Top-down effect of apparent humanness on vocal alignment toward human and device interlocutors

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

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

Abstract

Humans are now regularly speaking to voice-activated artificially intelligent (voice-AI) assistants. Yet, our understanding of the cognitive mechanisms at play during speech interactions with a voice-AI, relative to a real human, interlocutor is an understudied area of research. The present study tests whether top-down guise of "apparent humanness" affects vocal alignment patterns to human and text-to-speech (TTS) voices. In a between-subjects design, participants heard either 4 naturally-produced or 4 TTS voices. Apparent humanness guise varied within-subject. Speaker guise was manipulated via a top-down label with images, either of two pictures of voice-AI systems (Amazon Echos) or two human talkers. Vocal alignment in vowel duration revealed top-down effects of apparent humanness guise: participants showed greater alignment to TTS voices when presented with a device guise ("authentic guise"), but lower alignment in the two inauthentic guises. Results suggest a dynamic interplay of bottom-up and top-down factors in human and voice-AI interaction.

Keywords: vocal alignment; apparent guise; voice-activated artificially intelligent (voice-AI) systems; human-computer interaction

Introduction

Humans use speech as a way of conveying our abstract thoughts and intentions. Yet, speech is more than just a signal to emit words and phrases; our productions are shaped by intricate cognitive and social processes underlying and unfolding over the interaction. One large mediator of these processes is *who* we are talking to: their social characteristics (e.g., gender in Babel, 2012), humanness (e.g., computer or human in Burnham et al., 2010), and even our attitudes toward our interlocutors (e.g., speech accommodation in Chakrani, 2015) can shape our productions. Examining speech behavior can serve as a window into cognitive-social dimensions of communication and is particularly relevant for examining different types of interlocutors.

Recently, humans have begun interacting with voice-activated artificially intelligent (voice-AI) assistants, such as Amazon's Alexa and Apple's Siri. For example, over the past several years, tens of millions of "smart speakers" have been brought into people's homes all over the world (Bentley et al., 2018). Current voice-AI technology has advanced dramatically in recent years; common systems can now generate highly naturalistic speech in a productive way and respond to spontaneous verbal questions and commands by users. In some speech communities, voice-AI systems are omni-present and used on a daily basis to perform a variety of tasks, such as send and read text messages, make phone calls, answer queries, set timers and reminders, and control internet-enabled devices around the home (Hoy, 2018). Despite the prevalence of voice-AI, our scientific understanding of the socio-cognitive mechanisms shaping human-device interaction is still limited. Specifically, as humans engage in dyadic speech behaviors with these non-human entities, how are our speech behaviors toward them similar or dissimilar from how we talk with humans? In this paper, we examine the effect of apparent "humanness" category — i.e., top-down information that the interlocutor is a device or human — on people's vocal alignment toward text-to-speech (TTS) synthesized and naturally-produced voices.

Computer personification

On the one hand, classic theoretical frameworks of technology personification, such as the "Computers are social actors" (CASA) theory (Nass et al., 1997), hypothesize that our attitudes and behavior toward technological agents mirror those toward real human interactors. This has been tested by examining whether social patterns of behavior observed in human-human interaction apply when people interact with a computer. For instance, people's responses to questions vary based on the social characteristics of the interviewer; for example, biases in responses are observed based on the gender and ethnicity of the (human) interviewer (Athey et al., 1960; Hutchinson & Wegge, 1991). This phenomenon has been described as a social desirability effect: people seek to align their responses to the perceived preferences of their interviewer because to do otherwise would be impolite (Finkel et al., 1991). Nass and colleagues (1999) examined how people apply this politeness pattern to interactive machines. After performing a task (either text- or voice-based) with a computer system, consisting of a tutoring session by the computer on facts about culture, and then subsequently being tested on those facts, participants were asked questions about the performance of the computer. There were two critical conditions: either the same computer that performed the tutoring asked for an evaluation of its performance as a tutor, or the evaluation of that computer was solicited by a different computer located in another room. Nass and colleagues found that people provided more positive evaluations in the same-computer condition, relative to the different-computer condition, indicating that the social desirability effect applies when we interact with a machine. From empirical observations such as this, the CASA framework argues that our interactions with computers include a social component, positing that social responses to computers are automatic and unconscious especially when they display cues associated with being human, in particular "the use of **language**" (Nass et al., 1999, p. 1105, emphasis ours).

