Embodiment and gender interact in alignment to TTS voices

Michelle Cohn (mdcohn@ucdavis.edu)
Department of Linguistics, Phonetics Lab, UC Davis, 1 Shields Avenue
Davis, CA 95616 USA

Patrik Jonell (pjjonell@kth.se)
Division of Speech, Music, & Hearing, KTH Royal Institute of Technology, SE-100 44
Stockholm, Sweden

Taylor Kim (tsekim@ucdavis.edu)
Department of Linguistics, Phonetics Lab, UC Davis, 1 Shields Avenue
Davis, CA 95616 USA

Jonas Beskow (beskow@kth.se)
Division of Speech, Music, & Hearing, KTH Royal Institute of Technology, SE-100 44
Stockholm, Sweden

Georgia Zellou (gzellou@ucdavis.edu)
Department of Linguistics, Phonetics Lab, UC Davis, 1 Shields Avenue
Davis, CA 95616 USA

Abstract

The current study tests subjects' vocal alignment toward female and male text-to-speech (TTS) voices presented via three systems: Amazon Echo, Nao, and Furhat. These systems vary in their physical form, ranging from a cylindrical speaker (Echo), to a small robot (Nao), to a human-like robot bust (Furhat). We test whether this cline of personification (cylinder < mini robot < human-like robot bust) predicts patterns of gender-mediated vocal alignment. In addition to comparing multiple systems, this study addresses a confound in many prior vocal alignment studies by using identical voices across the systems. Results show evidence for a cline of personification toward female TTS voices by female shadowers (Echo < Nao < Furhat) and a more categorical effect of device personification for male TTS voices by male shadowers (Echo < Nao, Furhat). These findings are discussed in terms of their implications for models of device-human interaction and theories of computer personification.

Keywords: vocal alignment; embodiment; human-device interaction; gender; text-to-speech

Introduction

Recent advancements in robotics and conversational AI has led to the development of more human-like robotic systems, such as those with expressive facial movements (e.g., Sophia by Hanson Robotics) and speech synthesis systems that yield hyper-naturalistic voices (e.g., Amazon Echo). The presence of and variation across these different systems allows for an empirical test of aspects of computer personification theories, such as the *Computers are Social Actors* (CASA) framework which proposes that humans apply the social behavior norms from human-human interaction to their interactions with technology when they detect a cue of humanity in a digital system (e.g., Nass, Steuer, and Tauber, 1994). Aspects of CASA have received support across many empirical studies, such as showing that people apply politeness norms to computer interlocutors (Nass et al., 1997). Yet, most studies do not directly compare human-computer and human-human interaction (e.g., Bell et al., 2003; Nass et al., 1999, 1994). Furthermore, no prior studies, to our knowledge, have tested the extent to which people's application of human-based social responses might be gradient. The current study was designed to fill this gap in the literature by investigating whether we see differences in application of social behavior from human-human interaction across systems that vary *gradiently* in apparent humanness.

Given that the main type of interaction with modern voice-activated artificially intelligent (voice-AI) devices is through speech, a relevant social behavior to examine is vocal alignment: when speakers adjust their pronunciations of words to more closely mirror their interlocutors' speech patterns. Greater degree of alignment has been argued to signal social closeness between interlocutors; one theory of human-human linguistic coordination is *Communication Accommodation Theory* (CAT) (Giles et al., 1991; Shepard et al., 2001): where speakers use degree of convergence to convey their closeness to an interlocutor – or, conversely, their divergence to signal greater social distance. For example, people align more to interlocutors if they find them attractive (Babel, 2012) and likeable (Chartrand & Bargh, 1996).

