Prosodic differences in human- and Alexa-directed speech, but similar local intelligibility adjustments

Michelle Cohn[1] and Georgia Zellou[1]

[1]Phonetics Lab, University of California, Davis, Davis, CA United States

Abstract

The current study tests whether individuals (n=53) produce distinct speech adaptations during pre-scripted spoken interactions with a voice-AI assistant (Amazon's Alexa) relative to those with a human interlocutor. Interactions crossed intelligibility pressures (staged word misrecognitions) and emotionality (hyper-expressive interjections) as conversation-internal factors that might influence participants' intelligibility adjustments in Alexa- and human-directed speech (DS). Overall, we find speech style differences: Alexa-DS has a decreased speech rate, higher mean f0, and greater f0 variation than human-DS. In speech produced toward both interlocutors, adjustments in response to misrecognition were similar: participants produced more distinct vowel backing (enhancing the contrast between the target word and misrecognition) in target words, and louder, slower, and higher mean f0, and higher f0 variation at the sentence-level. No differences were observed in human- and Alexa-DS following displays of emotional expressiveness by the interlocutors. Expressiveness, furthermore, did not mediate intelligibility adjustments in response to a misrecognition. Taken together, these findings support proposals that speakers presume voice-AI has a 'communicative barrier' (relative to human interlocutors), but that speakers adapt to conversational-internal factors of intelligibility similarly in human- and Alexa-DS. This work contributes to our understanding of human-computer interaction, as well as theories of speech style adaptation.

Keywords: voice-activated artificially intelligent (voice-AI) assistant, speech register, intelligibility, human-computer interaction, computer personification

1. Introduction

People dynamically adapt their speech according to the communicative context and (apparent) barriers present. In the presence of background noise, for example, speakers produce speech that is louder, slower, and higher pitched ('Lombard speech') (for a review, see Brumm & Zollinger, 2011), argued by some to be an automatic, non-socially mediated response (Junqua, 1993, 1996). Other work has shown that people adapt their speech to the type of *listener* they are engaging with. One stance is that speakers presume certain types of interlocutors to have greater communicative barriers (Branigan et al., 2011; Clark, 1996; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998). Supporting this account, prior work has shown that people use different speech styles when talking to non-native speakers (Hazan et al., 2015; Scarborough et al., 2007; Uther et al., 2007), hearing impaired adults (Knoll et al., 2015; Picheny et al., 1985; Scarborough & Zellou, 2013), and computers (Bell et al., 2003; Bell & Gustafson, 1999; Burnham et al., 2010; Lunsford et al., 2006; Mayo et al., 2012; Oviatt, Levow, et al., 1998; Oviatt, MacEachern, et al., 1998; Siegert et al., 2019; Stent et al., 2008). For example, computer-directed speech (DS) has been shown to be

louder (Lunsford et al., 2006), with durational lengthening (Burnham et al., 2010; Mayo et al., 2012), greater vowel space expansion (Burnham et al., 2010), and smaller pitch range (Mayo et al., 2012) than speech directed to a (normal hearing adult) human.

This paper explores whether speakers use a specific speech style (or 'register') when talking to a voice-activated artificially intelligent (voice-AI) assistant. Voice-AI assistants (e.g., Amazon's Alexa, Apple's Siri, Google Assistant) are now a common interlocutor for millions of individuals completing everyday tasks (e.g., "set a timer for 5 minutes", "turn on the lights", etc.) (Ammari et al., 2019; Bentley et al., 2018). A growing body of research has begun to investigate the social, cognitive, and linguistic effects of humans interacting with voice-AI (Arnold et al., 2019; Burbach et al., 2019; Cohn, Ferenc Segedin, et al., 2019; Purington et al., 2017). For example, recent work has shown that listeners attribute human-like characteristics to the text-to-speech (TTS) output used for modern voice-AI, including personality traits (Lopatovska, 2020), apparent age (Cohn, Jonell, et al., 2020; Zellou et al., 2021), and gender (Habler et al., 2019; Loideain & Adams, 2020). While the spread of voice-AI assistants is undeniable — particularly in the United States — there are many open scientific questions as to the nature of people's interactions with voice-AI.

There is some evidence for a different speech style used in interactions with 'voice-AI' assistants: several studies have used classifiers to successfully identify 'device-' and 'non-device-'directed speech from users' interactions with Amazon Alexa (Huang et al., 2019; Mallidi et al., 2018). Yet, in these cases, the linguistic content, physical distance from the device, and other factors were not controlled and might have contributed to differences that are not speech-style adaptations per se. Critically, holding the interaction constant across a voice-AI and human interlocutor can reveal if individuals have a distinct voice-AI speech style. Some groups have aimed to compare human and voice-AI speech styles in more similar contexts. For instance, the Voice Assistant Conversation Corpus (VACC) had participants complete the same type of communicative task (setting an appointment on a calendar and doing a quiz) with an Alexa Echo and a real human confederate (Siegert et al., 2018). Several studies measuring the acoustic-phonetic features of human- and Alexa-DS in the corpus found productions toward Alexa were louder (Raveh et al., 2019; Siegert & Krüger, 2020), higher in fundamental frequency (f0, perceived pitch) (Raveh et al., 2019), and contained different vowel formant characteristics[1] (Siegert & Krüger, 2020). Yet, similar to studies of individuals using Alexa in their homes (e.g., Huang et al., 2019), differences observed in the VACC might also be driven by physical distance from the device and conversational variations. The current study holds context and physical distance from the microphone constant for the two interlocutors to address these limitations in prior work.

Making a direct human- and Alexa-DS comparison in a scripted task can speak to competing predictions across different computer personification accounts: if speech styles differ because speakers have a 'routinized' way of talking to computers (in line with *routinized interaction accounts*) or if speech styles are the same (in line with *technology equivalence accounts*). *Routinized interaction accounts* propose that people have a 'routinized' way of interacting with technological systems (Gambino et al., 2020), borne out of real experience with the systems, as well as a priori expectations. As mentioned, there is ample evidence for a computer-DS register (e.g., Bell et al., 2003; Bell & Gustafson, 1999; Burnham et al., 2010). Specifically, some propose that the computer faces additional communicative barriers, relative to humans

[1] They do not report a directionality of difference.

(Branigan et al., 2011; Oviatt, MacEachern, et al., 1998). These attitudes appear to be a priori, developed before any evidence of communicative barriers in an interaction. For example, people rate TTS voices as 'less communicatively competent' (Cowan et al., 2015). Therefore, one prediction for the current study is that speakers might have overall different speech styles in human- and Alexa-DS, reflecting this presumed communicative barrier and a 'routinized' way of talking to voice-AI.