Yet, the extent to which individuals display differences in their socio-cognitive representations of real human versus computer interlocutors has resulted in mixed findings in the literature. For example, studies manipulating interlocutor guise – as human or computer – are one way to compare these interactions while holding features of the interaction, such as error rate and conversational context, constant. For example, Wizard-of-Oz studies, where a human experimenter is behind different interactions that are apparently with either a "computer system" or a "real person", demonstrate that people do produce different speech patterns toward humans and technological systems. In particular, Burnham and colleagues (2010) found increased vowel space hyperarticulation and slower productions in speech toward an apparent computer avatar, relative to apparent human interlocutor; this suggests that people have distinct representations for the two interlocutor types, computer versus human. Others have observed greater overlap: Heyselaar and colleagues (2017) found similar priming effects for real humans and human-like avatars (presented in virtual reality, VR); yet, they additionally found less priming for the computer avatar in general. Together, these results suggest that people's cognitive representations for humans and computerized interlocutors appear to be different and, further, they may be gradiently, rather than categorically, distinct.

Manipulating top-down "humanness" guise

In many studies designed to compare of human-human and human-computer interactions, there are multiple features that co-vary: the computer interlocutor has both a different form (e.g., digital avatar in Burnham et al., 2010) and a synthetic voice. This has been true for recent work exploring human-voice-AI interaction as well, where naturally produced and TTS voices are confounded with "apparent humanness" (Cohn et al., 2019; Snyder et al., 2019). One way to probe whether people have a distinct social category for voice-AI is to observe their speech behavior toward a set of voices of the same type while varying the top-down label, either device or human, provided with each voice. Manipulating apparent humanness can speak to the impact of both bottom-up (acoustic) and top-down (guise) factors on speech behavior toward humans and voice-AI: if it is driven by the characteristics of the speech (e.g., naturally produced vs. TTS) and/or by the extent to which we "believe" we are interacting with a human or a device voice. At the same time, presenting an inauthentic top-down label while presenting the original audio (e.g., "human" label with a TTS voice, and vice versa) results in cue incongruency. In the present study, we test whether a match or *mismatch* in voice (real human vs. TTS) and image (human vs. device) shape speech behavior — and whether speakers are more sensitive to incongruous cues by guise: apparent human vs. apparent device talker.

While prior work has examined using moving computer avatars (e.g., Burnham et al., 2010; Heyselaar et al., 2017), most modern voice-AI systems lack an avatar representation (i.e., "Siri" and "Alexa" have no faces). As such, we ask whether presenting a guise via images (either of a device or a human face) can trigger differences in speech behavior toward AI. Prior matched guise experiments have found that participants' perception of a given voice shifts according to minimal social information available. For example, seeing a photo depicting speakers of various ages and socio-economic statuses impacts how listeners categorize the same set of vowels (Hay et al., 2006). Thus, in the present study we predict that top-down influences (here, images of voice-AI devices and human faces), will impact how participants respond to identical stimuli. In particular, we hypothesize that participants will display different speech patterns toward the voices labeled as "devices" compared to "humans".

Vocal alignment

In the current study, to test these questions, we examine vocal alignment, a subconscious, yet pervasive, speech behavior in human interactions, while varying top-down guise ("human" or "device"). Vocal alignment is the phenomenon whereby an individual adopts the subtle acoustic properties of their interlocutor's speech during verbal exchanges. Vocal alignment appears to be, to some extent, an automatic behavior, argued by some to stem from the human tendency to learn language in part by modeling their speech patterns based on those they have experienced from others (Delvaux & Soquet, 2007; Meltzoff & Moore, 1989). Yet, vocal alignment patterns also appear to be mediated by factors in the linguistic and social context, indicating that it is more than just an automatic behavior (Babel, 2012; Nielsen, 2011; Zellou et al., 2016). More specifically, we explore how human interlocutors' vocal behavior reveals the *social* role they assign to apparent voice-AI interlocutors, relative to apparent human interlocutors.

