Speech Communication 135 (2021) 66–75

Contents lists available at ScienceDirect

Speech Communication

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

Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers

Michelle Cohn *, Kristin Predeck, Melina Sarian, Georgia Zellou

Phonetics Laboratory, UC Davis Department of Linguistics, 1 Shields Avenue, Davis, CA, 95616, United States

ARTICLE INFO

Keywords:
vocal alignment to emotional expressiveness
human-computer interaction, voice-activated artificially intelligent (voice-AI) devices

ABSTRACT

This study tests whether individuals vocally align toward emotionally expressive prosody produced by two types of interlocutors: a human and a voice-activated artificially intelligent (voice-AI) assistant. Participants completed a word shadowing experiment of interjections (e.g., "Awesome") produced in emotionally neutral and expressive prosodies by both a human voice and a voice generated by a voice-AI system (Amazon's Alexa). Results show increases in participants' word duration, mean f0, and f0 variation in response to emotional expressiveness, consistent with increased alignment toward a general 'positive-emotional' speech style. Small differences in emotional alignment by talker category (human vs. voice-AI) parallel the acoustic differences in the model talkers' productions, suggesting that participants are mirroring the acoustics they hear. The similar responses to emotion in both a human and voice-AI talker support accounts of unmediated emotional alignment, as well as computer personification: people apply emotionally-mediated behaviors to both types of interlocutors. While there were small differences in magnitude by participant gender, the overall patterns were similar for women and men, supporting a nuanced picture of emotional vocal alignment.

1. Introduction

Speakers can readily convey their inner mental states and emotions to an interlocutor via acoustic-phonetic properties of their voice, such as using pitch and temporal variation. For example, happy speakers produce higher pitch and longer durations (Abadjieva et al., 1993; Murray & Arnott, 1993; Viscovich et al., 2003; Yildirim et al., 2004). While understudied, how listeners respond to this emotion subsequently *in their own speech* can reveal the mechanisms of speech and emotional alignment. When humans interact, there is a tendency for them to *vocally align*, subconsciously mirroring each others' (non-emotional) speech patterns (Babel, 2012; Nielsen, 2011; Pardo, 2006; Pardo et al., 2010; Zając, 2013), including prosodic features associated with emotion, such as speaking rate (Pardo et al., 2010), average pitch (Babel & Bulatov, 2012), and pitch range (Smith, 2007).

There is some evidence for alignment of emotional expressiveness: happiness or sadness conveyed in the prosodic patterns of an interlocutor appear to shape talkers' speech patterns (Arimoto & Okanoya, 2014; Vaughan et al., 2018; Xiao et al., 2015, 2013), though nearly all prior studies on this topic have been conducted in non-controlled settings (e.g., spontaneous speech while participants completed an interactive task in Arimoto & Okanoya, 2014; during a counseling session in Xiao et al., 2015). For example, Vaughan et al. (2018) observed that a female psychiatrist aligned in pitch, speaking rate, and vowel-spectral features toward six of her patients during clinical interviews. In particular, they found that during interactions with depressed speakers, the therapist adopted the more contracted vowel space of her patients, which is a characteristic of depressed speech (Scherer et al., 2016). Similarly, Yang et al. (2013) observed that interviewers aligned toward the lower pitch of depressed individuals when engaging with them. Together, alignment of features that signal emotional state has been proposed to convey empathy toward the emotional state of the interlocutor (Scherer et al., 2014; Vaughan et al., 2018; Xiao et al., 2015, 2013).

The present study compares vocal alignment toward words produced in emotionally expressive and neutral speech styles in a controlled laboratory setting with two types of talkers: a human voice and a voice-activated artificially intelligent (voice-AI) assistant voice. Voice-AI assistants (e.g., Apple's Siri, Amazon's Alexa, and Google Assistant) are increasingly prevalent in households in the United States (Ammari et al., 2019; Bentley et al., 2018). Unlike computer systems in the past, these systems display more apparent human characteristics. For example,

* Corresponding author.
E-mail addresses: mdcohn@ucdavis.edu (M. Cohn), kpredeck@ucdavis.edu (K. Predeck), msarian@ucdavis.edu (M. Sarian), gzellou@ucdavis.edu (G. Zellou).

https://doi.org/10.1016/j.specom.2021.10.003
Received 5 May 2020; Received in revised form 8 September 2021; Accepted 16 October 2021
Available online 20 October 2021

Amazon's Alexa voice is capable of producing *emotionally expressive* interjections. The Amazon Alexa Skills Kit (ASK), a system generated for 'voice app' developers, includes over 100 emotionally expressive words (e.g., "Awesome!") and phrases (e.g., "Whoops a daisy!") pre-recorded by the US-English female Alexa voice, which can be implemented with speech output tags (known as 'Speechcons': Amazon, 2018). Cohn & Zellou (2019) found that speakers showed significantly greater vocal alignment toward emotionally expressive Alexa interjections than neutral productions generated with the default Alexa voice. Beyond vocal alignment, there is other evidence that people treat computer systems with similar emotional/affective responses as they do for other humans (Brave et al., 2005; Nass et al., 1999; Bucci et al., 2018; Cohn et al., 2019; Nass et al., 1995). For example, in a study of interactions with a car navigation system, participants had fewer accidents, were less distracted, spoke more, and reported greater satisfaction with the system when the emotional speech patterns in the text-to-speech (TTS) voice matched their own mood (either happy or sad) (but note that these voices did not contain 'neutral', non-emotional productions for comparison) (Nass et al., 2005). Together, finding similar responses toward computer/voice-AI agents and toward humans supports computer personification theories; for example, the 'Computers are Social Actors' (CASA) theoretical framework (Nass et al., 1997, 1994) proposes that if a cue of humanity is detected in a computer system, people subconsciously apply the social norms from human-human interactions, even if it is clear that the system is non-human. Yet, critically, the vast majority of prior studies examining the effect of emotional expressiveness by a computer or voice-AI system have not contained a direct comparison to a human (Brave et al., 2005; Bucci et al., 2018; Cohn & Zellou, 2019; Liu & Sundar, 2018; Nass et al., 1999; Nass et al., 1995).

Indeed, work examining (non-emotional) vocal alignment has demonstrated *differences* in how individuals align toward device and human voices when direct comparisons are made (Cohn et al., 2019; Raveh et al., 2019; Snyder et al., 2019). In a study comparing alignment toward a voice-AI (Amazon's Alexa) and a human interlocutor, Raveh and colleagues (2019) found that people do align toward voice-AI; but, when a human confederate was present, participants aligned less toward the Alexa voice. Similarly, in two studies examining single-word shadowing of voice-AI and human interlocutors, participants displayed greater alignment toward human voices, relative to Apple's Siri voices (Cohn et al., 2019; Snyder et al., 2019). With respect to emotion, a recent study (Cohn et al., 2020) found that listeners perceived synthesized 'happiness' in a human and Alexa voice (from 'emotionally-neutral' productions) similarly in some respects (e.g., increased perceived arousal with 'happiness' manipulations) but differently for others: listeners did not hear the same increase in valence with the 'happiness' manipulation in the Alexa voice. Taken together, the alignment and emotion perception findings suggest that voice-AI systems might be a separate social category from humans, and thus serves as a relevant interlocutor comparison for vocal alignment toward emotional expressiveness.

In the current study, examining emotionally expressive (and neutral) speech by Amazon's Alexa and a human interlocutor can serve as a test of our scientific understanding of emotional mimicry, teasing apart theories that it is 'unmediated' or 'socially-mediated', such as by the characteristics of the speaker (as human or voice-AI).