Some prior work has explored whether alignment patterns differ for non-human interlocutors, comparing human-human and human-computer interaction (for a review, see Branigan et al., 2010). For example, Branigan and colleagues (2003) found that participants aligned in syntactic structure (e.g., "give the dog a bone" vs. "give a bone to the dog") to the same extent in typed interactions between an apparent 'computer' and 'human' interlocutor. Yet, in spoken language interaction, differences by interlocutor appear to be more pronounced: three recent studies found that people vocally align to both human and voice-AI assistants (Apple's Siri, Amazon's Alexa), but display less alignment to the assistant voices (Cohn et al., 2019; Raveh et al., 2019; Snyder et al., 2019). These findings suggest that our transfer of social behaviors to AI systems in speech interactions is tempered by their social category as not human. This differentiation of speech behavior based on humanness is in line with the theory of *Audience Design* (Bell, 1984; Clark & Murphy, 1982): whereby interlocutors strategically adapt their productions for the communicative needs of their listener. Combining aspects of *Audience Design* and *CASA* (Nass et al., 1997, 1994), we hypothesize that people's speech behavior toward voice-AI will vary gradiently as a function of their personification of the system. We predict that people will treat more naturalistic systems more like they would a real human, while less human-like systems will receive less human-based socially-mediated behaviors.

The present study tests this hypothesis – gradient application of social behaviors based on personification – by varying the physical embodiment of voice-AI systems. Current devices vary in how they embody humanness. For example, cylindrical smart speakers are now common household voice-AI systems (e.g., Amazon Echo; Google Home). Other types of voice-AI systems take on more human-like forms. For example, the Nao robot has a head, face, and body, but with clear physical and mechanical characteristics that make it distinct from a real human (see Figure 1). Related work has suggested that the Nao could be considered an intermediate type of robot along a cline of human-likeness: in a study by Brink and colleagues (2019), they found that participants found the Nao less uncanny than a more human-like robot face. In the present study, we consider a cline of personification from a smart speaker to a Nao robot to a Furhat robot (Al Moubayed et al., 2012), another type of robot that is more human-like (see Figure 1). The Furhat resembles a human bust, with a 3D printed face, and a back-projected video of a human face. These videos increase its realism: the eyes blink and make micro-movements, and the mouth shows appropriate articulation of speech sounds to match the audio.

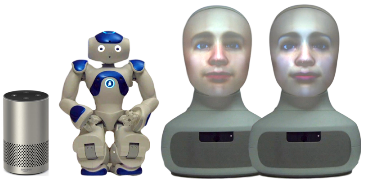

Figure 1: Systems used in the present study (L-R): Amazon Echo, Nao, Furhat (male), Furhat (female).

These three devices – a cylindrical speaker, mini-robot, and naturalistic bust – vary along a continuum of humanness in terms of embodying a human form and displaying human-like features. Our AI personification hypothesis is that simply varying the humanness of the device will lead to changes in vocal alignment toward the system. Identical stimuli recordings will be presented across systems to avoid any confound that might arise from using different voices. In this case, we expect an increasing degree of alignment, signaling greater application of this human-based, socially-mediated behavior, as the personification of the device increases: greatest alignment toward the Furhat device, less toward the Nao, and least toward the Echo speaker.

An alternative hypothesis is that increasing personification of AI may lead to less alignment – or even divergence – from the speech produced by the most human-like system (e.g., Furhat) as a consequence of the *Uncanny Valley* effect (Mori et al., 2012): as non-human, robotic entities display greater human-like characteristics, there is a tendency for people to assess them more positively. Yet, there is a point at which there is a steep drop-off and likeability plunges, a function known as the 'uncanny valley'. An example of this is the response to seeing a nearly human-like face in a non-human device, triggering feelings of disgust or uneasiness. Speakers' patterns of vocal alignment are one way to test the uncanny valley; prior work has shown that speakers show more convergence toward an interlocutor they feel socially close with, while they show divergence from those they want to distance themselves from socially. If a human-like voice is paired with a hyper-naturalistic robotic entity, this might trigger an uncanny valley-like effect, causing participants to align less than they would for a real human.

Gender-mediated alignment toward AI?