Vocal alignment has been posited to serve pro-social goals and motivations (e.g., Babel, 2012). In particular, "Communication Accommodation Theory" (CAT, Shepard et al., 2001) proposes that speech convergence is used by speakers to foster social closeness with their interlocutor, increasing rapport (Babel, 2012; Pardo, 2006). One question is whether the social distance between humans and AI will be comparable to that between two humans. A CASA account might predict the same application of vocal alignment behavior to humans and to technological systems that engage with participants using language. Indeed, there is some support for alignment in human-computer interaction (Branigan et al., 2010, 2011; Cowan et al., 2015). For example, Branigan and colleagues (2003) found that humans display syntactic alignment toward apparent computer and human interlocutors, but with greater syntactic alignment toward the computer, than human, guise.

There is some work showing vocal alignment toward computers/voice-AI as well. For example, Bell and colleagues (2003) found that participants vocally aligned to the speech rate produced by a computer avatar they interacted with. At the same time, multiple studies have found that individuals tend to show *less* vocal alignment toward modern voice-AI systems than toward human voices (Cohn et al., 2019; Raveh et al., 2019; Snyder et al., 2019). For example, Snyder et al. (2019) found that participants showed greater vowel duration alignment toward human voices, relative to Apple's Siri voices. Furthermore, these vowel-durational patterns mirror those reported in a perceptual similarity ratings task of alignment: less overall vocal alignment toward voice-AI (relative to human voices) overall (Cohn et al., 2019). This suggests that rate/durational cues could be used as a window into the cognitive/social dynamics in vocal alignment toward human and voice-AI interlocutors.

Taken together, these findings suggest that our linguistic interactions with devices/computers are distinct from how we talk to human interlocutors — in some cases triggering less alignment toward voice-AI (vocal alignment) and in other cases more alignment toward computers (lexical, syntactic alignment). In other words, contra CASA, one possibility is that AI actors hold a social status that is distinct from that of humans and subtle differences in our behavior toward them reflect this.

Current Study

The current study was designed to investigate a specific research question: *What is the effect of apparent humanness on vocal alignment patterns toward human and voice-AI interlocutors*? Manipulating apparent social information about the speaker has not, to our knowledge, been used in vocal alignment paradigms. To that end, in the current study, participants completed a word shadowing paradigm (Babel, 2012) consisting of four distinct model talkers. Across two between-subjects conditions, the 4 model talkers were either all real human voices or all TTS voices (Amazon Polly TTS voices). These TTS voices were developed to be used in Amazon Alexa Skills (e.g., interactive apps through Alexa-enabled devices). In each of these conditions, participants were told that two of the model talkers were humans and two of the talkers were devices. By crossing the actual status of the voices (*Voice Type*: human or TTS) with the top-down apparent guise label (*Apparent Guise*: human or device), we aim to probe the cognitive and social role of voice-AI in human-AI verbal interactions (relative to human-human interactions). There are several possible outcomes of this experimental design, each of which can speak to the status of voice-AI in speech interactions.

One hypothesis is that differences in patterns of vocal alignment will be driven by low-level acoustic differences across naturally produced and TTS voices, such as that used in voice-AI systems. In other words, there are inherent acoustic differences in synthetic and naturally produced speech which may lead to different alignment patterns toward the voices. While advances in TTS have created more naturalistic human-sounding voices, TTS voices used in modern voice-AI systems (e.g., Amazon's Alexa, Apple's Siri) are still perceptibly distinct from natural voices, most likely since they contain less variability than human voices for some acoustic features. For example, some have pointed to prosodic irregularities in TTS productions as a marker of a synthetic voice (Németh et al., 2007). If it is the case that bottom-up acoustic properties of synthetic speech, even for more realistic, modern voice-AI TTS, drive the way humans interact with interlocutors, regardless of their apparent status as humans or devices, then we predict that apparent guise will not affect listeners' behavior. In the design of our current study, this would lead to observe only a main effect of Voice Type.