Technology equivalence accounts, on the other hand, propose that people automatically and subconsciously apply social behaviors from human-human interaction to their interactions with computer systems (e.g., Lee, 2008). For example, 'Computers are Social Actors' (CASA) (Nass et al., 1997, 1994) specifies that this transfer of behaviors from human-human interaction is triggered when people detect a 'cue' of humanity in the system, such as engaging with a system using language. For example, people appear to apply politeness norms from human-human interaction to computers: giving more favorable ratings when a computer directly asks about its own performance, relative to when a different computer elicits this information (Hoffmann et al., 2009; Nass et al., 1994). In line with *technology equivalence accounts*, there is some evidence for applied social behaviors to voice-AI in the way people adjust their speech, such as gender-mediated vocal alignment (Cohn, Ferenc Segedin, et al., 2019; Zellou et al., 2021). In the present study, one prediction from *technology equivalence accounts* is that people will adjust their speech patterns when talking to voice-AI and humans in similar ways if the communicative context is controlled.

1.1. Different strategies to improve intelligibility following a misrecognition?

To probe *routinized interaction* and *technology equivalence* accounts, the present study further investigates if speakers adapt their speech differently after a human or a voice-AI assistant 'mishears' them. There is evidence that speakers monitor communicative pressures during an interaction, varying their acoustic-phonetic output to improve intelligibility when there is evidence listeners might mishear them (Smiljanić & Bradlow, 2009; Hazan & Baker, 2011). Lindblom's (1990) Hyper- & Hypo-articulation (H&H) model proposes a real-time trade-off between speakers' needs (i.e., to preserve articulatory effort) and listeners' needs (i.e., to be more intelligible). While the majority of prior work examining speakers' adaptations following a computer misrecognition has lacked a direct human comparison, many of the adjustments parallel those observed in human-human interaction; for example, speakers produce louder and slower speech after a dialog system conveys that it 'heard' the wrong word (Bell & Gustafson, 1999; Oviatt, Levow, et al., 1998; Swerts et al., 2000). Additionally, some studies report vowel adaptations in response to a misunderstanding that are consistent with enhancements to improve intelligibility, including vowel space expansion (Bell & Gustafson, 1999; Maniwa et al., 2009) and increase in formant frequencies (Vertanen, 2006). There is also evidence of targeted adjustments: speakers produce more vowel-specific expansion (e.g., high vowels produced higher) in response to misrecognitions by a dialog system (Stent et al., 2008). Will speakers use different strategies to improve intelligibility following a staged word misrecognition based on who their listener is? One possibility is that speakers might have a 'routinized' way of improving their intelligibility following a misrecognition made by a voice-AI assistant, which would support *routinized interaction accounts*. At the same time, Burnham et al. (2010) found no difference between speech adjustments post-misrecognition for an (apparent) human and digital avatar, but only more global differences for the computer interlocutor (i.e., speech with longer segmental durations and with

greater vowel space expansion). Therefore, it is possible that speakers will produce similar intelligibility adjustments in response to a staged misrecognition made by either a voice-AI or human listener, supporting *technology equivalence accounts*.

Additionally, the current study adds a novel manipulation in addition to intelligibility pressures: emotional expressiveness. When an interlocutor 'mishears', they might be disappointed and express it (e.g., "Darn! I think I misunderstood."); when they get it correct, they might be enthusiastic and convey that in their turn (e.g., "Awesome! I think I heard boot."). Emotional expressiveness is a common component of naturalistic human conversations, providing a window into how the listener is feeling (Ameka, 1992; Goffman, 1981). This 'socio-communicative enhancement' might increase the pressure for speakers to adapt their speech for the listener. On the one hand, this enhanced emotional expressiveness might result in even more similar adjustments for voice-AI and human interlocutors, since adding expressiveness might increase the perception of human-likeness for the device, which could strengthen *technology equivalence*. Indeed, there is some work to suggest that emotional expressiveness in a computer system is perceived favorably by users. For instance, Brave and colleagues (2005) found when computer systems expressed empathetic emotion, they were rated more positively. For voice-AI, there is a growing body of work testing how individuals perceive emotion in TTS voices (Cohn, Chen, et al., 2019; Cohn, Jonell, et al., 2020). For example, an Amazon Alexa Prize socialbot was rated more positively if it used emotional interjections (Cohn, Chen, et al., 2019). Alternatively, the presence of emotionality might lead to distinct clear speech strategies for the human and voice-AI interlocutors. For example, a study of phonetic alignment (using the same corpus in the current study) found that vowel duration alignment differed both by the social category of interlocutor (human vs. voice-AI) and based on emotionality (Zellou & Cohn, 2020): participants aligned more in response to a misrecognition, consistent with H&H theory (Lindblom, 1990), which increased even more when the voice-AI talker was emotionally expressive when conveying their misunderstanding (e.g., "Bummer! I'm not sure I understood. I think I heard sock or sack."). Still, that study examined just one acoustic difference in speech behavior (vowel duration alignment). The present study investigates whether emotionality similarly mediates targeted speech adjustments to voice-AI, an underexplored research question.

1.2. Current study

The present study examines a corpus of speech directed at a human and voice-AI interlocutor which crossed intelligibility factors (staged misrecognitions) and emotionality of the interlocutor's responses in identical pre-scripted tasks (Zellou & Cohn, 2020). This is the first study, to our knowledge, to test both intelligibility and emotional expressiveness factors in speech style adaptations for a voice-AI assistant and human. Here, the Amazon Alexa voice (US-English, female) was selected for its ability to generate emotionally expressive phrases recorded by the voice actor, common in Alexa Skills Kit apps ('Speechcons'). To determine overall differences between Alexa- and human-DS, as well as more local intelligibility adjustments in response to a staged misrecognition, we measure several acoustic features associated with computer-DS and/or 'clear' speech: intensity, speech rate, mean f0, f0 variation, and vowel formant characteristics (F1, F2).

2. Methods

2.1. Participants

Data were taken from a corpus (Zellou & Cohn, 2020) containing 53 native English speaking participants (27 female, 26 male; mean age of 20.28 years old, sd = 2.42 years; range: 18-34) talking to a voice-AI and human interlocutor in an identical interactive task. None reported having any hearing impairment. Nearly all participants (*n*=49) reported using a voice-AI system: Alexa (*n*=35), Siri (*n*=13), Google Assistant (*n*=1). Participants were recruited from the UC Davis psychology subjects pool and completed informed consent, in pursuance with the UC Davis Institutional Review Board (IRB).

2.2. Target words

Sixteen target words, presented in Table 1, were selected from Babel (2012) who had chosen the items for being low frequency in American English; higher frequency items have been shown to be more phonetically reduced in production (e.g., Pluymaekers et al., 2005). Target words were all CVC words containing either /i, æ, u, ow, a/ and a word-final obstruent (e.g., /z/, /p/) (a subset of the words used in Babel, 2012). In addition, we selected a real-word vowel minimal pair, differing in vowel backness, to be used in the interlocutor responses in the misrecognition condition.