1.1. Unmediated, matched motor accounts

On the one hand, *unmediated, matched motor accounts* propose that the mechanism underlying emotional alignment is embodied cognition, or a matched motor response (De Waal, 2007; Decety & Jackson, 2006; Preston, 2007). For example, Arias and colleagues (Arias et al., 2018) found that participants listening to a 'smiling' voice produced more micro-activations of the zygomatic muscle (used to pull the mouth widthwise in smiling) than when they heard a 'frowning' voice. Further, individuals align to smiles produced by both in-group and out-group members (Van Der Schalk et al., 2011) and smile or laugh along with laugh tracks (Fuller & Sheehy-Skeffington, 1974), supporting an *unmediated* motor mechanism (i.e., without a social mediator). There are multiple proposed mechanisms for this interpersonal alignment, including mirror neurons, which are thought to fire when a person either completes an action or sees another person complete an action (Decety & Jackson, 2006). Other proposed mechanisms include spreading activation of exemplars, where an experience (e.g., hearing a word produced in a certain way) updates the listener's own mental representations. In the domain of speech, the updating of exemplars has been one proposed mechanism for vocal alignment more generally (Goldinger, 1996, 1998). While the aim of the current investigation is not to tease apart views about the neural underpinnings of emotional alignment, as based in spreading activation or mirror neurons, these types of accounts make a similar prediction for the present study: that motor and/or linguistic representations are equally 'activated' upon hearing emotional speech from different interlocutors (here, human vs. voice-AI). Furthermore, emotional speech (relative to non-emotional speech) often contains more exaggerated acoustic-phonetic features: longer duration and higher f0 for 'happy' speech, relative to 'neutral' speech (Abadjieva et al., 1993; Murray & Arnott, 1993; Viscovich et al., 2003; Yildirim et al., 2004). If participants are merely 'mirroring' the input, consistent with *unmediated accounts*, then we might expect them to display greater alignment for these features as a function of their acoustic distance (e.g., greater alignment toward longer segments in emotionally expressive speech). Such a prediction would apply regardless of whether the speaker is a human versus AI system, which is consistent with recent work. For example, Gazzola and colleagues (2007) found identical engagement of regions associated with motor-action perception for household movements produced by both human and robot agents. Thus, in the present study we might predict no overall difference in alignment toward emotionally expressive speech produced by human voices and voice-AI TTS, reflecting a general motor-perception mechanism.

1.2. Socially mediated accounts

On the other hand, *socially mediated accounts* propose that the social relationship between interacting humans mediates patterns of emotional alignment (Fischer et al., 2019; Hess & Fischer, 2013, 2014). Work in linguistic alignment more generally (i.e., not necessarily emotional) has demonstrated that the social dynamics of an interaction shape convergence and divergence between interlocutors (Abrego-Collier et al., 2011; Babel, 2012; Yu et al., 2013). Speech coordination is often examined through the framework of 'Communication Accommodation Theory', or CAT (Giles et al., 1991; Giles & Baker, 2008; Shepard, 2001). CAT proposes that speakers demonstrate their social closeness via increased alignment, or that they increase social distance via divergence. For example, individuals display greater (non-emotional) alignment toward interlocutors they are socially close to: over the course of a year, college roommates who reported stronger feelings of closeness also displayed greater vocal alignment (Pardo et al., 2012). In the emotion alignment literature, there is some evidence for differences based on the social dynamics between individuals. For instance, there is greater reported emotional alignment toward in-group versus out-group members (Matsumoto, 2002; Thibault et al., 2006; Weisbuch & Ambady, 2008). Group-mediated alignment is also observed even within a lab setting, where participants are arbitrarily assigned to 'teams' (Lakin et al., 2003). In the current study, socially-mediated alignment patterns might be realized as greater emotional vocal alignment toward the 'in-group' human voice, relative to the voice-AI talker, which represents a distinct social category. This would be consistent with prior findings for (non-emotional) alignment where greater alignment toward human, compared to voice-AI, interlocutors was observed (Cohn et al., 2019; Raveh et al., 2019; Snyder et al., 2019). Alternatively, participants might find emotionally expressive productions by voice-AI to be

'uncanny' (Mori, 1970; Mori et al., 2012), and subsequently diverge from those productions only. In either scenario, *socially-mediated accounts* would predict categorical differences in emotional vocal alignment towards humans and voice-AI.

Another social factor that has received more attention in the emotional alignment literature is the role of gender (Arimoto & Okanoya, 2014; Cohn, Ferenc Segedin, et al., 2019; Cohn & Zellou, 2019; Doherty et al., 1995). On the one hand, some have proposed that 'emotional contagion' is stronger for women, relative to men, based on differences in socialization (Doherty et al., 1995; Sonnby-Borgström et al., 2008). This is in line with findings reporting that women showed greater vocal alignment toward emotion in a competitive dialog game (Arimoto & Okanoya, 2014) and with (non-emotional) vocal alignment of single-word shadowing (Namy et al., 2002). In the present study, one prediction is that female participants will show greater (emotional) alignment than male participants. However, there is some evidence of the opposite pattern: greater speech alignment by *males*. For example, Cohn & Zellou (2019) found that male participants displayed more alignment toward emotionally expressive Alexa productions in a single word shadowing study. Broadly, greater alignment by males is also consistent with studies of (non-emotional) vocal alignment demonstrating similar asymmetries (Dijksterhuis & Bargh, 2001; Pardo, 2006). Therefore, an alternative prediction for the present study is that men will show greater alignment toward emotionally expressive speech, relative to women.

1.3. Current Study

The current study examines vocal alignment toward neutral and emotionally expressive interjections. In particular, participants completed a word shadowing task (Goldinger, 1998) where they repeated isolated words (presented over headphones). We measured three prosodic properties associated with vocal emotional expression: word duration, mean fundamental frequency (f0; perceived pitch), and f0 variation (Abadjieva et al., 1993; Murray & Arnott, 1993; Viscovich et al., 2003; Yildirim et al., 2004). Specifically, we test whether acoustic alignment toward emotional expressiveness differs based on interlocutor (human vs. voice-AI) and speaker gender (male or female), which can speak to theories of emotional alignment (as an *unmediated, matched motor response* or *socially-mediated*). We compare two types of interlocutors: a human voice (naturally produced) and an Amazon Alexa voice (generated from the US-English TTS voice). As mentioned, the Alexa voice is capable of producing hyper-expressive interjections (e.g., "Wow!") ('Speechcons', Amazon, 2018).

For each acoustic feature, we assess change from the speaker's baseline (elicited during a pre-exposure phase). Specifically, we centered each participant's shadowed production to their baseline production (Cohn et al., 2021). In doing so, we test whether, relative to their baseline speech characteristics, speakers increase their word duration, mean f0, and f0 variation when shadowing emotionally expressive speech. While prior studies have frequently used difference-in-distance (DID) measures to quantify alignment (Babel, 2012; Snyder et al., 2019; Zellou & Cohn, 2020), recent work suggests that DID can be biased to find larger differences for participants with larger baseline distances from the model talkers and can also result in apparent divergence for speakers who are more similar to the model talkers at baseline (Cohen Priva & Sanker, 2019; MacLeod, 2021). Using a baseline-centered approach allows us to test if speakers make general prosodic adjustments (relative to their pre-exposure productions) in response to a particular interlocutor and in response to emotional expressiveness which might otherwise obscure (or spuriously enhance) alignment effects if assessed using DID. For example, there is work showing that people produce a higher mean f0 in Alexa-directed speech, in a direct comparison with human-directed speech (Raveh et al., 2019; Siegert and Krüger, 2021), and smaller f0 range and shorter productions (Siegert et al., 2019). These prosodic differences are argued to be, in part, driven by differences in perceptions of the voice-AI/computer as being less communicatively competent (compared to the human) (Branigan et al., 2011; Oviatt et al., 1998), which can be triggered by hearing a TTS voice (Cowan et al., 2015). Here, we might similarly predict interlocutor-based effects (shorter words, higher mean f0, and smaller f0 range toward Alexa).

2. Methods

2.1. Participants

A total of 66 native English speakers were recruited from the UC Davis Psychology subjects pool (30 females, 36 males; mean age = 20.64 ± 2.43 years). Table 1 provides a summary of the demographic characteristics of the participants. Nearly all participants had experience with a voice-AI system (e.g., Apple's Siri, Google Assistant, etc.).