In addition to social factors, such as likeability, human-human vocal alignment has been shown to be mediated by gender. For example, participants show stronger vocal alignment toward human male voices than female voices (Pardo, 2006). However, this gender effect is sometimes mixed (Pardo et al., 2017), suggesting that idiosyncratic properties of voices can influence the degree of alignment as well. Nevertheless, there is some evidence that gender-mediated alignment patterns may also transfer to human-device interaction: humans display greater alignment toward male, relative to female, voices for both human and Apple Siri model talkers (Cohn et al., 2019; Snyder et al., 2019). This supports the hypothesis that humans transfer gender-mediated patterns of vocal alignment from human-human conversations to their interactions with voice-AI systems, supporting predictions made by CASA (Nass et al., 1994). Yet, the properties of the voices themselves (e.g., idiosyncrasies of the human speakers, TTS synthesis) pose a confound between apparent human-likeness and degree of alignment seen in prior studies.

Based on our proposal that people's vocal alignment behavior toward AI will vary as a function of the personification of the system, we can explore more specific predictions by varying the apparent gender of the voice. Prior work reports that male voices are imitated to a greater degree than female voices, which is realized to a lesser extent for TTS voices (Cohn et al., 2019; Snyder et al., 2019). Thus, we predict that this gender-mediated pattern will vary gradiently as a function of the personification of the AI. More specifically, we predict gender-mediated patterns of alignment to be realized to the largest extent for the AI system with the most human-like physical features (Furhat) and the least amount of gender-mediated alignment patterns for the Echo speaker, with the Nao receiving alignment patterns in-between the others.

Current Study

The present study examines 1) the influence of degree of human-likeness on extent of vocal alignment toward voice-AI interlocutors, and 2) how AI personification interacts with apparent gender on vocal alignment patterns. We conducted a shadowing experiment for identical sound files produced by two TTS voices presented across three embodied robotic systems: a Furhat (Al Moubayed et al., 2012), a Nao robot (SoftBank Robotics), and an Amazon Echo. In doing so, we address a limitation of many vocal alignment studies, where comparisons are made across a small subset of different model talkers, leading to mixed, and often conflicting findings about the influence of gender on alignment in the literature (cf. Pardo et al., 2017), allowing us to specifically test for the role of system personification, while holding the voice characteristics constant across model talkers. Furthermore, using these three systems also serves as a stronger cue that these are indeed separate interlocutors. The current study consists of two experiments: the first is a single word shadowing paradigm, where participants were first asked to record baseline productions of words and then asked to repeat (to shadow) words produced by the systems. Experiment 2 is an AXB similarity rating task where a separate group of listeners rate the speakers' baseline and shadowed productions from Experiment 1, providing a holistic assessment of vocal alignment (cf. Cohn et al., 2019).

Experiment 1. Shadowing

Methods

Subjects. Subjects were 10 native English speakers (mean age = 35.1 ± 8.5 years old; 5 female, 5 male). Six participants reported prior use of one or more voice-AI systems (e.g., Amazon's Alexa, Apple's Siri, Google Assistant, etc.); four reported no prior interaction with any voice-AI system. Participants received a $15 Amazon gift card for their time.

Stimuli. Twelve target CVN words were selected for the current study, taken from related studies of phonetic alignment by talker gender and humanness (Cohn et al., 2019; Snyder et al., 2019): *bomb, chime, hem, pun, sewn, shone, shun, tame, vine, wane, wren, yawn*. The 12 target words were generated with two Amazon Polly TTS voices (US-English): a male voice ("Matthew") and a female voice ("Salli"). For the Furhat talkers, two face "textures" were selected (male texture: "Marty", female texture: "Fedora"). These faces were selected as they were the most human-like available (see Figure 1). For each of the 6 gender/system pairings, we generated instructions where each model talker introduced themselves with a different gender-matching name (e.g., 6 different apparent speakers: "Rebecca", "Matthew", "Mary", "Michael", etc.).