Table 1. Target words and their (minimal pairs) used in the experiment dialogue.

bat (boat)	boot (beat)	cheek (choke)	coat (Kate)
cot (cat)	deed (dude)	dune (dean)	hoop (heap)
moat (meet)	pod (pad)	soap (seep)	sock (sack)
tap (top)	toot (teat)	tot (tat)	weave (wove)

2.3. Interlocutor Recordings

The human and voice-AI interlocutor responses were pre-recorded. For the human, a female native California English speaker recorded responses in a sound attenuated booth, with a head-mounted microphone (Shure WH20 XLR). The Alexa productions were generated with the default female Alexa voice (US-English) with the Alexa Skills Kit. Both interlocutors generated introductions ("Hi! I'm Melissa. I'm a research assistant in the Phonetics Lab." / "Hi! I'm Alexa. I'm a digital device through Amazon.") and voice-over instructions for the task. We recorded each interlocutor producing two responses for each target word: a 'correctly understood' response ("I think I heard bat") and an 'misrecognition' response ("I'm not sure I understood. I think I heard bought or bat."). Figure 1 provides an example of the different interlocutor responses. Order of target word and misheard word was counterbalanced across sentences, such that the 'correct' word did not always occur in the same position in these response types.

Both interlocutors generated 16 emotionally expressive interjections as well: 8 positive interjections (*bam, bingo, kapow, wahoo, zing, awesome, dynamite, yipee*) and 8 negative interjections (*argh, baa, blarg, oof, darn, boo, oy, ouch*) selected from the Speechcons website[2] at the time of the study. We generated these interjections for the Alexa text-to-speech (TTS) output

[2] https://developer.amazon.com/en-US/docs/alexa/custom-skills/speechcon-reference-interjections-english-us.html

using synthesis markup language (SSML) tags. The human produced these interjections in an expressive manner (independently, not imitating the Alexa productions). We randomly assigned each interjection to the interlocutor responses, matching in whether the response was correctly understood (positive interjection) or misunderstood (negative interjection). The full set of interjections was used twice in each block (e.g., 8 positive interjections randomly concatenated to 16 correct productions). The full set of interlocutor productions are available on Open Science Framework[3].

2.4. Procedure

Participants completed the experiment while seated in a sound-attenuated booth, wearing a head-mounted microphone (Shure WH20 XLR) and headphones (Sennheiser Pro), and facing a computer screen. First, we collected citation forms of the target words produced in sentences. Participants read the word in a sentence ("The word is bat.") presented on the screen. Target words were presented randomly.

Following the Citation block, participants completed identical experimental blocks with both a human talker and an Alexa talker (block order counterbalanced across subjects). First, the interlocutor introduced themselves and then went through voice-over instructions with the participant. Participants saw an image corresponding to the interlocutor category: stock images of 'adult female' (used in prior work; Zellou et al., 2021) and 'Amazon Alexa' (2nd Generation Black Echo).

Each trial consisted of 4 turns. Participants first read a sentence aloud containing the target word sentence-finally (e.g., "The word is bat."). Then, the interlocutor responded in one of four possible Staged Misunderstanding (correctly heard/misrecognition) and Emotionality (neutral/expressive) Conditions (see Figure 1). Next, the participant responded to the interlocutor by repeating the sentence (e.g., "The word is bat."). This is the response that we acoustically analyze. Finally, the interlocutor provides a confirmation, randomized ("Thanks", "Perfect", "Okay", "Uh huh", "Got it", etc.).

[3] doi: 10.17605/OSF.IO/3Y59M

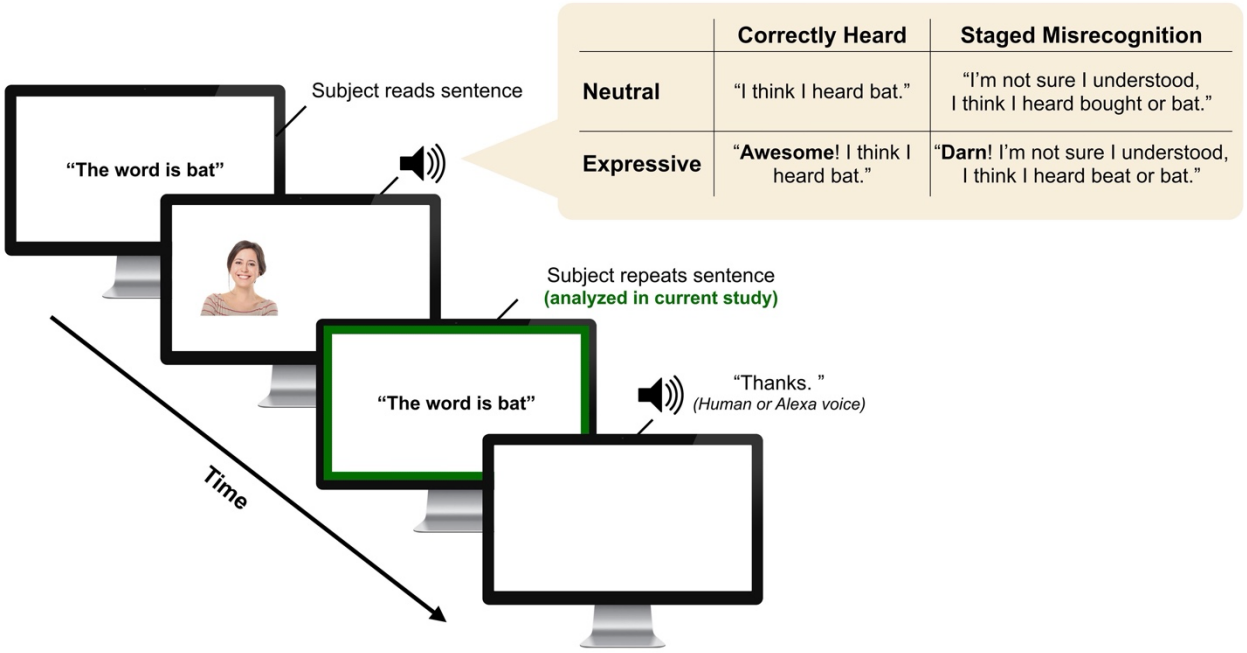

Figure 1. Interaction trial schematic. After participants read a sentence, the interlocutor (human or Alexa) responds in one of the Staged Misunderstanding Conditions (correctly heard, misrecognition) and Emotionality Conditions (neutral, emotionally expressive). Then, the subject responds (the production we analyze). Finally, the interlocutor provides a follow-up response.