2.2. Stimuli

The words were selected from the available set of Alexa 'Speechcons' (Amazon, 2018), consisting of interjections and phrases spoken by the Amazon Alexa voice actor in an emotionally expressive way[1]. Interjections can be used to express the speaker's mental or emotional state (Ameka, 1992) and to convey the disposition or attitude of the speaker (Goffman, 1981). Following Cohn & Zellou (2019), stimuli consisted of 18 interjections generated in neutral and emotionally expressive manners (*awesome, bravo, cheers, cool, ditto, dynamite, eureka, great, howdy, hurray, jinx, roger, splash, super, wow, wowzer, yum, zing*). All the items can be classified as having a positive emotional valence based on the words' lexical and/or prosodic qualities. An additional six interjections with negative valence were presented in the experiment (e.g., "darn"; see Cohn & Zellou, 2019 for full list), but not included in the final analysis which aimed to examine imitation of tokens with a consistent emotional valence. Using the Alexa Skills Kit (ASK), the 'neutral' Alexa productions of the words were generated with the default prosody, while the 'emotionally expressive' Alexa productions were generated using the Speech Synthesis Markup Language (SSML) tags (e.g., <say-as interpret-as= "interjection" > awesome! </say-as>). For the human model talker condition, a 24 year-old white, female native English speaker of American English (from California) was recorded producing the same set of words with neutral and emotionally expressive prosody. The recording took place in a sound attenuated booth, where the speaker wore a head-mounted microphone (Shure WH20 XLR). The human speaker naturally produced the words in her own neutral and emotionally expressive manners; she did not imitate the productions by the Alexa voice. Both the Alexa and human productions were amplitude normalized in Praat (70 dB). The stimuli are available for audio illustration at Open Science Foundation (OSF)[2].

Acoustic analyses of the Alexa and human productions in Praat are

Table 1
Subject demographics.

Gender	n	Mean age (sd)	Experience with Alexa	Used voice-AI (#)
Females	30	20.66 yrs (1.67)	24 Yes; 6 No	28 Yes; 2 No
Males[a]	36	20.63 yrs (2.95)	27 Yes; 9 No	36 Yes
Total	**66**	**27.17 yrs (4.96)**	**51 Yes; 15 No**	**64 Yes; 2 No**

[a] Including 1 trans-male.

[1] **At the time of the study, the only digital assistant voice capable of producing both a 'neutral' and an 'emotionally expressive' production was the US-English female Amazon Alexa voice. Due to this constraint, we used one human speaker, also producing neutral and emotionally expressive productions.**

[2] **https://doi.org/10.17605/OSF.IO/GDWPR**

provided in Table 2. These measurements confirm that, relative to neutral productions, expressive productions for both Alexa and human speakers are longer [$t(53.18)=3.85$, $p<0.001$], have higher mean f0 [$t(51.54)=4.55$, $p<0.001$], and have greater f0 variation [$t(66.66)=4.30$, $p<0.001$], consistent with acoustic features of happy speech (Abadjieva et al., 1993; Murray & Arnott, 1993; Viscovich et al., 2003; Yildirim et al., 2004). Two-sample t-tests for each acoustic property revealed that the human and Alexa stimulus items did not significantly differ in their overall word duration [$t(59.69)=-1.12$, $p=0.27$] or mean f0 [$t(60.86)=-0.80$, $p=0.43$], but did for f0 variation [$t(67.97)=2.95$, $p<0.01$], with more variation produced overall by the Alexa voice.

Comparisons of emotionally expressive productions by the Alexa and human revealed no difference for word duration [$t(27.24)=0.11$, $p=0.92$] or f0 variation [$t(33.68)=0.60$, $p=0.56$], but a lower mean f0 for the emotionally expressive Alexa [$t(33.92)=-3.11$, $p<0.01$].

Comparisons of neutral productions also showed differences across talkers: shorter word duration for Alexa [$t(31.24)=-3.17$, $p<0.01$], a lower mean f0 for the human [$t(27.71)=4.34$, $p<0.001$], and a larger f0 variation for the Alexa neutral voice [$t(22.53)=6.06$, $p<0.001$].

2.3. Procedure

Participants began with a pre-exposure word production block, where they saw and read aloud each of the target words (randomly presented, one at a time, in three repetition blocks). Next, they completed the word shadowing blocks. First, participants were introduced to the model talkers: either the voice-AI system ('Alexa') or a human ('Melissa'), accompanied by a corresponding image of an Echo device or a female human stock photo to strengthen the guise, as clearly a human or a voice-AI interlocutor (shown in Fig. 1). On each trial, participants were told to "repeat the word" produced by each talker (ISI = 1000ms). All items across both expressiveness conditions (Emotionally Expressive or Neutral) and Model Talker were randomly presented within each repetition block. In total, participants completed two repetition blocks.

2.4. Analysis

2.4.1. Acoustic Analysis

Participants' baseline and shadowed productions were force-aligned with FAVE (Rosenfelder et al., 2011) and hand-corrected, focusing on the start and end of the word, by the second and third author. For each word, the mean f0[3] and f0 variation[4] (standard deviation over the word) values were taken using a Praat script adapted from DiCanio (2007) in semitones[5] (ST, relative to 100 Hz), with plausible maxima and minima f0 for each gender (78-150 Hz for males, 150-350 Hz for females). Model talkers' productions (i.e., Alexa and human) were also FAVE aligned, hand-corrected, and measured with the same methods. Table 3 provides the pre-exposure means for the three acoustic properties of interest, separately for female and male participants.

For each acoustic feature (word duration, mean f0, and f0 variation), we centered each subject's production in the shadowing experiment, relative to their pre-exposure productions. The third repetition of each word in the pre-exposure phase were selected as participants' 'baseline' productions, as the participants would be familiar with the words by that point and differences in production due to initial word novelty would be reduced. For the shadowed productions, the second repetition of the word for that given interlocutor was selected, following the same reasoning. For each feature, we calculated the mean 'baseline' value for each participant (capturing their baseline speech characteristics), which we subtracted from each 'shadowed' production. These centered values were used as the dependent variables in the linear mixed effects models.

2.4.2. Statistical Analyses

We modeled each (centered) shadowed acoustic property in separate linear mixed effects models with the *lme4* R package (Bates et al., 2015). Fixed effects included Expressiveness Condition (2 levels: neutral, expressive), Model Talker (2 levels: Alexa, human), Gender (2 levels: male, female), and all possible interactions. Random effects included by-Subject and by-Word random intercepts, as well as by-Subject random slopes for Expressiveness Condition (more complex random effects structure resulted in a singularity error, indicating overfit for all three models). Contrasts were sum coded. (Lmer syntax: Shadowed.c[6] ~ Emotion Condition*Interlocutor*Gender + (1+Condition|Subject) + (1|Word).)

Table 2
Acoustic properties of regular and expressive tokens.

Model talker	Expressiveness condition	Word duration (sd)	Mean f0	F0 Variation
Human	Neutral	595.7 ms (120.9)	195.5 Hz	1.5 ST
	Expressive	699.0 ms (161.7)	244.9 Hz	3.1 ST
Alexa	Neutral	483.6 ms (89.0)	212.4 Hz	2.8 ST
	Expressive	707.3 ms (279.3)	215.0 Hz	3.3 ST

3. Results

The word duration model output is shown in Table 4 and the values are plotted in Fig. 2.A. The model revealed a significant intercept: relative to baseline productions (in the pre-exposure), participants produced longer word durations on average. We also observe an effect of Expressiveness Condition: participants produce longer words in response to emotionally expressive productions. A main effect of Model Talker reveals that participants produce less of an increase in word duration toward Alexa. Furthermore, there is an interaction between Expressiveness Condition and Model Talker: speakers' word duration (on average) increases more when shadowing Alexa emotionally expressive productions. No other effects or interactions were significant in the model[7].

The mean f0 model is provided in Table 5 and values are plotted in Fig. 2.B. First, there is a significant intercept, indicating that participants increase their mean f0 in the shadowing experiment (relative to their baseline productions). There is also an effect of Expressiveness Condition indicating that speakers increase their mean f0 when shadowing emotionally expressive productions. While there is no main effect of Model Talker, it interacted with Expressiveness Condition: participants produce less of a mean f0 increase when shadowing Alexa Expressive productions. Additionally, while there is no main effect of Gender, it interacts with Expressiveness Condition: female participants show a larger mean f0 increase when shadowing the expressive productions. No other effects or interactions were significant in the model.