A second hypothesis is that we will observe asymmetries in how apparent humanness guises affect vocal alignment patterns of TTS and human voices. This would be expected if there are both bottom-up (acoustic-level) and top-down (humanness category) influences on how people interact with human and device interlocutors. In other words, if listeners are sensitive to the acoustic-phonetic differences between synthetic and naturally produced speech and if the top-down labels of device and human speaker lead listeners to apply different expectations and social rules to the interaction, we predict asymmetrical influences of the guises "device" and "human" on synthetic and naturally produced voices. Support for this possibility comes from observations of the "uncanny valley" phenomenon: when an entity that humans know is not a natural human takes on enough similarity to human-like features that it elicits feelings of unease or discomfort (Mori et al., 2012). One interpretation of this phenomenon is that it is due to cue incongruence (Moore, 2012). Often, the uncanniness is assessed and explored experimentally using listener likeability ratings: likeability increases as the human-likeness of a device increases, but drops, creating a non-linear function, when human-likeness approaches actual "human" levels. We will explore the uncanny valley through vocal alignment. Since greater degrees of vocal alignment has been shown to correlate with higher ratings of likeability and attractiveness, it can also serve as a way to explore people's subtle reactions to incongruency of cues for humanness. Our interpretation is that dips in degree of alignment for the inauthentic guise reflects a negative reaction toward the top-down and bottom-up pairing for a given condition. Thus, a finding of an interaction between Voice Type and Apparent Guise, where a device or human voice receives unequal patterns of alignment behavior in an inauthentic guise, would support the cue incongruence hypothesis.

Methods

Stimuli

Target words consisted of 30 CVC monosyllabic English low usage frequency real words, balanced by vowel categories (/i/: *weave, teethe, deed, cheek, peel, key*; /æ/: *wax, wag, vat, tap, nag, bat*; /ɑ/: *wad, tot, sod, sock, pod, cot*; /o/: *woe, soap, moat, hone, comb, coat*; /u/: *zoo, toot, hoop, dune, doom, boot*). Target words were selected as a subset of the original 50 words used in Babel (2012) by omitting words with complex codas or open syllables. Stimuli consisted of recordings of the target words produced by 8 distinct voices. For the real human voices, target words were recorded by 4 humans (2 females, 2 males), native English speakers of American English in their 20s, using a Shure WH20 XLR head-mounted microphone in a sound-attenuated booth. For the device voices, recordings of the 4 TTS voices (2 females, 2 males) were generated with standard parametric TTS in Amazon Polly (US-English): “Joanna”, “Matthew”, “Salli”, and “Joey”. All recordings were amplitude normalized to 60 dB.

Participants

Overall, 92 participants were recruited from the UC Davis undergraduate subjects’ pool and received course credit for their participation. The participants’ mean age was 20.3 years (range=18-42 years old). All participants were native English speakers and reported having no visual or hearing impairments. 76 participants reported experience using a digital device on at least a weekly basis (either Apple’s Siri, Amazon’s Alexa, Microsoft’s Cortana, and/or Google Assistant), while 16 participants reported no device usage.

Procedure

Participants were assigned to either the Human Voice Type condition (n=53) or the TTS Voice Type condition (n=39) in a between-subjects design. The procedure and instructions were identical across conditions, except in the Human Voice Type condition participants heard only the human voices while in the TTS Voice Type condition, participants heard only the Alexa TTS voices. Within the Human Voice Type and TTS Voice Type conditions, 2 of the voices was assigned a Human Guise and 2 of the voices was assigned a Device Guise (for the Human Guise, male and female voices were assigned images and names corresponding to their apparent gender). The matching of each voice to a human or device guise was counterbalanced across 4 lists for each Voice Type condition (the 4 lists contained all possible different pairings of a voice to a guise). Participants were randomly assigned to a Voice Type condition and list.

Before beginning the study, all participants were told they would be repeating words produced by both device and human voices. Each participant was presented with four talkers: two apparent device models (a female and male) and two apparent human models. An introductory slide presented this information and presented the four talkers, including names “Joanna” and “Matthew” (female and male device voices, respectively) and “Melissa” and “Carl” (female and male human talkers) (see Figure 1). Along with names, pictures of the talkers were also provided. The pictures served to reinforce the apparent humanness guise for each of the voices. The images for the voice-AI interlocutors consisted of two Amazon Echo devices, while the images for the humans were two stock photos of adult humans of corresponding genders (photos were selected from the first Google results of “male” and “female stock image” at the time of the study design).