In 50% of trials, the interlocutor (human, Alexa) 'misunderstood' the speakers, while in the other 50% they heard correctly. Additionally, in 50% of trials, the interlocutor responded with an expressive production (distributed equally across correctly heard and misrecognition trials). Order of target words was randomized, as well as trial correspondence to the Misunderstanding and Emotionality Conditions. In each block, participants produced all target sentences once for all conditions for a total of 128 trials for each interlocutor (16 words x 2 misunderstanding conditions x 2 emotionality conditions). Participants completed the task with both interlocutors (256 total target sentences). After the speech production experiment ended (and while still in the soundbooth), participants used a sliding scale (0-100) to rate how human-like each interlocutor sounded (order of interlocutor was randomized) ("How much like a real person did [Alexa/Human] sound?" (0=not like a real person, 100=extremely realistic)". The overall experiment took roughly 45 minutes.

2.5. Acoustic Analysis

Four acoustic measurements were taken over each target sentence in both the Citation and Interaction blocks using Praat scripts (De Jong et al., 2017; DiCanio, 2007): intensity (dB), speech rate (syllables/second), mean fundamental frequency (f0) (semitones, ST, relative to 100 Hz), and f0 variation (ST). We centered the measurements from the Interaction blocks within-speaker, subtracting their Citation speech mean value (within-speaker, within-word). This measurement indicates changes from the speakers' citation form for that feature.

To extract vowel-level features, recordings were force-aligned (using the Forced Alignment and Vowel Extraction (FAVE) suite) (Rosenfelder et al., 2014). Next, vowel boundaries were hand-corrected by trained research assistants: vowel onsets and offsets were

defined by the presence of both higher format structure and periodicity. Following hand-correction, we measured vowel duration and vowel formant frequency values (F1, F2) at vowel midpoint with FAVE-extract (Rosenfelder et al., 2014) for the subset of 13 words containing corner vowels: /i/ (cheek, weave, deed), /u/ (boot, hoop, toot, dune), /a/ (pod, cot, sock, tot), and /æ/ (bat, tap). We additionally scaled the formant frequency values (from Hertz) using a log base-10 transformation and centering each value to the subject's citation production values for that word (Nearey, 1978).

In order to assess whether speech changes made by participants were not simply alignment toward the interlocutors, the same sentence-level (rate, mean f0, f0 variation) and target vowel measurements (duration, F1, F2) were also taken over each interlocutor's production in Turn 2 (e.g., "I think I heard weave."). In order to compare across the interlocutors, formant frequency values (F1, F2) were centered relative to each interlocutor's mean value for that word (log mean normalization: Nearey, 1978).

2.6. Statistical Analysis

Participants' sentence-level values for each acoustic feature (centered to speaker citation form values) were modeled in separate linear mixed effects models with the *lme4* R package (Bates et al., 2015), with identical model structure: fixed effects of Interlocutor (voice-AI, human), Staged Misunderstanding Condition (correctly heard, misrecognition), Expressiveness (neutral, expressive), and all possible interactions, with by-Sentence and by-Speaker random intercepts.

Participants' vowel-level features (F1, F2) were also modeled in separate linear mixed effects models with a similar structure as in the sentence-level models: Interlocutor, Staged Misunderstanding Condition, Expressiveness Condition, with by-Word and by-Speaker random intercepts. In both the F1 and F2 model, we included an additional predictor of Vowel Category (For the F1 (height) model, this factor included 2 height levels: high vs. low vowels; for the F2 (backness) model, this factor included 2 levels: front vs. back vowels) and all possible interactions with the other predictors (Vowel Category*Interlocutor*Misunderstanding*Emotion). The formant models (F1, F2) additionally included a fixed effect of Vowel Duration (centered within speaker).

3. Results

3.1. Human-likeness rating

Figure 2 provides the mean values for participants' human-like ratings of the voices. A t-test on participants' ratings of the voices confirmed that the Alexa voice was perceived as less human-like ($\bar{x}$=31.06) than the human ($\bar{x}$=87.67) [t(104.87)=-12.84, p<0.001].

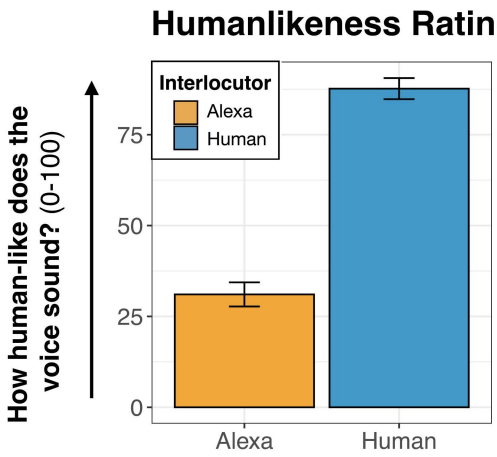

Figure 2. Mean 'human-like' ratings of each interlocutor. Error bars depict the standard error.

3.2. Interlocutor stimuli acoustics

T-tests of the interlocutors' productions found no overall difference between the Alexa and Human speaking rate (Human $\bar{x}$=2.53 syll/s; Alexa $\bar{x}$=2.68 syll/s) [t(124.27)=-1.87, p=0.06], but there was a significant difference in mean f0: the human had a higher mean f0 ($\bar{x}$ =14.42 ST) than Alexa ($\bar{x}$=13.16 ST) [t(106.25)=9.21, p<0.001]. Additionally, the human produced greater f0 variation ($\bar{x}$ =3.27 ST) than Alexa ($\bar{x}$=2.86 ST) [t(132.97)=7.06, p<0.001]. T-tests comparing formant frequency characteristics revealed no difference in vowel height (F1) for the interlocutors for high vowels (Human $\bar{x}$=-0.37 log Hz; Alexa $\bar{x}$=-0.41 log Hz) [t(35.28)=-1.38, p=0.18] or low vowels (Human $\bar{x}$=0.43 log Hz; Alexa $\bar{x}$=0.47 log Hz) [t(45.42)=1.75, p=0.09]. Additionally, there was no difference in vowel fronting (F2) for the interlocutors for front vowels (Human $\bar{x}$=0.30; Alexa $\bar{x}$=0.35)[t(34.66)=0.67, p=0.51] or back vowels (Human $\bar{x}$=-0.18; Alexa $\bar{x}$=-0.22)[t(47.73)=0.71, p=0.48].

T-tests comparing the Expressiveness Conditions (neutral vs. emotionally expressive) confirmed differences: expressive productions were produced with a slower speaking rate (Expressive $\bar{x}$=2.45 syll/s; Neutral $\bar{x}$= 2.76 syll/s) [t(153.88)=-4.25, p<0.001] and with a lower mean f0 (Expressive $\bar{x}$=13.55 ST; Neutral $\bar{x}$=14.03 ST) [t(145.44)=-2.89, p<0.01]. However, there was no difference for f0 variation (Expressive $\bar{x}$=3.04 ST; Neutral $\bar{x}$=3.09 ST) [t(157.44)=-0.60, p=0.55].