The summary of the model run on f0 variation is provided in Table 6

[3] **Total of 144 excluded mean f0 observations due to creak (10 females: n=78, 9 males: n=66).**

[4] Total of 152 excluded f0 variation observations due to creak (10 females: n=78; 10 males: n=74).

[5] Semitones (ST) are used so that f0 values are on the same scale (across speakers/genders). We use ST in t-tests and in the full analysis.

[6] ".c" is used to indicate that this continuous value has been centered.

[7] Note that the effects are still present even if the model does not include the fixed effect of Gender.

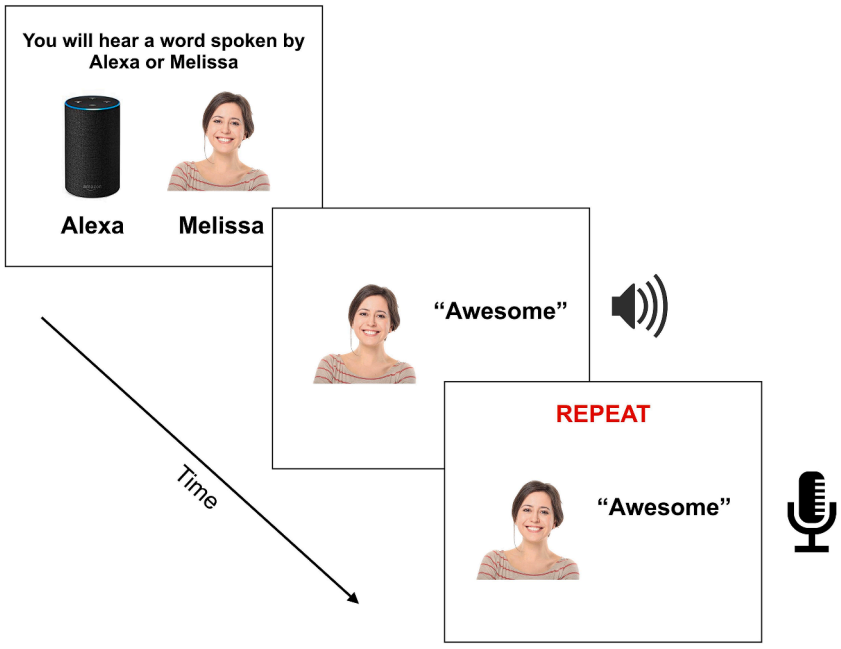

Fig. 1. Introduction of model talkers and general shadowing trial design. (Color online.)

Table 3
Acoustic properties of participants' pre-exposure productions.

	Duration (sd)	F0 mean (sd)	F0 variation (sd)
Females	540.6 ms (134.5)	205.2 Hz	1.25 ST
Males	525.8 ms (144.5)	106.7 Hz	1.26 ST
Pairwise comparison of F vs. M	t(2306.7)=2.57, p=0.10	t(2206.4)= 120.94, p<0.001	t(2300.90)=-0.52, p=0.61

Table 4
Word duration. Summary statistics for the linear mixed effects model.

	Coef	*SE*	*df*	*t*	*p*	
(Intercept)	54.83	24.47	24.02	2.24	0.03	*
Expressiveness Condition (Expressive)	25.53	3.59	63.98	7.1	<0.001	***
Model (Alexa)	-13.74	1.45	4600.03	-9.46	<0.001	***
Gender (F)	3.32	9.99	64	0.33	0.74	
Condition (Expressive) * Model (Alexa)	9.17	1.45	4600.03	6.31	<0.001	***
Condition (Expressive) *Gender (F)	1.68	3.59	63.98	0.47	0.64	
Model (Alexa)*Gender (F)	-0.85	1.45	4600.03	-0.58	0.56	
Condition (Expressive) *Model (Alexa) *Gender (F)	-1.37	1.45	4600.03	-0.94	0.35	

Num. observations = 4,753, Num. subjects = 66, Num. words = 18
Duration.c ~Condition*ModelTalker*Gender + (1+Condition|Subject) + (1|Word)

and values are plotted in Fig. 2.C. The model revealed a main effect of Expressiveness Condition indicating that relative to their baseline productions, speakers increase f0 variation when shadowing Expressive productions. There is also an effect of Model Talker revealing that participants produce a larger increase in f0 variation when shadowing the Alexa voice. Furthermore, there is an interaction between Expressiveness Condition and Model Talker wherein participants produce less of an increase in f0 variation in response to emotional expressiveness by the Alexa voice. Finally, there is an interaction between Expressiveness Condition, Model Talker, and Gender: female participants produce a larger f0 variation increase for the Alexa Expressive condition. No other effects or interactions were significant.

4. Interim Discussion

This study revealed that speakers adapt their speech toward an emotional speech style — with longer durations, higher f0, and larger f0 variation (Abadjieva et al., 1993) — when shadowing isolated words produced by human and Alexa talkers. Yet, how speakers 'emotionally align' differs by the model talker. Specifically, we see a larger increase in word duration when speakers shadow emotionally expressive Alexa productions, but a smaller increase in mean f0 and f0 variation. At first glance, this appears to be *socially-mediated emotional alignment* (Fischer et al., 2019; Hess & Fischer, 2013, 2014), with differences based on the social category of talker.

Yet, an alternative explanation is that differences for voice-AI versus human model talkers could be explained by magnitude of acoustic distance across the voices, which would support *unmediated, matched motor accounts* (De Waal, 2007; Decety & Jackson, 2006; Preston, 2007). As summarized in Section 2.2. and Table 2, the acoustic properties of the model talkers' productions varied. For example, the (mean) word duration difference from neutral-to-expressive was larger for Alexa talker (Δ = 223.7 ms) than for the human talker (Δ = 103.3 ms), the condition where we greater more lengthening during emotional alignment toward Alexa. Similarly, there are larger differences from neutral-to-expressive for the human talker for both f0 properties, consistent with the emotional alignment effects: larger alignment toward the human ($\Delta_{\text{mean f0}}$ = 3.9 ST; $\Delta_{\text{f0 var}}$ = 1.6 ST) than Alexa ($\Delta_{\text{mean f0}}$ = 0.2 ST; $\Delta_{\text{f0 var}}$ = 0.5 ST). Together, these observations suggest that differences in emotional alignment by model talker category (voice-AI vs. human) might be driven by the magnitude of acoustic distance between their neutral-to-expressive productions.

This leads to a related question: do the two Model Talker main effects — words are shorter and have a larger f0 variation when shadowing the Alexa voice — simply reflect mirroring the acoustics, rather than interlocutor-specific adaptations (e.g., human- vs. computer-directed speech: Burnham et al., 2010; (Raveh et al., 2019)? Here, the much