Procedure. Subjects completed the experiment in a semi-soundproof room with a head-mounted microphone. First, a pre-exposure production of the words was recorded from each of the subjects in order to get their baseline speech patterns prior to exposure to the model talkers. Participants produced each of the 12 target words (repeated 2 times), reading from a pseudo-randomized list.

Next, participants completed the word-shadowing portion of the study with the Amazon Echo, Nao, and Furhat (order counterbalanced across subjects). The same experiment was designed on all three systems, using the Amazon Alexa Skills Kit, Nao Choregraphe, and Furhat Blockly, respectively. For each interlocutor (Echo, Nao, Furhat), subjects completed two blocks: a male and female speaker (gender ordering was counterbalanced across subjects). For each subject, the voice gender ordering (e.g., M-F, M-F, M-F) was consistent across the interlocutors; this was to avoid consecutively presenting an identical voice for two different interlocutors (e.g., male Furhat, male Echo). On a given trial, subjects heard the system produce a target word (e.g., "wren") followed by a 3000ms silence, providing the subject time to respond. Note that the systems' responses were not contingent on the subjects' productions to avoid ASR errors.

Finally, subjects completed a short ratings survey about each talker (randomly presented). For each, they saw the "name" of the talker, a picture of the face/system, and an example word recording. Using a sliding scale, they rated each voice on four dimensions: age, friendliness (0=not friendly, 100=extremely friendly), human-likeness (0=machine-like, 100=human-like), and interactiveness (0=inert, 100=extremely interactive.)

Ratings Analysis & Results

We analyzed participants' ratings of the talkers with separate mixed effects linear regressions, with main effects of Model Talker System (a 3-level predictor: Echo, Nao, Furhat) and Model Talker Gender (a two-level predictor: Female, Male), and their interaction, and by-Subject random intercepts. Average ratings for the male and female TTS voices across the systems are plotted in Figure 2.

First, we observe differences of Model Talker Gender for age rating: female voices were rated as being younger than male voices [β=-7.27, t=-9.13, p<0.001]. Additionally, there was a top-down effect of Model Talker System: voices presented through the Furhat were rated as being younger [β=-4.03, t=-3.58, p<0.001]. Yet, this is driven by an interaction between Model Talker Gender and System, where the male voice was rated as being younger when presented through the Furhat device, compared to when it was presented through the other systems [β=-3.27, t=2.90, p<0.01]. For friendliness ratings of the voices, the model showed only a main effect of Model Talker System: voices presented through the Furhat were rated as being friendlier

[β=6.53, t=2.97, p<0.01]. For friendliness ratings of the voices, the model showed only a main effect of Model Talker System: voices presented through the Furhat were rated as being friendlier [β=6.53, t=2.97, p<0.01]. For ratings of human-likeness and interactiveness of the voices (bottom two panels), there were no significant differences by the Model Talker System or Model Talker Gender.

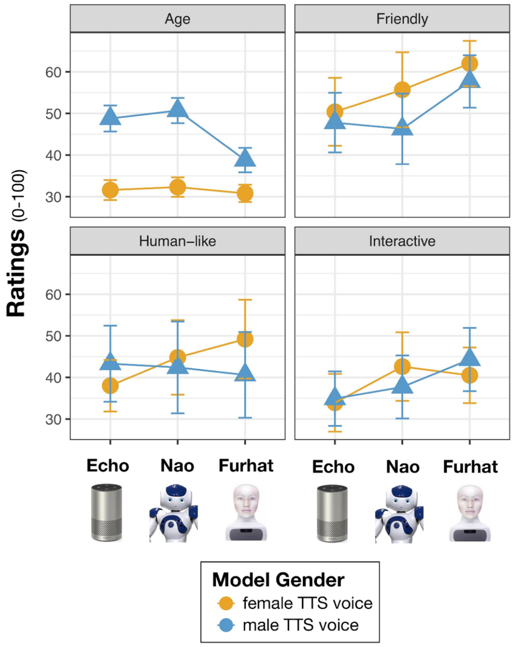

Figure 2: Mean ratings of age, friendliness, human-likeness and interactiveness of the TTS voices when presented across systems (Echo, Nao, Furhat). Error bars show standard error of the mean.