T-tests comparing the Misunderstanding Conditions (correctly heard vs. misrecognition) showed no significant difference in speaking rate (Correct $\bar{x}$=2.64 syll/s; misunderstood $\bar{x}$=2.57 syll/s) [t(139.38)=0.89, p=0.37] or mean f0 (Correct $\bar{x}$=13.92 ST; misunderstood $\bar{x}$=13.65 ST) [t(114.62)=1.60, p=0.11]. However, they did vary in terms of f0 variation: larger for correctly understood ($\bar{x}$=3.15 ST) than misrecognized ($\bar{x}$=2.98 ST) [t(141.59)=2.58, p<0.05].

3.3. Participants' sentence-level measurements

Figure 3 displays the mean acoustic values for participants' sentence-level measurements (centered to speakers' Citation form values). Model output tables are provided in Appendices A1-A4.

The Intensity model showed a significant intercept: participants increased their intensity in the interaction (relative to their citation form) [*Coef*=2.64, *SE*=0.45, t=5.86, p<0.001]. There was also a main effect of Misunderstanding Condition: as seen in Figure 3, participants' productions of

sentences that the system did not understand correctly were louder than repetitions of utterances that the system understood correctly [*Coef*=0.20, *SE*=0.04, *t*=5.13, *p*<0.001]. No other effects or interactions were significant in the Intensity model.

The Speech Rate model showed no difference from 0 for intercept: overall, speakers did not speed up or slow down their speech in interlocutor interactions, relative to their citation form productions. The model also revealed a main effect of Interlocutor, producing a slower speech rate (indicated by fewer syllables per second) in Alexa-DS [*Coef*=-0.03, *SE*=0.01, *t*=-2.87, *p*<0.01]. There was also a main effect of Misunderstanding Condition wherein speakers decreased their speech rate in response to a misrecognition [*Coef*=-0.02, *SE*=0.01, *t*=-1.96, *p*<0.05]. These effects can be seen in Figure 3. No other effects or interactions were significant in the model.

The Mean F0 model had a significant intercept, indicating that speakers increased their mean f0 in the interactions relative to the citation form productions [*Coef*=0.83, *SE*=0.15, *t*=5.65, *p*<0.001]. The model also showed an effect of Interlocutor: speakers produced a higher mean f0 toward the Alexa interlocutor [*Coef*=0.03, *SE*=0.01, *t*=2.40, *p*<0.05]. Additionally, there was an effect of Misunderstanding wherein responses to misunderstood utterances were produced with a higher f0 [*Coef*=0.06 *SE*=0.01, *t*=5.04, *p*<0.001], as seen in Figure 3. Furthermore, there was a main effect of Expressiveness Condition wherein speakers produced a higher mean f0 in response to emotionally expressive utterances [*Coef*=0.03, *SE*=0.01, *t*=2.49, *p*<0.05]. No other effects or interactions were observed in the Mean f0 model.

The F0 Variation model also had a significant intercept: relative to their citation form productions, speakers increased their f0 variation in the interaction [*Coef*=0.34, *SE*=0.07, *t*=4.94, *p*<0.001]. There was also a main effect of Interlocutor: speakers produced greater f0 variation in responses directed to the Alexa voice [*Coef*=0.02, *SE*=0.01, *t*=2.79, *p*<0.01]. Additionally, there was an effect of Misunderstanding: responses to misrecognitions were produced with greater f0 variation [*Coef*=0.01, *SE*=0.01, *t*=1.98, *p*<0.05]. No other effects or interactions were significant in the F0 Variation model.