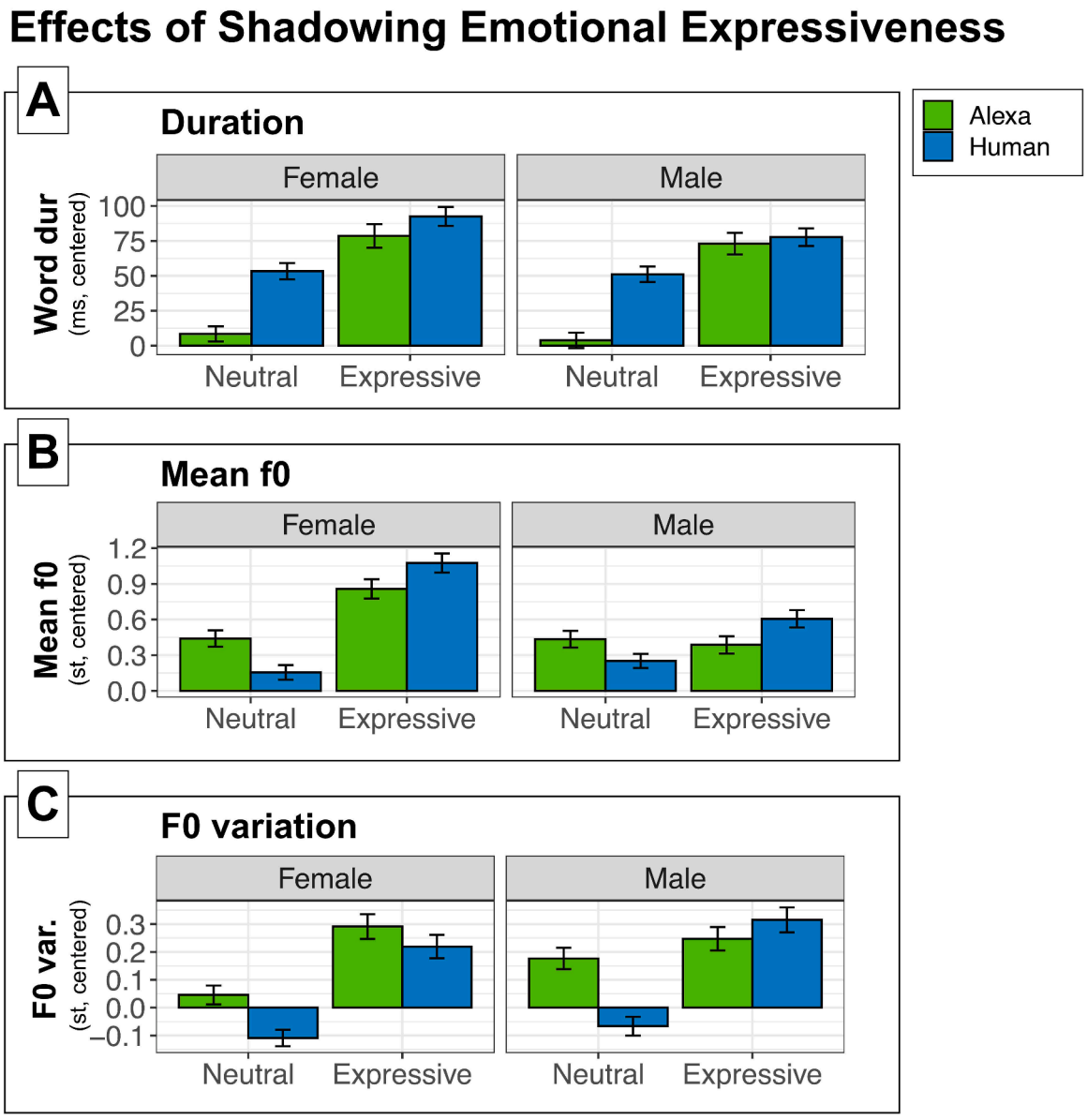

Fig. 2. Prosodic changes by Emotional Expressiveness Condition (neutral vs. expressive), Participant Gender (female vs. male), and Model Talker (Alexa vs. Human) for word-level acoustic measurements: (A) duration (milliseconds, ms), (B) mean f0 (semitones, st), and (C) f0 variation (st). Values are centered to the participants' mean in the pre-exposure phrase. Values higher than 0.0 indicate a relative increase, while values lower than 0.0 indicate a relative decrease, during shadowing.

Table 5
Mean f0. Summary statistics for the linear mixed effects model.

	Coef	*SE*	*df*	*t*	*p*	
(Intercept)	0.53	0.14	74.29	3.76	<0.001	***
Expressiveness Condition (Expressive)	0.20	0.04	64.13	4.63	<0.001	***
Model (Alexa)	4.4e-03	0.02	4536.57	0.22	0.83	
Gender (F)	0.11	0.12	64.04	0.89	0.38	
Condition (Expressive) * Model (Alexa)	-0.11	0.02	4536.53	-5.65	<0.001	***
Condition (Expressive) *Gender (F)	0.13	0.04	64.13	2.93	<0.001	***
Model (Alexa)*Gender (F)	0.01	0.02	4536.61	0.61	0.54	
Model (Alexa)*Condition (Expressive)*Gender (F)	-0.01	0.02	4536.53	-0.61	0.54	

Num. observations = 4,689, Num. subjects = 66, Num. words = 18
Meanf0.c ~Condition*ModelTalker*Gender + (1+Condition|Subject) + (1|Word)

Table 6
F0 variation. Summary statistics for the linear mixed effects model.

	Coef	*SE*	*df*	*t*	*p*	
(Intercept)	0.14	0.08	68.63	1.68	0.10	
Expressiveness Condition (Expressive)	0.13	0.02	64.26	5.38	<0.001	***
Model (Alexa)	0.05	0.01	4533.42	4.52	<0.001	***
Gender (F)	-0.03	0.07	63.99	-0.45	0.65	
Condition (Expressive) * Model (Alexa)	-0.05	0.01	4533.39	-4.41	<0.001	***
Condition (Expressive) *Gender (F)	0.01	0.02	64.26	0.58	0.56	
Model (Alexa)*Gender (F)	0.01	0.01	4533.45	0.50	0.62	
Model (Alexa)*Condition (Expressive)*Gender (F)	0.03	0.01	4533.39	2.53	0.01	*

Num. observations = 4,686, Num. subjects = 66, Num. words = 18
F0var.c ~Condition*ModelTalker*Gender + (1+Condition|Subject) + (1|Word)

shorter word durations in the Alexa Neutral production (averaging 483.6 ms vs. Human Neutral: 595.7 ms; see Table 2) might bias the Model Talker effects. Likewise, the Alexa voice has larger f0 variation overall (2.8 ST in Neutral, 3.3 ST in Expressive), compared to the Human voice (1.5 ST in Neutral, 3.1 ST in Expressive), suggesting this might be driving the overall larger f0 variation produced towards the Alexa model talker.

5. Post hoc analyses

To test whether model talker main effects are consistent across both emotional expressiveness conditions, we analyzed 'neutral' and 'emotionally expressive' productions in separate post hoc models for word duration and f0 variation. As in the main models, the dependent variables were the participants' shadowed value (centered). (Lmer syntax: Shadowed ~ Model Talker + (1|Subject)+ (1|Word).)

The post hoc duration models revealed that speakers' decrease in word duration for Alexa was consistent across the expressiveness conditions. When shadowing Alexa productions, participants produce shorter words in both the Neutral [*Coef*=-22.92, t=-13.64, $p<0.001$] and Expressive conditions [*Coef*=-4.56, t=-2.00, $p<0.05$]. The f0 variation post hoc models showed differences by Model Talker only when shadowing neutral productions: participants produce more f0 variation after hearing Alexa Neutral [*Coef*=0.10, t=6.84, $p<0.001$]. No difference for Model Talker was observed for f0 variation in the expressive condition dataset [*Coef*=1.09, t=0.07, p=0.95].

Together, these post hoc analyses reveal that the shorter word durations when shadowing the Alexa talker were stable across both expressive and neutral conditions. On the other hand, the Model Talker effects in the main f0 variation model — with larger f0 variation when shadowing Alexa — appears to be driven by the Alexa Neutral conditions (which has a larger f0 variation than the Human Neutral condition).

6. General Discussion

The present study is the first, to our knowledge, to examine vocal alignment patterns toward emotionally expressive and neutral productions in a controlled laboratory setting, and toward two types of interlocutors in the same study: human and voice-AI. We observed that participants converged more toward the prosodic features (word duration, mean f0, and f0 variation) of emotionally expressive productions. In particular, when shadowing, participants shift their pronunciations toward the longer duration, higher f0, and larger f0 variation in the emotionally expressive productions, which are characteristics of 'happy' speech (Abadjieva et al., 1993; Murray & Arnott, 1993; Viscovich et al., 2003; Yildirim et al., 2004). This result extends prior findings of emotional alignment in spontaneous, dyadic interactions (e.g., therapists' office in Vaughan et al., 2018) and in-lab approaches examining physiological responses to emotional speech (e.g., zygomatic, 'smile', muscle activation in Arias et al., 2018) to speech shadowing of isolated words. A summary of the effects of the main and post hoc analyses is provided in Table 7.