Experiment 2. AXB Similarity

In this experiment, we assessed global similarity between the participants' baseline productions of the words (produced at the beginning of the experiment, prior to exposure to the model talkers) and their shadowed productions for each model talker from Experiment 1 with an AXB similarity ratings task (Cohn et al., 2019).

Methods

Subjects. 51 native English speakers participated in the AXB study. Subjects were recruited through a university Psychology subjects' pool (37 females, 14 males; mean age = 19.9 ± 1.7 years old). All subjects received course credit for their participation.

Stimuli. The stimuli consisted of a baseline and shadowed production by the 10 speakers who completed Experiment 1. For each speaker, we selected one of their pre-exposure and shadowed productions of each word for each of the six model talkers (i.e., Furhat female, Furhat male, Echo female, etc.). Due to speakers' confusions about the TTS production of 'yawn', and speaker mispronunciations for several other words, we only had a full set of pre-exposure and correct shadowed productions from each model talker of 8 words for the AXB study: *bomb, chime, hem, pun, shun, tame, wane, wren.*

Procedure. Participants completed the AXB similarity ratings experiment in a sound-attenuated booth, wearing headphones (Seinheiser Pro) and sitting in front of a computer screen and button box. On a given trial, raters heard three words separated by a short silence (ISI =1s): a speaker's production of a word at baseline (e.g., "A"), the model talker's production of that same word ("X"), and the speaker's shadowed production of that word for that model talker (e.g., "B"). Their task was to select the speaker's token that sounded most similar to "X" (i.e., the model talker). Order of pre-exposure and shadowed token (i.e., "A" and "B") was balanced within each subject and counterbalanced across both system and interlocutor gender. In total, raters completed 480 AXB similarity ratings (10 speakers x 8 words x 3 systems x 2 genders). Trials were presented in four blocks of 120 trials; after each block, subjects could take a short break. In total, the experiment lasted roughly 45 minutes.

Analysis. We coded whether the raters selected the shadowed token as more "similar" to the model talker (=1) or not (=0) and analyzed their responses with a mixed effects logistic regression (glmer). Fixed effects included the Model Talker System (3 levels: Echo, Nao, Furhat), the Model Talker Gender (2 levels: female, male), and the Shadower Gender (2 levels: female, male), and the interaction between them. Random effects structure included by-Shadower random intercepts and by-Shadower random slopes for Model Talker System, and by-Rater and by-Word random intercepts.

Results

The mean AXB similarity ratings for each of the three systems and two TTS voices is displayed in Figure 3. Overall, the model computed several main effects and interactions. First, there was a main effect of Model Talker Gender: shadowers showed significantly less alignment to the female TTS voice (in orange, Figure 3) than the male TTS voice (in blue, Figure 3) [β=-0.01, t=-4.8, p<0.001].

While there was a trend toward significance for less alignment toward the Echo (p=0.054), there was not a main effect of Model Talker System. Similarly, there was not a main effect of Shadower Gender. However, Model Talker System and Model Talker Gender did participate in several significant interactions. First, we observed greater alignment toward Furhat by female shadowers [β=0.01, t=3.08, p<0.001]. Yet, this effect was mediated by a three-way interaction: female shadowers imitated the female Furhat more [β=0.02, t=4.29, p<0.001] (see Figure 3, left panel).

