactions on cybernetics* (2020).

[40] Jan Ondřej, Cathy Ennis, Niamh A Merriman, and Carol O'sullivan. 2016. Franken-Folk: Distinctiveness and attractiveness of voice and motion. *ACM Transactions on Applied Perception (TAP)* 13, 4 (2016), 1–13.

[41] 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*. 1010–1018.

[42] Lukasz Piwek, Lawrie S McKay, and Frank E Pollick. 2014. Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition* 130, 3 (2014), 271–277.

[43] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.

[44] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In *International conference on social robotics*. Springer, 31–41.

[45] Ari Shapiro, Petros Faloutsos, and Victor Ng-Thow-Hing. 2005. Dynamic animation and control environment. In *Proceedings of graphics interface 2005*. Canadian Human-Computer Communications Society, 61–70.

[46] Harrison Jesse Smith and Michael Neff. 2017. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[47] Shiva Sundaram and Shrikanth Narayanan. 2003. An empirical text transformation method for spontaneous speech synthesizers. In *Eighth European Conference on Speech Communication and Technology*.

[48] Ann K Syrdal, Alistair Conkie, Yeon-Jun Kim, and Mark C Beutnagel. 2010. Speech acts and dialog TTS. In *Seventh ISCA Workshop on Speech Synthesis*.

[49] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019. Spontaneous conversational speech synthesis from found data. In *Interspeech*.

[50] James C Thompson, J Gregory Trafton, and Patrick McKnight. 2011. The perception of humanness from the movements of synthetic agents. *Perception* 40, 6 (2011), 695–704.

[51] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The Effect of Multimodal Emotional Expression and Agent Appearance on Trust in Human-Agent Interaction. In *Motion, Interaction and Games*. 1–6.

[52] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society*. 1–6.

[53] Burcu A Urgen, Marta Kutas, and Ayse P Saygin. 2018. Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia* 114 (2018), 181–185.

[54] Lawrence Weschler. 2002. Why Is This Man Smiling? *Wired* 10 (2002). http://www.wired.com/wired/archive/10.06/face.html

[55] Yanzhe Yang, Jimei Yang, and Jessica Hodgins. 2020. Statistics-based Motion Synthesis for Social Conversations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 201–212.

[56] Yanzhe Yang, Jimei Yang, and Jessica Hodgins. 2020. Statistics-based Motion Synthesis for Social Conversations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 201–212.

[57] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[58] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.

[59] Unai Zabala, Igor Rodriguez, José María Martínez-Otzeta, Itziar Irigoien, and Elena Lazkano. 2021. Quantitative analysis of robot gesticulation behavior. *Autonomous Robots* (2021), 1–15.