We additionally compared speech behavior toward two types of model talkers: a naturally produced human voice and an Amazon Alexa TTS voice. Here, we found that speakers produce *shorter* words when shadowing Alexa. Post hoc analyses confirmed that speakers' decrease in word duration for Alexa was reliable across the emotion conditions. While shorter durations contrast with prior work on computer-directed speech more generally (e.g., Burnham et al., 2010), it does align with a recent finding of a direct comparison of a human and modern voice-AI system (Siegert et al., 2019). One possible explanation is that this finding is driven by 'convergence-to-expectation'. Wade (2020) found that imitators adopt expected, but not heard, features of an interlocutor's speech. For instance, they observed monophthongization of /aɪ/ after exposure to a model talker with a Southern American English accent, a feature that was not directly heard in that model talker's speech. In the current study, speakers might have perceived the Amazon Alexa TTS voice as 'sounding' shorter overall, as part of their expectations about the voice. Indeed, prior work has shown that people hear differences in duration for more robotic-sounding speech (i.e., utterances modified to contain audible prosodic disfluencies) relative to smooth-sounding synthetic speech (Boril et al., 2017). While we did not see a lengthening effect for TTS voices here, future work systematically varying the voice — as well as assessing listener's *perception* of the segments (as sounding 'longer' or 'shorter') — can tease apart these possibilities.

In addition to word duration, the main model also revealed differences in f0 variation overall when participants were shadowing the Alexa voice, relative to the human voice. While at first glance these differences appear to reflect systematic differences in Alexa-directed speech (as seen for overall word duration), post hoc analyses provide evidence that these changes reflect *acoustically driven alignment* since f0 variation only varied by model talker in the Neutral subset (not Expressive). This aligns with the acoustic differences in the stimuli wherein the Alexa voice had greater f0 variation in neutral conditions (2.8 ST), compared to the human (1.5 ST).

To examine sources of a possible *socially-mediated* emotional alignment response (human vs. device social categories), we compared speakers' adjustments for the two interlocutors' emotional expressive productions. In response to emotional expressiveness, we find that speakers adapt their speech in similar directions for the human and voice-AI interlocutors, with small differences in magnitude. While it is possible these differences could be a socially-mediated effect (as argued in related work; e.g., Cohn et al., 2019; Snyder et al., 2019), as mentioned in the Interim Discussion (Section 4), the degree of difference can be explained by acoustic differences between 'neutral' and 'expressive' productions by the Alexa and human model talkers. In cases where the Alexa model talker has a larger difference from neutral-to-expressive (e.g., word duration), participants show larger increases toward Alexa Expressive productions. The converse was also true: when acoustic differences are smaller for the Alexa voice from neutral-to-expressive (e.g., mean f0, f0 variation), participants show *weaker* increases toward Alexa Expressive productions. Thus, we interpret these model talker-based differences in emotionality as driven by the acoustics (rather than a difference in social category). This is

Table 7
Summary of findings.

	Expressiveness Condition	Model Talker Category
Duration	Increases for Expressive ● larger increase toward Alexa	Shorter for Alexa (post hoc: shorter in both Neural and Expressive)
Mean f0	Increases for Expressive ● smaller increase toward Alexa ● larger increase by Female participants	No difference
F0 variation	Increases for Expressive ● smaller increase toward Alexa ● larger increase toward Alexa by Female participants (post hoc: larger toward Alexa Expressive by Female participants)	Larger toward Alexa (post hoc: larger toward Alexa Neutral)

consistent with *unmediated alignment accounts* of emotional alignment (e.g., De Waal, 2007; Decety & Jackson, 2006; Preston, 2007), wherein participants are simply aligning toward changes in acoustic features as they are realized in the stimuli.

More generally, observing overall increases in prosodic features associated with 'positive-emotional' speech for *both* human and Alexa voices supports computer personification theories (CASA: Nass et al., 1997, 1994). Here, people appear to apply human-human speech behaviors in response to emotional expressiveness by a non-human entity: voice-AI. These results are in line with prior work that described similar responses for emotional/affective behaviors in human-human and human-computer interaction (Brave et al., 2005; Bucci et al., 2018; Nass et al., 1999, 1995; Vaughan et al., 2018; Xiao et al., 2013). It also supports prior work that participants' alignment toward virtual interlocutors is, in part, an automatic behavior (Staum Casasanto et al., 2010). While one possibility we raised was that participants might find the emotionally expressive Alexa voice to be 'uncanny' (Mori, 1970; Mori et al., 2012) and diverge from it, we did not find evidence to support that.

At the same time, we observed several differences by speaker gender in emotional alignment: women show larger increases in mean f0 toward expressiveness (overall) and in f0 variation (toward Alexa). These increases might reflect differences in socialization, where women display stronger 'emotional contagion' (Doherty et al., 1995; Sonnby-Borgström et al., 2008), consistent with *socially-mediated accounts* of emotional alignment (Hess and Fischer, 2013, 2014; Fischer et al., 2019). Broadly, observing greater alignment by female participants (whether due to socialization and/or acoustic tracking) is in line with prior work in the (non-emotional) vocal alignment literature (Arimoto & Okanoya, 2014; Namy et al., 2002). At the same time, it contrasts with recent work reporting greater alignment by male speakers (than female speakers) toward emotionally expressive Alexa productions (Cohn & Zellou, 2019). Why might this be the case? One possibility is that women might produce more pitch-based adjustments (here, mean f0 and f0 variation) to align to an interlocutor. Furthermore, it is possible that listener's *perception* of alignment (e.g., using AXB in Cohn & Zellou, 2019) might differ for speaker gender (Babel & Bulatov, 2012); raters might perceive 'more' alignment by interlocutors whose baseline differences start farther away (here, males with lower f0 converging toward female f0), parallel to arguments in vocal alignment that speakers with larger baseline distances have more 'room' to converge (Babel, 2010; Walker & Campbell-Kibler, 2015; but see Cohen Priva & Sanker, 2019; MacLeod, 2021).

Taken together, our results suggest a nuanced picture of emotional alignment. In general, we find support for *unmediated, motor accounts* (e.g., De Waal, 2007; Decety & Jackson, 2006; Preston, 2007), where speakers 'match' the acoustic input, in responses to emotionally expressive model talkers. At the same time, we see some possible support for *socially-mediated accounts* of emotional alignment (Hess and Fischer, 2013, 2014; Fischer et al., 2019) in the domain of speaker gender. The present examination of emotional vocal alignment, while novel, has a number of limitations that can set up many directions for future research. While one of the innovations of the present study is the comparison of emotional alignment across human and voice-AI interlocutors, this focus limited the number of model talkers. At the time of the study, only the Amazon Alexa default female voice was capable of producing both neutral and emotionally expressive productions in US English. This default TTS voice was likely to be familiar to participants, as most (51/66) had prior experience with Amazon's Alexa specifically. Familiarity may mediate alignment toward emotion and also perhaps why we find that it is largely comparable toward the human and the Alexa across acoustic features. Future work examining more model talkers (e.g., varying in gender, 'recognizability', etc.) can uncover the extent to which these effects generalize to more voices. In particular, recent work has pointed to a large degree of idiosyncratic variation across speakers, some of which would also likely be present among different TTS voices (Lee et al., 2019).

Additionally, the present study used both audio and visual cues to cue the model talker categories. While the aim was to provide clear guise information (such that it was unambiguous that the talker was a human or device), there is a body of work showing visual cues shape auditory perception (Babel & Russell, 2015; D'Onofrio, 2019; Hay et al., 2006; Zellou et al., 2020). Recently, there is also work showing differences in vocal alignment based on physical form: speakers show stronger vocal alignment toward TTS voices when they are presented with a more human-like form (e.g., Furhat or Nao robot) relative to a form that lacks human features (e.g., Amazon Echo) (Cohn, Jonell, et al., 2020). In the current study, the visual information for the human (a smiling female) might have provided stronger emotion-congruent information with the positive-valence stimuli (e.g., "Awesome!"). There is related work showing processing costs when cues of emotion conflict: for example, Nygaard & Queen (2008) found word-naming latencies when the word's meaning and how it was spoken conflicted (e.g., 'happy' word produced with 'sad' prosody). The extent to which emotional mismatch might shape vocal alignment — and vary for different types of interlocutors (human vs. device) — remain avenues for future work.