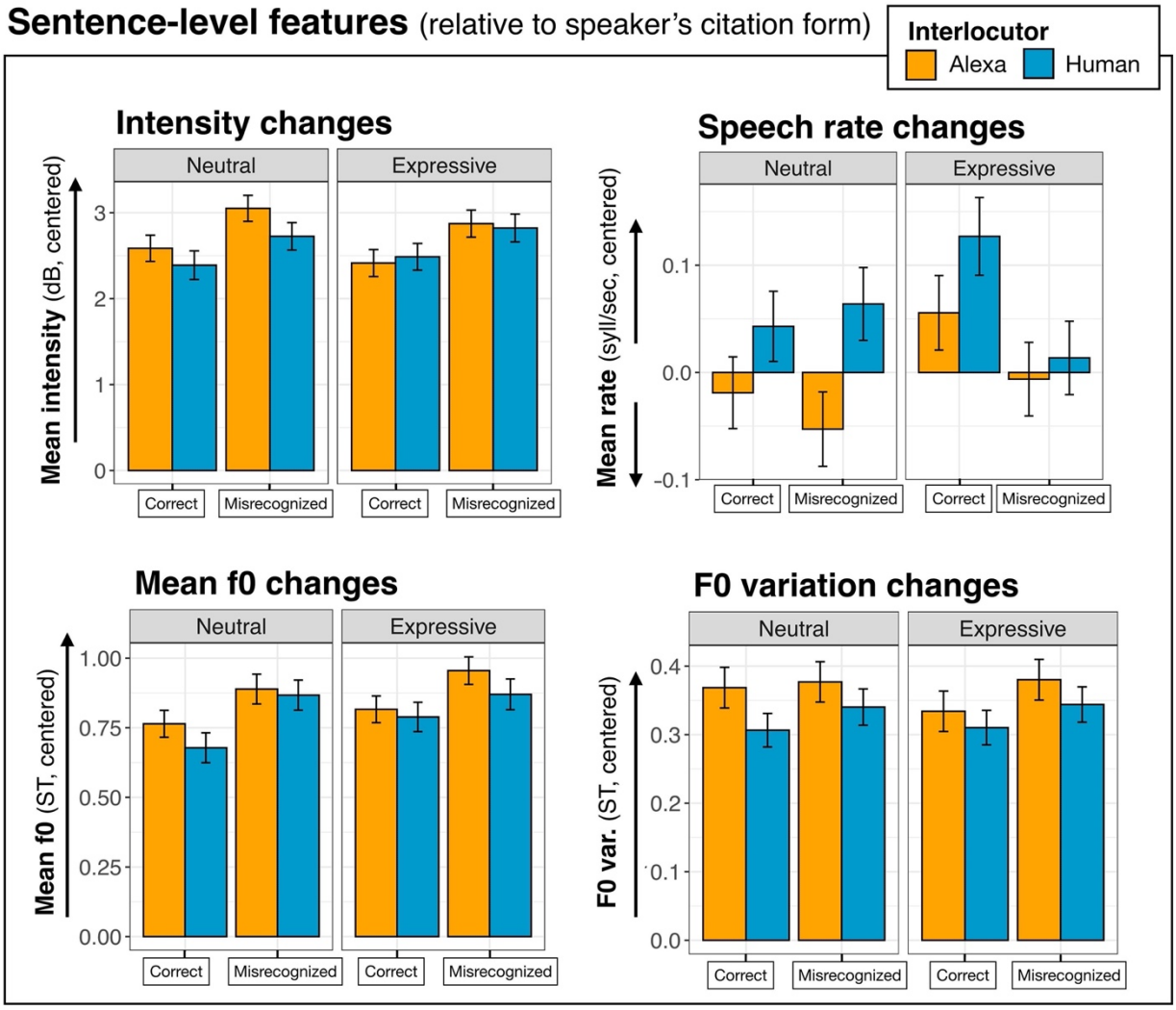

Figure 3. Mean acoustic changes from speaker's citation form productions to the interaction with the Interlocutors (Alexa vs. human) for sentence intensity (in decibels, dB), speech rate (syllables per second), f0 (semitones, ST, rel. to 100 Hz), and f0 variation (ST). The x-axis shows Staged Misunderstanding Condition (correctly heard vs. misrecognized), while Expressiveness Condition is faceted. Values higher than 0.0 indicate an increase (relative to speakers' citation form), while values lower than 0.0 indicate a relative decrease. Error bars depict the standard error.

3.4. Participants' vowel-level measurements

Figure 4 displays participants' mean vowel-level values across conditions. Model output tables are provided in Appendices A5 and A6.

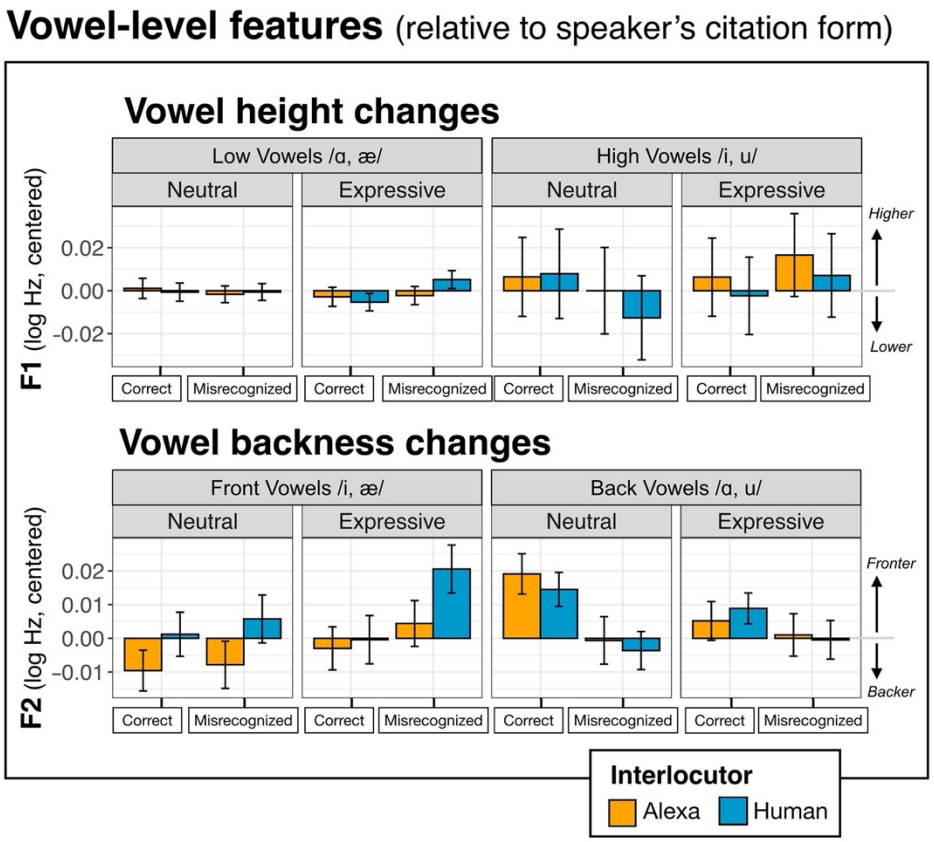

Figure 4. Mean acoustic changes from speaker's citation form productions to the interaction with the Interlocutors (Alexa vs. human) for vowel duration (milliseconds, ms), F1 (log Hertz, Hz), and F2 (log Hertz, Hz). Formant plots are additionally faceted by Vowel Category: F1 (by vowel height: low vs. high vowels) and F2 (by vowel backness: front vs. back vowels). The x-axis shows Staged Misunderstanding Condition (correctly heard vs. misrecognized), while Expressiveness Condition is faceted. Values higher than 0.0 indicate an increase (relative to speakers' citation form), while values lower than 0.0 indicate a relative decrease. Error bars depict the standard error.

The F1 model testing changes in vowel height (where a smaller F1 values indicate raising) showed no significant intercept; relative to the citation forms, speakers did not change their vowel height. The model revealed only an effect of Vowel Duration: speakers produce lower vowels (higher F1) with increasing duration [*Coef*=2.1e-04, *SE*=7.8e-05, *t*=2.62, *p*<0.01]. No other effects or interactions were significant.

The F2 model, testing changes in vowel backness, showed several significant effects. While there was no significant intercept (indicating no general change in vowel backness from citation form), participants produced more backed vowels (i.e., lower F2 values) with increasing vowel duration [*Coef*=-1.8e-04, *SE*=3.4e-05, *t*=-5.41, *p*<0.001]. There was also an interaction between Misunderstanding Condition and Vowel Category. As seen in Figure 4, back vowels were produced even farther back (lower F2) in response to a staged word misrecognition [*Coef*=-0.01, *SE*=1.5e-03, *t*=-3.46, *p*<0.001]. No other effects or interactions were observed[4].

4. Discussion

[4] Note that while there is a numerical F2 increase in the Front Vowels in response to Misrecognized Expressive productions, this was not significant in the main model or in a post hoc model (with the subset of Front Vowels).

The current study examined whether participants use a different speech style when talking to an Alexa interlocutor, relative to a human interlocutor, in a computer-mediated interaction (a summary of the main effects is provided in Table 2). We systematically controlled functional and socio-communicative pressures in real-time during interactions with both interlocutors who made the same types and rates of staged word misrecognitions, and responded in emotionally expressive and neutral manners. This approach serves to complement studies done with users talking to devices in their home (e.g., Huang et al., 2019; Mallidi et al., 2018) and also pinpoint differences that might be present due to other factors in the situation (e.g., physical distance from the microphone; rate and type of automatic speech recognition (ASR) errors). While TTS methods have advanced in recent years (e.g., Wavenet in Van Den Oord et al., 2016), our participants rated the two talkers as distinct in their human-likeness: Alexa was less human-like than the human voice, consistent with prior work (Cohn, Sarian, et al., 2020; Cohn & Zellou, 2020).