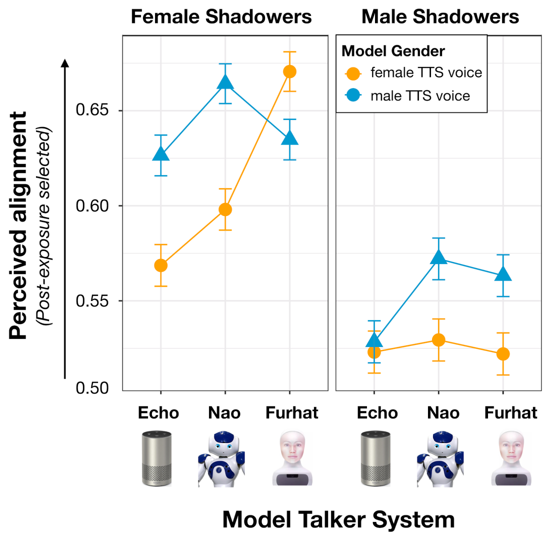

Figure 3: Mean ratings of perceived degree of vocal alignment in the AXB similarity ratings task for the three systems (Echo, Nao, Furhat) and two TTS voices. Error bars show standard error of the mean.

There was also a three-way interaction for the Echo: female shadowers showed less alignment toward the Echo with the female TTS voice than to the male TTS voice [β=-0.01, t=-3.03, p<0.001]. The releveled model (ref = Echo) showed only a two-way interaction for the Nao: both male and female shadowers aligned to the male Nao more than the female Nao [β=1.24e-02, t=2.85, p<0.01]. The releveled model for gender (ref = female) showed that males aligned to the male Furhat more [β=1.92e-02, t=4.39, p<0.001] and to the male Echo less [β=-1.32e-02, t=-3.03, p<0.01] (in blue, Figure 3). A post-hoc analysis on the data for male shadowers/male model talkers confirmed no significant difference for the Nao and Furhat, but both showed greater alignment than the Echo [β=-2.61e-02, t=-2.96, p<0.01].

Post-hoc Analysis: Alignment and Ratings

We additionally conducted post-hoc analyses to test whether participants' ratings of the model talkers (e.g., age, friendliness, human-likeness, interactiveness) mediated their alignment patterns. The four ratings were included as additional independent variables in separate logistic regression models run on similarity ratings responses, with identical fixed and random effects structure: Model Talker System*Model Talker Gender*ShadowerGender*Rating + (1 + Model Talker System | Shadower) + (1|Rater) + (1|Word).

Results

None of the models revealed significant main fixed effects for any of the ratings; yet, there were interactions between all ratings and Model Talker System. For age, we found an interaction with age ratings of the model talkers and degree of alignment: participants showed more alignment toward the female Echo when they rated the voice as being older [β=2.45e-03, t=2.16, p<0.05]. Age ratings did not influence alignment patterns toward the Furhat model talkers. The releveled model showed that female shadowers displayed less alignment toward the female Nao voice as ratings of its age increased [β=-1.75e-03 t=-2.04, p<0.05].

Friendliness also interacted with alignment patterns toward particular model talkers: participants showed less alignment toward the Furhat talkers when they were rated as being friendlier [β=-9.8e-04, t=-2.8, p<0.05] and no effect for the Echo talkers. The releveled model showed greater alignment toward the Nao voices as they were rated as being friendlier [β=6.8e-04, t=2.5, p<0.05].

For ratings of human-likeness of the system, several interactions were computed as significant: first, there was less alignment toward the female Furhat if it was rated as being more human-like [β=-3.7e-04, t=-2.1, p<0.05]. Additionally, the model revealed there was less alignment for female shadowers toward the female Echo if it was rated as being more human-like [β=6.0e-04, t=-2.6, p<0.001]. The model, releveled in order to unpack the comparison with the Nao system and the other devices, also revealed that shadowers displayed alignment patterns toward the Nao system as a function of their human-likeness ratings: there was more alignment toward the female Nao if it was rated as being more human-like [β=8.0e-04, t=3.4, p<0.0001].