Another direction for future work is to examine additional sources of socially-mediated variation, including language background and cultural attitudes toward voice-AI systems that might influence emotional vocal alignment. There is some work, for example, demonstrating that emotional expressiveness varies cross-linguistically and cross-culturally (Abelin & Allwood, 2000; Batliner et al., 2004). Furthermore, future work examining individual differences in response to emotion — such as in vocal alignment — can further probe the cognitive and social dynamics of human-computer interaction (HCI). Broadly, understanding individual variation in HCI is important for developing comprehensive models of human behavior toward AI, and addressing a gap in the HCI literature, where fewer studies have examined individual differences in participants' vocal interactions with technology (for a review, see Snyder et al., 2019), as well as for possible practical applications. There is already some work suggesting that depressed patients' speech with interactive voice response (IVR) technology can track their recovery (Mundt et al., 2007); this suggests that biomarkers in speech toward voice-AI might be useful in clinical applications.

7. Conclusion

Overall, this study sheds light on the underlying mechanisms of emotional vocal alignment: even in a laboratory setting, people align toward the positive-emotional speech style they hear. Here, the social category of the talker — as a human or device — did not serve as a social factor guiding emotional alignment. Rather, we see that magnitude of acoustic difference (from neutral-to-expressive) can explain the small differences in alignment toward the Alexa and human model talkers. Observing a similar response for these categories further supports computer personification accounts: people appear to apply similar *emotional* behaviors from human-human interactions to speech interactions with voice-AI. While more work is needed to test the extent of this overlap, this raises many important scientific questions as to the nature of anthropomorphization, and can serve practical applications in voice user interface design.

CRediT authorship contribution statement

Michelle Cohn: Conceptualization, Formal analysis, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Kristin Predeck:** Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Melina Sarian:** Data curation, Writing – original draft, Writing – review & editing. **Georgia Zellou:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

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

Acknowledgments

Thank you to editor Dellwo and two anonymous reviewers for their guidance and feedback on the manuscript. Thank you also to our undergraduate research assistants, who helped collect data on the project: Patricia Sandoval, Eleanor Lacaze, and Maria Carroll. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship to MC under Grant No. 1911855. Additionally, this work was partially supported by a 2019 Amazon Faculty Research Award to GZ.

References

Abadjieva, E., Murray, I.R., & Arnott, J.L. (1993). Applying analysis of human emotional speech to enhance synthetic speech. Third European Conference on Speech Communication and Technology.

Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Abrego-Collier, C., Grove, J., Sonderegger, M., Alan, C.L., 2011. Effects of Speaker Evaluation on Phonetic Convergence. ICPhS 192–195.

Amazon. (2018). Speechcon Reference (Interjections): English (US) | Custom Skills. https://developer.amazon.com/docs/custom-skills/speechcon-reference-interjections-english-us.html.

Ameka, F., 1992. Interjections: The universal yet neglected part of speech. Journal of Pragmatics 18 (2-3), 101–118.

Ammari, T., Kaye, J., Tsai, J.Y., Bentley, F., 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. ACM Transactions on Computer-Human Interaction (TOCHI) 26 (3), 1–28.

Arias, P., Belin, P., Aucouturier, J.-J., 2018. Auditory smiles trigger unconscious facial imitation. Current Biology 28 (14). R782-R783.

Arimoto, Y., & Okanoya, K. (2014). Emotional synchrony and covariation of behavioral/physiological reactions between interlocutors. 2014 17th Oriental Chapter of the International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA), 1-6.

Babel, M., 2010. Dialect divergence and convergence in New Zealand English. Language in Society 39 (4), 437–456.

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

Babel, M., Bulatov, D., 2012. The role of fundamental frequency in phonetic accommodation. Language and Speech 55 (2), 231–248.

Babel, M., Russell, J., 2015. Expectations and speech intelligibility. The Journal of the Acoustical Society of America 137 (5), 2823–2833.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M.J., & Wong, M. (2004). "You Stupid Tin Box"-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. Lrec.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., Lottridge, D., 2018. Understanding the long-term use of smart speaker assistants. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2 (3), 1–24.

Boril, T., Sturm, P., Skarnitzl, R., Volin, J., 2017. Effect of formant and F0 discontinuity on perceived vowel duration: Impacts for concatena- tive speech synthesis. Proceedings of Interspeech 2998–3002.

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.

Brave, S., Nass, C., Hutchinson, K., 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. International Journal of Human-Computer Studies 62 (2), 161–178. https://doi.org/10.1016/j.ijhcs.2004.11.002.

Bucci, P., Zhang, L., Cang, X.L., & MacLean, K.E. (2018). Is it Happy? Behavioural and Narrative Frame Complexity Impact Perceptions of a Simple Furry Robot's Emotions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-11.

Burnham, D.K., Joeffry, S., & Rice, L. (2010). Computer-and human-directed speech before and after correction. Proceedings of the 13th Australasian International Conference on Speech Science and Technology, 13-17. http://handle.uws.edu.au:8081/1959.7/504796.

Cohen Priva, U., Sanker, C., 2019. Limitations of difference-in-difference for measuring convergence. Laboratory Phonology: Journal of the Association for Laboratory Phonology 10 (1).

Cohn, M., Chen, C.-Y., & Yu, Z. (2019). A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog. Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, 293-306. https://www.sigdial.org/files/workshops/conference20/proceedings/cdrom/pdf/W19-5935.pdf.

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. https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_202.pdf.

Cohn, M., Jonell, P., Kim, T., Beskow, J., & Zellou, G. (2020). Embodiment and gender interact in alignment to TTS voices. Proceedings of the Cognitive Science Society, 220-226. https://cogsci.mindmodeling.org/2020/papers/0044/0044.pdf.

Cohn, M., Liang, K.-H., Sarian, M., Zellou, G., Yu, Z., 2021. Speech Rate Adjustments in Conversations With an Amazon Alexa Socialbot. Frontiers in Communication 6, 1–8. https://doi.org/10.3389/fcomm.2021.671429.

Cohn, M., Raveh, E., Predeck, K., Gessinger, I., Möbius, B., Zellou, G., 2020. Differences in Gradient Emotion Perception: Human vs. Alexa Voices. Proc. Interspeech 2020, 1818–1822.

Cohn, M., & Zellou, G. (2019). Expressiveness influences human vocal alignment toward voice-AI. Proc. Interspeech 2019, 41-45. https://doi.org/10.21437/Interspeech.2019-1368.

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. International Journal of Human-Computer Studies 83, 27–42.

De Waal, F.B. (2007). The 'Russian doll'model of empathy and imitation. On Being Moved: From Mirror Neurons to Empathy, 35-48.

Decety, J., Jackson, P.L., 2006. A social-neuroscience perspective on empathy. Current Directions in Psychological Science 15 (2), 54–58.

DiCanio, C. (2007). Extract Pitch Averages. https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_pitch.praat.

Dijksterhuis, A., & Bargh, J.A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In Advances in experimental social psychology (Vol. 33, pp. 1-40). Elsevier.

Doherty, R.W., Orimoto, L., Singelis, T.M., Hatfield, E., Hebb, J., 1995. Emotional Contagion: Gender and Occupational Differences. Psychology of Women Quarterly 19 (3), 355–371. https://doi.org/10.1111/j.1471-6402.1995.tb00080.x.

D'Onofrio, A., 2019. Complicating categories: Personae mediate racialized expectations of non-native speech. Journal of Sociolinguistics 23 (4), 346–366. https://doi.org/10.1111/josl.12368.

Fischer, A.H., Pauw, L.S., & Manstead, A.S.R. (2019). Emotion Recognition as a Social Act: The Role of the Expresser-Observer Relationship in Recognizing Emotions. In U. Hess & S. Hareli (Eds.), The Social Nature of Emotion Expression: What Emotions Can Tell Us About the World (pp. 7-24). Springer International Publishing. https://doi.org/10.1007/978-3-030-32968-6_2.

Fuller, R.G.C., Sheehy-Skeffington, A., 1974. Effects of Group Laughter on Responses to Humourous Material, a Replication and Extension. Psychological Reports 35 (1), 531–534. https://doi.org/10.2466/pr0.1974.35.1.531.

Gazzola, V., Rizzolatti, G., Wicker, B., Keysers, C., 2007. The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. Neuroimage 35 (4), 1674–1684.

Giles, H., & Baker, S.C. (2008). Communication accommodation theory. The International Encyclopedia of Communication.