Overall, we found prosodic differences across Alexa- and human-DS, consistent with *routinized interaction accounts* that propose people have a 'routinized' way of engaging with technology (Gambino et al., 2020), and in line with prior studies finding differences in computer and voice-AI speech registers (e.g., Burnham et al., 2010; Huang et al., 2019; Siegert & Krüger, 2020). In the present study, speakers showed a systematic Alexa-DS speech style: when talking to Alexa, speakers produced sentences with a slower rate, higher mean f0, and higher f0 variation, relative to human-DS. These differences align with prior work showing slowed speech rate toward Alexa socialbot (Cohn et al., 2021), increased higher mean f0 in speech toward voice-AI (Raveh et al., 2019), and greater segmental lengthening in computer-DS (Burnham et al., 2010). Furthermore, both an increased mean f0 and f0 variation are consistent with increased vocal effort in response to a presumed communicative barrier; for instance, prior work has reported that speakers produce greater f0 variation in response to a word misrecognition in computer-DS (Vertanen, 2006), as well as higher mean f0 and a larger f0 range in Lombard speech (Brumm & Zollinger, 2011; Marcoux & Ernestus, 2019). Furthermore, in contrast to other work reporting greater intensity in Alexa-DS (Raveh et al., 2019; Siegert & Krüger, 2020), we did not see a difference in intensity in the present study. This might reflect the controlled interaction, where participants were recorded with a head-mounted microphone (such that it was equidistant from their mouths for the entire experiment) and heard amplitude normalized stimuli over headphones. Additionally, the lack of an intensity effect suggests that adjustments in Alexa-DS differ from strict 'Lombard' effects (e.g., louder in Brumm & Zollinger, 2011).

While one possibility was that these adjustments reflect alignment toward the Alexa talker, we did not find support for this: acoustic analyses demonstrated that the Alexa productions had lower mean f0 and less f0 variation than the human productions (speech rate did not significantly differ for the Alexa and human productions). Hence, speakers appear to produce more effortful prosodic adjustments in response to an interlocutor with presumed communicative barriers (Branigan et al., 2011; Clark & Murphy, 1982; Cowan et al., 2015; Oviatt, Levow, et al., 1998), even while the 'actual' misunderstandings were matched across the two talker types.

Table 2. Summary of effects in main analysis, comparing interlocutor acoustics.

		Speaking Style Changes	*Interlocutor acoustics*
Sentence-level	Intensity	Louder for Misrecognition	--

	Speech rate	Decreased rate in Alexa-DS Decreased rate for Misrecognition	*Alexa vs. human N.S.* *Correct vs. misrecognized N.S.*
	Mean f0	Higher mean f0 in Alexa-DS Higher mean f0 for Misrecognition Higher mean f0 for Expressive	*Human - higher mean f0 ($p<0.001$)* *Correct vs. misrecognized N.S.* *Expressive - lower mean f0 ($p<0.01$)*
	F0 variation	More f0 variation in Alexa-DS More f0 variation for Misrecognition	*Human- larger f0 var. ($p<0.001$)* *Correct - greater f0 var. ($p<0.05$)*
Vowel-level	F1 (Vowel height)	No diff.	*Alexa vs. human N.S.*
	F2 (Vowel backness)	Back vowels backed for Misrecognition	*Alexa vs. human N.S.*

Do the differences in human- and Alexa-DS reflect *distinct* functionally-oriented speech registers? Examining responses to misrecognized utterances suggests that some of these adjustments might be part of a more general speech intelligibility strategy. When either interlocutor 'misheard' the word, participants responded by producing many of the same adjustments they did in Alexa-DS, including slower rate, higher f0, and higher f0 variation. These adjustments are in line with proposals that the speech adjustments people make in communicatively challenging contexts are listener-oriented (Lindblom, 1990; Smiljanić & Bradlow, 2009; Hazan & Baker, 2011). Thus, for these particular features, the adjustments made when there is a local communicative pressure parallel those made globally in Alexa-DS, suggesting that speakers make adjustments following misrecognitions and toward Alexa to improve intelligibility.

Yet, we see other adjustments in response to word misrecognitions not seen globally in Alexa-DS: increased intensity and F2 adjustments. These F2 adjustments, in particular, are predicted based on the type of misunderstanding created in the experimental design: when the interlocutor 'misheard' the participant, they always produced the correct target word alongside its minimal pair counterpart which differed in backness (e.g., "mask" (front vowel) versus "mosque" (back vowel)). Producing back vowels further back is consistent with vowel space expansion. In particular, one possibility is that these F2 adjustments are targeted specifically for clarity, making the vowels more distinct from the distractor minimal pair. This aligns with findings from Stent et al. (2008) who found that speakers repaired misrecognitions of high vowels by a dialog system (e.g., "deed") by producing even *higher* vowels. That the same effect is not seen for front vowels in the current study could come from the dialectal variety of the speakers: participants were California English speakers, a variety with back vowel fronting (Hall-Lew, 2011). Thus, it is possible that there is more room for these speakers to make back vowels more back, rather than to adjust the front vowels, though further work exploring dialect-specific intelligibility strategies can shed light on this question (cf. Clopper et al., 2017; Zellou & Scarborough, 2019). Future work varying vowel height, as well as hyperarticulation of consonants (e.g., flapping vs. /t/ release in Stent et al., 2008) can further explore targeting effects in response to word misunderstandings.

However, if people produce global register differences in speech toward Alexa that parallel those seen in response to misrecognitions, why don't we see *greater* speech adjustments in response to misrecognitions made by Alexa? One possible explanation for the similarities is the rate: in the current study, the interlocutors both had staged word misrecognitions in 50% of trials.

Related work has shown that rate of misrecognition can change speakers' global and local adaptations (Oviatt, MacEachern, et al., 1998; Stent et al., 2008); at a high rate of word misrecognitions, speakers might produce more similar intelligibility-related adjustments across interlocutors. Additionally, this high misrecognition rate — as well as random occurrence of the misunderstandings — might be interpreted by the speaker that the listener (human or Alexa) is not benefiting from these adjustments, which might drive similarities. In the current study, speakers might produce a word as clearly as they can and the human/voice-AI listener still misunderstands them half the time. The extent to which these patterns hold at a lower misrecognition rate — or an adaptive misrecognition rate, improving as the speaker produces 'clearer' speech — are avenues for future work.