The model with how interactive-inert the talker was revealed just one interaction: shadowers displayed even less alignment toward the female Echo with less interactive ratings [β=-7.2e-04, t=-3.0 p<0.001]. There was no effect for Furhat talkers. The releveled model showed a different pattern toward the Nao: shadowers displayed *greater* alignment toward the female Nao with increasing interactiveness ratings [β=9.0e-04, t=4.1 p<0.001].

Discussion

This study was designed to test whether patterns of vocal alignment toward male and female TTS voices are realized gradiently, on the basis of the physical form of the device producing the speech, varying from very non-human-like (a cylindrical speaker) to more human-like (a human-shaped bust). In general, participants aligned toward the male TTS voices to a greater extent, in line with gender-mediated patterns observed in prior work on human and Siri voices (e.g., Cohn et al., 2019; Pardo, 2006). That we see applications of this gender-mediated social 'rule' from human-human interaction to human-AI interaction supports predictions made by the CASA framework (Nass et al., 1997, 1994): participants applied gender-mediated patterns to their alignment during interactions with AI systems.

Additionally, we observed gender asymmetries based on both shadower and model talker gender. Female participants, in general, displayed greater alignment to the male TTS voice across systems. These findings parallel those in the human-human literature. For example, in a shadowing experiment with disembodied voices (and no images), female shadowers aligned more to the male talkers, while male shadowers aligned equally toward both male and female voices (Namy et al., 2002). We see these patterns borne out in the present study for the Amazon Echo responses (females aligning to the male TTS voice more; males aligning to both genders equally). Yet, when more social cues are available (e.g., in a more human-like form: Furhat), we observe that alignment may vary based on same- versus mixed-gender shadower/model talker pairs. This suggests that the amount of social information available and characteristics of the participants may shape the degree of alignment more generally.

Furthermore, we found some evidence in support of our proposal of AI personification gradience: degree of vocal alignment increased as degree of personification of the device increased (cylinder < mini-robot < human-like robot) for female shadowers toward the female TTS voice. Male shadowers also showed evidence that personification of the system mediates alignment, but in a more categorical way: Male shadowers aligned more toward the male TTS voice presented in the Nao and the Furhat, the two more pseudo-anthropomorphic systems, relative to the Echo. These results support our hypothesis, that the *degree* to which a device embodies a human-like form, the more people will apply the norms of human-human communication to human-AI dyadic interactions: in this case, alignment. This finding supports CAT (Giles et al., 1991; Shepard et al., 2001), where speakers strategically adapt their degree of convergence toward their interlocutor based on their social relationship. Additionally, our findings are broadly in line with *Audience Design* (Bell, 1984; Clark & Murphy, 1982): speakers adjust their speech differently based on the apparent communicative needs of their interlocutor (here, based on their physical form as, possibly, a cue of more ‘human-like’ competence).

Our proposal of AI personification gradience also receives support from our post-hoc analyses; all four ratings of the interlocutors (age, friendliness, human-likeness, and interactiveness) interacted with the degree of human embodiment of the system to explain vocal alignment patterns. For one, increasing human-likeness ratings of the Nao system led to *increased* alignment; in contrast, increasing ratings of human-likeness led to *decreased* degree of alignment toward the Furhat device. The reversal of the expected pattern of increasing alignment with increasing human-likeness, might be interpreted as an ‘uncanny valley’ effect (Mori et al., 2012), where increasing human-likeness of a non-human entity leads to increasing positive feelings toward the entity until a threshold where it elicits feelings of discomfort and/or disgust. Some participants may have felt a sense of eeriness in seeing a more human-like face realized on a device. Age ratings were also linked to patterns of vocal alignment: participants aligned more to the Echo if they rated the voice as being from an older speaker, but displayed *less* alignment to the Nao if it was rated as being older. This may also be related to the uncanny valley effect, where cue incongruency drives a sense of uneasiness: the Nao has an infant-like form which contrasts with the voice ages (adult TTS parameters) (~30s for the female TTS voice, ~40-50s for male TTS voice). These observations lead us to refine our AI personification hypothesis: people’s application of human-based behavior norms during speech interaction with voice-AI will increase as a function of the personification of the device, until the AI anthropomorphism reaches realism levels that trigger feelings of discomfort. The finding of uncanny valley realized in patterns of vocal alignment is novel and opens up new ways of exploring and investigating behavioral responses to embodied AI.