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

Goffman, E., 1981. Response cries. Forms of talk. University of Pennsylvania Press, pp. 78-122.

Goldinger, S.D., 1996. Words and voices: episodic traces in spoken word identification and recognition memory. Journal of experimental psychology: Learning, memory, and cognition 22 (5), 1166.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychological Review 105 (2), 251.

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

Hess, U., Fischer, A., 2013. Emotional mimicry as social regulation. Personality and Social Psychology Review 17 (2), 142–157.

Hess, U., Fischer, A., 2014. Emotional mimicry: Why and when we mimic emotions. Social and Personality Psychology Compass 8 (2), 45–57.

Lakin, J.L., Jefferis, V.E., Cheng, C.M., Chartrand, T.L., 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. Journal of Nonverbal Behavior 27 (3), 145–162.

Lee, Y., Keating, P., Kreiman, J., 2019. Acoustic voice variation within and between speakers. The Journal of the Acoustical Society of America 146 (3), 1568–1579.

Liu, B., Sundar, S.S., 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychology, Behavior, and Social Networking 21 (10), 625–636.

MacLeod, B., 2021. Problems in the Difference-in-Distance measure of phonetic imitation. Journal of Phonetics 87, 101058. https://doi.org/10.1016/j.wocn.2021.101058.

Matsumoto, D. (2002). Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence.

Mori, M., 1970. Bukimi no tani [the uncanny valley]. Energy 7, 33–35.

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

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via

interactive voice response (IVR) technology. Journal of Neurolinguistics 20 (1), 50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001.
Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America 93 (2), 1097–1108.
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., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. CHI'05 Extended Abstracts on Human Factors in Computing Systems, 1973-1976.
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., Fogg, B.J., Reeves, B., & Dryer, C. (1995). Can computer personalities be human personalities?, Conference Companion on Human Factors in Computing Systems, 228-229.
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. https://doi.org/10.1145/259963.260288.
Nielsen, K., 2011. Specificity and abstractness of VOT imitation. Journal of Phonetics 39 (2), 132–142.
Nygaard, L.C., Queen, J.S., 2008. Communicating emotion: Linking affective prosody and word meaning. Journal of Experimental Psychology: Human Perception and Performance 34 (4), 1017.
Oviatt, S., MacEachern, M., Levow, G.-A., 1998. Predicting hyperarticulate speech during human-computer error resolution. Speech Communication 24 (2), 87–110.
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., Gibbons, R., Suppes, A., Krauss, R.M., 2012. Phonetic convergence in college roommates. Journal of Phonetics 40 (1), 190–197.
Pardo, J.S., Jay, I.C., Krauss, R.M., 2010. Conversational role influences speech imitation. Attention, Perception, & Psychophysics 72 (8), 2254–2264.
Preston, S.D. (2007). A perception-action model for empathy. Empathy in Mental Illness, 428-447.
Raveh, E., Siegert, I., Steiner, I., Gessinger, I., Möbius, B., 2019. Three's a Crowd? Effects of a Second Human on Vocal Accommodation with a Voice Assistant. Proc. Interspeech 4005–4009. https://doi.org/10.21437/Interspeech.2019-1825, 2019.
Raveh, E., Steiner, I., Siegert, I., Gessinger, I., & Möbius, B. (2019). Comparing phonetic changes in computer-directed and human-directed speech. Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019, 42-49.
Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (forced alignment and vowel extraction) program suite. URL Http://Fave. Ling. Upenn. Edu.
Scherer, S., Hammal, Z., Yang, Y., Morency, L.-P., & Cohn, J.F. (2014). Dyadic behavior analysis in depression severity assessment interviews. Proceedings of the 16th International Conference on Multimodal Interaction, 112-119.
Scherer, S., Lucas, G.M., Gratch, J., Rizzo, A.S., Morency, L.-P., 2016. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Transactions on Affective Computing 1, 59–73.
Shepard, C.A., 2001. Communication accommodation theory. The new handbook of language and social psychology. John Wiley & Sons, Ltd (W. P. Robinson, H. Giles, pp. 33-56).
Siegert, I., & Krüger, J. (2021). "Speech Melody and Speech Content Didn't Fit Together"—Differences in Speech Behavior for Device Directed and Human Directed Interactions. In Advances in Data Science: Methodologies and Applications (1st ed., Vol. 189, pp. 65-95). Springer. https://doi.org/10.1007/978-3-030-51870-7_4.
Siegert, I., Nietzold, J., Heinemann, R., & Wendemuth, A. (2019). The restaurant booking corpus-content-identical comparative human-human and human-computer simulated telephone conversations. Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019, 126-133.
Smith, C. (2007). Prosodic accommodation by French speakers to a non-native interlocutor. Proceedings of the XVIth International Congress of Phonetic Sciences, 313-348.
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.
Sonnby-Borgström, M., Jönsson, P., Svensson, O., 2008. Gender differences in facial imitation and verbally reported emotional contagion from spontaneous to emotionally regulated processing levels. Scandinavian Journal of Psychology.
Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. 32nd Annual Meeting of the Cognitive Science Society (CogSci 2010), 127-132.
Thibault, P., Bourgeois, P., Hess, U., 2006. The effect of group-identification on emotion recognition: The case of cats and basketball players. Journal of Experimental Social Psychology 42 (5), 676–683.
Van Der Schalk, J., Fischer, A., Doosje, B., Wigboldus, D., Hawk, S., Rotteveel, M., Hess, U., 2011. Convergent and divergent responses to emotional displays of ingroup and outgroup. Emotion 11 (2), 286.
Vaughan, B., De Pasquale, C., Wilson, L., Cullen, C., & Lawlor, B. (2018). Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients. International Symposium on Pervasive Computing Paradigms for Mental Health, 150-159. https://doi.org/10.1007/978-3-030-01093-5_19.
Viscovich, N., Borod, J., Pihan, H., Peery, S., Brickman, A.M., Tabert, M., Schmidt, M., Spielman, J., 2003. Acoustical Analysis of Posed Prosodic Expressions: Effects of Emotion and Sex. Perceptual and Motor Skills 96 (3), 759–771. https://doi.org/10.2466/pms.2003.96.3.759.
Wade, L., 2020. The Linguistic and the Social Intertwined: Linguistic Convergence Toward Southern Speech. Dissertation.
Walker, A., Campbell-Kibler, K., 2015. Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. Frontiers in Psychology 6. https://doi.org/10.3389/fpsyg.2015.00546.
Weisbuch, M., Ambady, N., 2008. Affective divergence: Automatic responses to others' emotions depend on group membership. Journal of Personality and Social Psychology 95 (5), 1063.
Xiao, B., Georgiou, P.G., Imel, Z.E., Atkins, D.C., Narayanan, S., 2013. Modeling therapist empathy and vocal entrainment in drug addiction counseling. Interspeech 2861–2865.
Xiao, B., Imel, Z.E., Atkins, D.C., Georgiou, P.G., & Narayanan, S.S. (2015). Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. Sixteenth Annual Conference of the International Speech Communication Association.
Yang, Y., Fairbairn, C., Cohn, J.F., 2013. Detecting Depression Severity from Vocal Prosody. IEEE Transactions on Affective Computing 4 (2), 142–150. https://doi.org/10.1109/T-AFFC.2012.38.
Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Deng, Z., Lee, S., Narayanan, S., & Busso, C. (2004). An acoustic study of emotions expressed in speech. Eighth International Conference on Spoken Language Processing.
Yu, A.C.L., Abrego-Collier, C., Sonderegger, M., 2013. Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and "Autistic" Traits. PLOS ONE 8 (9), e74746. https://doi.org/10.1371/journal.pone.0074746.
Zając, M., 2013. Phonetic imitation of vowel duration in L2 speech. Research in Language 11 (1), 19–29.
Zellou, G., Cohn, M., 2020. Social and functional pressures in vocal alignment: Differences for human and voice-AI interlocutors. Proc. Interspeech 2020, 1634–1638 https://doi.org/10.21437/Interspeech.2020-1335.
Zellou, G., Cohn, M., Block, A., 2020. Does top-down information about speaker age guise influence perceptual compensation for coarticulatory/u/-fronting? Cognitive Science Society 3483–3489.