Figure 1: Voice-AI device (Amazon Echos) and human guises used in the present study.

The first part of the study was a baseline word production block. The purpose of this block was to collect the pre-exposure production of each of the 30 target words by each participant. In this block, each of the 30 target words were shown on a computer screen one at a time and participants were instructed to produce the word aloud. Participants read each word aloud twice, randomly selected across 2 blocks.

Following the pre-exposure production block, participants completed the shadowing block: on each trial, they heard one of the four interlocutor voices saying one of the target words and were instructed to repeat the word. On the screen, they saw the speaker guise and a sentence showing the name of the speaker with the target word (e.g., “Joanna says ‘doom’”). Each trial consisted of a randomly selected word-talker pairing. A block containing one repetition of each item per talker was repeated twice in the experiment. In total, participants shadowed the 30 words twice for each interlocutor voice (4 model talkers x 30 words x 2 repetitions = 240 shadowed word productions per participant). Each word production was recorded and digitized at a 44kHz sampling rate using a Shure WH20 XLR head-mounted microphone in a sound-attenuated booth. The entire experiment took roughly 45 minutes.

Analysis

Measuring alignment: Difference in distance (DID)

In this experiment, we focused on durational alignment, given the close correspondence between overall vocal alignment patterns for human/Siri voices observed across a global similarity paradigm (cf. AXB similarity in Cohn et al., 2019) and difference in distance (DID) vowel duration measures (Snyder et al., 2019). Following Snyder et al. (2019), we used a DID acoustic measure to assess degree of vocal alignment toward the model talkers. First, recordings were force-aligned; vowel boundaries were hand-corrected by a trained phonetician based on the presence of voicing and higher formant structure. Using a Praat script, we measured vowel duration (in milliseconds) from participants' pre-exposure and shadowing productions, as well as from the model talkers' productions. Degree of alignment was then tabulated as "difference in distance" (DID, Pardo et al., 2013): DID = |baseline duration - model duration| - |shadowed duration - model duration|. First, the absolute difference between the participant's baseline production and the model's production was calculated. Then, the absolute value of the difference between the model's production and the participant's shadowed production was calculated. Note that we matched first and second baseline repetition to first and second shadowed production, respectively. Finally, we calculated the difference of these two differences (DID). This measure assesses overall alignment. Positive DID values indicate alignment toward the model talker's production, while negative values indicate divergence from the model. Additionally, the magnitude of DID reflects *degree* of alignment: larger positive values indicate greater convergence, while smaller positive values indicate weaker convergence.

Statistical analysis

DID scores were modeled using a mixed effects linear regression including main fixed effect predictors of Voice Type (TTS, Human) and Humanness Guise Authenticity (Authentic, Inauthentic), as well as their interaction. Random effects structure consisted of by-Participant random intercepts and by-Participant random slopes for Humanness Guise.

Results

The model did not compute significant main effects of either Voice Type [β=-1.5, SE=1.1 p=.1] or Guise [β=.07, SE=.34 p=.8]. Yet, the model did compute a significant interaction between Voice Type and Guise [β=-.85, SE=.34 p=.01]. This interaction is illustrated in Figure 2. Overall, mean DID values for each condition were above 0, meaning that shadowers did align to the vowel duration properties of both device and human model talkers. Yet, there were differences in degree of alignment across conditions. DID values were largest (indicating greatest degree of convergence) in shadowed productions of TTS voices presented in the Authentic guise (i.e., with a picture of a device). Yet, when the TTS voices were presented in the inauthentic guise (i.e., presented with a picture of a human), shadowers displayed less alignment to these same word productions. Human voices presented in the authentic guise received the smallest DID values, indicating least degree of convergence. When the human voices were presented in the inauthentic guise (i.e., with a picture of a device), however, degree of convergence increased.