There are several limitations of the current study that can serve as avenues for future work. For one, differences observed for the Furhat faces may have been driven by those particular images displayed: future studies using additional face textures and having participants rate the attractiveness of the faces can tease apart the contribution of this visual social information. Previous work has reported a link between shadower’s attractiveness ratings of faces and their degree of alignment toward that voice (Babel, 2012), further suggesting this may have played a role.

Additionally, while an advantage of an AXB similarity rating is that we make no a priori assumptions as to which acoustic-phonetic features may be imitated, our overall number of shadowers was limited in order to allow for raters to make similarity judgments on the full set of stimuli (all shadowed tokens, across the three systems; 480 trials, taking roughly 45 minutes). While some groups have split AXB ratings into separate experiments by groups of speakers, the results were less than clear (Pardo et al., 2017). One benefit of the current approach is that the patterns are easily identifiable and comparable for future work (cf. Cohn et al. 2019; Snyder et al., 2019).

Furthermore, one limitation and avenue for future study is the number and variety of TTS voices. While using two Amazon Polly voices allowed us to address a confound in previous work (by using identical voices across the three systems during the shadowing experiment), it may have affected the ratings that the participants provided (age, friendliness, etc.) if they recognized that the same voice was used on each system. This was, in part, mitigated by never presenting the same voice consecutively. Having the speakers shadow a greater variety of TTS voices across different systems could lessen this possible effect.

Finally, subtle differences in the speaker systems between the Echo, Nao, and Furhat may also have contributed to differences in perceived human-likeness. Future work using computer-mediated methods, such as that presenting videos of the three interlocutors, can control more aspects of the interaction (e.g., intensity) and be compared to the present study to assess the degree to which embodied versus computer-mediated interactions may shape vocal alignment.

Overall, this study provides a first step in exploring the nature of AI personification and its relationship with vocal alignment, and sets the groundwork for future research.

References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems* (pp. 114–130). Springer.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189.

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. *Proceedings of ICPHS*, *3*, 833–836.

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 186–191.

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development*, *90*(4), 1202–1214.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*(3), 464.

Clark, H. H., & Murphy, G. L. (1982). Audience Design in Meaning and Reference. In J.-F. Le Ny & W. Kintsch (Eds.), *Advances in Psychology* (Vol. 9, pp. 287–299). North-Holland. https://doi.org/10.1016/S0166-4115(09)60059-5

Cohn, M., Ferenc Segedin, B., & Zellou, G. (2019). Imitating Siri: Socially-mediated alignment to device and human voices. *Proceedings of International Congress of Phonetic Sciences*, 1813–1817.

Giles, H., Coupland, N., & Coupland, I. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, *1*.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems 1. *Journal of Applied Social Psychology*, *29*(5), 1093–1109.

Nass, C., Moon, Y., Morkes, J., Kim, E.-Y., & Fogg, B. J. (1997). Computers are social actors: A review of current research. *Human Values and the Design of Computer Technology*, *72*, 137–162.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00559

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659.

Raveh, E., Siegert, I., Steiner, I., Gessinger, I., & Möbius, B. (2019). Three'sa Crowd? Effects of a Second Human on Vocal Accommodation with a Voice Assistant. *Proc. Interspeech 2019*, 4005–4009.

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. In *The new handbook of language and social psychology* (W. P. Robinson, H. Gile, pp. 33–56). John Wiley & Sons, Ltd.

Snyder, C., Cohn, M., & Zellou, G. (2019). Individual variation in cognitive processing style predicts differences in phonetic imitation of device and human voices. *Proceedings of the Annual Conference of the International Speech Communication Association*, 116–120.