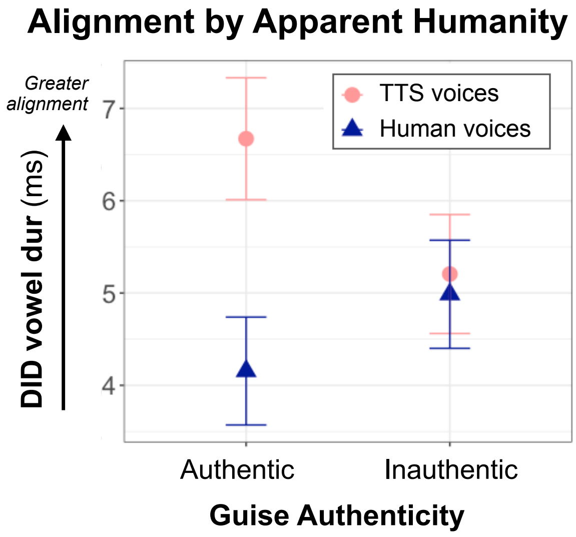

Figure 2: Vowel duration DID scores (means and standard errors of the mean) for text-to-speech (TTS) synthesized and Human Voices by Guise Authenticity (Authentic = Guise matches Voice Type, Inauthentic = mismatch between Voice Type and Guise).

Discussion

The current study explored the effect of *humanness guise* (i.e., telling people voices were produced by either a human or a device) on degree of vocal alignment to voices naturally produced by humans and voices generated by a common voice-AI system (Amazon Polly TTS voices). In doing so, we aimed to examine humanness (device vs. human) as a social category – how it predicts whether people interacting with devices will adopt the speech patterns produced by an apparent AI system, or not, relative to how they would with an apparent human interlocutor.

First, we hypothesized that acoustic differences between TTS and naturally produced voices might predict alignment patterns. We do not find support for this hypothesis; the model did not find an effect of Voice Type. Overall, we observe that participants align toward both voice types (naturally produced and TTS). That we observe some vocal alignment for both voice types is broadly in line with the "Computers are Social Actors" (CASA) framework (Nass et al., 1997): participants apply human-human social behaviors to voice-AI. In this case, the behavior is vocal alignment – a robust behavior with social motivations in human-human interaction (cf. Babel, 2012).

Yet, contra a strict interpretation of CASA, we observe that patterns of vocal alignment are *mediated* by top-down representations of the voices — as produced by either real humans or devices. This supports our second hypothesis, that we would see an interaction between voice type and guise authenticity. Participants were more likely to vocally align to a device TTS voice when the guise was authentic, yet, alignment was reduced when given the human guise for the synthetic voices (i.e., the inauthentic guise). Meanwhile, we observe the reverse pattern for the real human voices: in the authentic guise (seeing a human picture, hearing a human voice) participants show less alignment than when they are told the voice is from a device talker.

The asymmetry observed between the authentic and inauthentic guises suggests that both the low-level acoustic differences between human and device voices and the top-down effect of categorizing the interlocutor as a human or non-human entity are factors that influence our speech behavior. Interestingly, in the current study, these effects appear to be additive: participants aligned more to the TTS than the human voices in the authentic guise; yet switching the guise leads participants to align to an equal extent toward both voices, reducing alignment toward the same TTS voices and increasing alignment toward the same human voices. These findings support our proposal of a gradient personification of voice-AI based on the social cues available — one that is not categorically applied to technology that exhibits cues of "humanity" (cf., CASA; Nass et al., 1997).

This particular pattern of alignment toward the TTS voices as a result of humanness guise additionally supports the proposal by Moore (2012) that cue congruence can explain human behavior interactions with non-human entities: the presence of two cues which provide conflicting information about the realism of an entity leads people to react in a negative way (i.e., with feelings of disgust or discomfort). In the current study, this is realized as decreasing degree of alignment toward the device voice when presented in a human guise. The cues to non-humanness are realized in the synthetic quality of the voices, thus the false guise leads participants to align less to the device voices.

Yet, our finding of *less* alignment toward the human voices in the authentic guise is in contrast with recent findings for human/voice-AI in a similar population of university-age students (Cohn et al., 2019; Snyder et al., 2019). It is possible that there were idiosyncrasies in the particular human voices used in the current study, all of whom were undergraduates at the time of recording. The TTS voices, on the other hand, consisted of Amazon Polly voices, which differ in various ways with the quality of the Siri voices used in prior work. One area for future study is to control for and vary the style of speech across the two interlocutor types (e.g., using more formal/expressive TTS and human productions, relative to more neutral productions). It may be that the more casual-sounding productions by the humans were favored less than more formal and/or expressive productions by the Amazon Polly voices, even if the TTS voices were synthetic.

Another possibility is that the apparent *age* of the talkers may have mediated degree of vocal alignment. In a follow-up post-hoc ratings study, we found that the 4 human voices were rated as younger (mean age rating = 29 years old) than the 4 TTS voices (mean age rating = 35.7 years old). The extent to which acoustic indices of speaker age interact with top-down effects (e.g., varying age guises) has, to our knowledge, not yet been explored in human-computer/voice AI interaction, or in vocal alignment more generally.

Another possibility for the asymmetry in the present study is that our "authentic" human guise was still computer-mediated. Participants completed the experiment through a computer (via E-Prime), perhaps making it a more ecologically valid interaction with the device (TTS) voices, but less natural way to interact with the humans. This interpretation would be in line with Branigan et al. (2003), who found greater syntactic alignment toward computers when the participants thought they were interacting (via text) with a computer versus with a real human. This suggests that *matching* expectations — such as an individual's expectation to engage with a computer via typing — may impact degree of alignment.

On the other hand, our observation of greater alignment toward device voices in the authentic guise parallels prior studies' observations that people display greater syntactic/lexical alignment toward apparent computer, relative to apparent human, interlocutors (e.g., Cowan et al., 2015; Branigan et al., 2010, 2011). The interpretation from these studies, was that the functional difficulty related to communicating with a computer, since computers are believed to be less communicatively able than humans, lead to greater alignment (Cowan et al., 2015). It is possible that the same motivation could be argued to be at play in the present study, leading to greater vocal alignment toward to the TTS voice in the device guise, where authenticity lead to a more believable scenario where the interlocutor was communicatively disadvantaged. Thus, across multiple studies where top-down guise is manipulated, people display greater alignment, across multiple linguistic levels, toward technological agents than toward humans. Future work exploring this possibility can shed light on this possibility.

Overall, this study provides a first step in examining how bottom-up and top-down influences of apparent humanness shape vocal alignment, serving as a window into participants' subconscious social behaviors toward AI and human interlocutors.

References

Athey, K., Coleman, J. E., Reitman, A. P., & Tang, J. (1960). Two experiments showing the effect of the interviewer's racial background on responses to questionnaires concerning racial issues. *J. Applied Psychology*, *44*(4), 244.

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

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

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proc. Inter., Mobile, Wearable and Ubiquitous Tech.*, *2*(3), 1–24.

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

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57.

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. *Proc. Cognitive Science Society*, 186–191.

Burnham, D., Joeffry, S., & Rice, L. (2010). " Does-Not-Compute": Vowel Hyperarticulation in Speech to an Auditory-Visual Avatar. *A-V Speech Processing 2010*.

Chakrani, B. (2015). Arabic interdialectal encounters: Investigating the influence of attitudes on language accommodation. *Language & Communication*, *41*, 17–27.

Cohn, M., Ferenc Segedin, B., & Zellou, G. (2019). Imitating Siri: Socially-mediated alignment to device and human voices. *Proc. ICPhS*, 1813–1817.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *Int'l J. Human-Computer Studies*, *83*, 27–42.

Delvaux, V., & Soquet, A. (2007). The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*, *64*(2–3), 145–173.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll Virginia 1989. *Public Opinion Quarterly*, *55*(3), 313–330.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phonetics*, *34*(4), 458–484.

Heyselaar, E., Hagoort, P., & Segaert, K. (2017). In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior Research Methods*, *49*(1), 46–60.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88.

Hutchinson, K. L., & Wegge, D. G. (1991). The Effectiveness of Interviewer Gender Upon Response in telephone Survey Research. *J. Social Behavior and Personality*, *6*(3), 573.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Dev. Psych.*, *25*(6), 954.

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports*, *2*, 864.

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

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems 1. *J. Applied Social Psych*, *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.

Németh, G., Fék, M., & Csapó, T. G. (2007). Increasing prosodic variability of text-to-speech synthesizers. *Proc. ISCA*.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *JASA*, *119*(4), 2382–2393.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *JML, 69*(3), 183–195.

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. *Proc. Interspeech*, 116–120.

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *JASA*, *140*(5), 3560–3575.