Furthermore, another possible reason for the similar intelligibility adjustments in response to a misunderstanding (in both Alexa- and human-DS) is that the speakers did not have access to information about the source of these perceptual barriers. For example, Hazan and Baker (2011) found that speakers dynamically adjust their speech to improve intelligibility when they are told their listener is hearing them in competing background speakers or as noise-vocoded speech (simulating the auditory effect of cochlear implants), relative to when the listener experienced no barrier. Furthermore, the *type* of adjustments varied according to the type of barrier (e.g., more f0 adjustments when the listener was in 'babble' than 'vocoded speech'). In the present study, speakers were left to 'guess' what the source of the communicative barrier was, based on observed behavior of the human or voice-AI interlocutor. Indeed, when the speaker does not have information about the listener, adaptations might not be advantageous. For example, computer-DS adaptations have been shown in some work to lead to worse outcomes for some ASR systems, leading to a cycle of misunderstanding (e.g., Wade et al., 1992; for a discussion, see Stent et al., 2008 and Oviatt et al., 1998). Future work examining intelligibility for the intended listener (here, a human or ASR system) can further shed light on the extent local intelligibility adjustments in Alexa- and human-DS are equally beneficial.

Another possible factor why we see similar local intelligibility adjustments in response to misunderstandings (across Alexa- and human-DS) is that the experiment was computer-mediated. Recent work has shown differences in linguistic behavior across contexts: for example, participants show stronger style convergence toward their interlocutor in the in-person condition, relative to a (text-based) computer-mediated interaction (Liao et al., 2018). In line with this possibility, Burnham et al. (2010) found similar adjustments in response to a misrecognition made by a computer- and human-DS (but overall differences in computer-DS, paralleling our findings). At the same time, in the current study, the human-likeness ratings for the interlocutors collected at the end of study suggest that the participants found the interlocutors to be distinct. Future work manipulating rate of misunderstanding and embodiment (Cohn, Jonell, et al., 2020; Staum Casasanto et al., 2010) can investigate what conditions lead to greater targeted intelligibility strategies for distinct interlocutor types.

We also explored whether emotional expressiveness mediates speech styles for Alexa- and human-DS. Here, we found the same speech adjustments in response to expressiveness by both interlocutors: higher mean f0 in response to utterances containing emotional expressiveness. First, speakers' overall higher f0 in their sentences does not appear to reflect an alignment toward the interlocutors (who actually produced lower mean f0 in their expressive productions). One possible explanation for the increased f0 following the expressive responses is that it reflects a positivity bias in reaction to stimuli (but see Jing-Schmidt (2007) for work on biases toward negative valence). Indeed, work has shown that smiling is associated with higher mean f0 (Tartter, 1980;

Tartter & Braun, 1994) (but we did not see formant shifts, which are also associated with smiled speech, in response to Expressiveness). Here, one explanation for similarities in response to emotion by both interlocutors is that speakers are applying the social behaviors toward voice-AI as they do toward humans, as proposed by *technology equivalence accounts* (Lee, 2008; Nass et al., 1997, 1994). For instance, here people are reacting to emotional expressiveness by both types of interlocutors similarly. This explanation is consistent with work showing similar affective responses to computers as seen in human-human interaction (e.g., Brave et al., 2005; Cohn, Chen, et al., 2019; Cohn & Zellou, 2019).

Additionally, we did not observe differences in how participants adapted their speech following an emotionally expressive or neutral word misrecognition. This contrasts with related work on this same corpus (Zellou & Cohn, 2020) that found greater vowel duration alignment when participants responded to an emotionally expressive word misunderstanding made by a voice-AI system. Thus, it is possible that emotional expressiveness might shape vocal alignment, but it might not influence speech style adjustments. That emotion appears to have an effect on vocal alignment toward humans and voice-AI (e.g., Cohn & Zellou, 2019; Vaughan et al., 2018) could be explained by proposals that alignment is used as a means to communicate social closeness (Giles et al., 1991). While conveying affect is thought to be part of infant- and pet-DS registers (Trainor et al., 2000), listener-oriented speech styles directed toward human adults (non-native speakers, hearing impaired speakers) and computers are generally not associated with increased emotionality. Furthermore, conveying affect is generally not associated with clear speech strategies. Indeed, classic perspectives on clear speech (H&H theory) do not account for emotionality in predicting hyperspeech behavior (e.g., Lindblom, 1990). Yet, one possibility for a lack of difference in the current study is based on how emotion was added in the stimuli: emotional expressiveness was conveyed only in the interjection. Since the time this study was run, there are now more ways to adapt the Alexa voice in terms of positive and negative emotionality (at low, medium, and high levels[5]), which can serve as avenues for future research.

There were also several limitations of the present study which open directions for future work. For instance, one possible factor in the lack of difference detected for emotionality across Alexa- and human-DS is the communicative context: the current study consisted of fully scripted interactions in a lab setting. While this controlled interaction was intentional as we were interested in word misrecognitions (which might otherwise be difficult to control in voice-AI interactions), it is possible that differences based on emotional expressiveness might be seen in a non-scripted conversation with voice-AI, as well as one conducted outside a lab context (e.g., Cohn et al., 2019). Additionally, the present study used two types of voices; it is possible that other paralinguistic features of those voices might have mediated speech style adjustments. For example, recent work has shown that speakers align speech differently toward TTS voices that 'sound' older (e.g., Apple's Siri voices, rated in their 40s and 50s) (Zellou et al., 2021). Furthermore, there is work showing that introducing 'charismatic' features from human speakers' voices shapes perception of TTS voices (Fischer et al., 2019; Niebuhr & Michalsky, 2019). The extent to which individual differences in speakers (human and TTS) and participants remain avenues for future research.

While here the findings align with those for another Germanic language (e.g., German in Raveh et al., 2019; Siegert & Krüger, 2020), the extent to which the same effects might be observed with other languages and other cultures is another open question for future work. For example, cultures might vary in terms of acceptance of voice-AI technology, such as due to privacy concerns

[5] https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html#amazon-emotion

(e.g., GDPR in Europe: Loideain & Adams, 2020; Voss, 2016). Additionally, cultures vary in terms of their expressions of emotion (Mesquita & Markus, 2004; Shaver et al., 1992; Van Hemert et al., 2007). How emotional expressiveness and 'trust' in voice-AI (Metcalf et al., 2019; Shulevitz, 2018) might interact remains an open question for future work.

5. Conclusion

Overall, this work adds to our growing understanding of the dynamics of human interaction with voice-AI assistants — still distinct from how individuals talk to human interlocutors. As these systems and other AI robotics systems are even more widely adopted, characterizing these patterns across different timepoints — and with diverse populations of participants — is important in our ability to track the trajectory of the influence of voice-AI on humans and human speech across languages and cultures.

Acknowledgments

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC and an Amazon Faculty Research Award to GZ.

References

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.

Arnold, R., Tas, S., Hildebrandt, C., & Schneider, A. (2019). Any Sirious Concerns Yet?–An Empirical Analysis of Voice Assistants' Impact on Consumer Behavior and Assessment of Emerging Policy Challenges. *An Empirical Analysis of Voice Assistants' Impact on Consumer Behavior and Assessment of Emerging Policy Challenges (July 25, 2019)*.

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

Bell, L., & Gustafson, J. (1999). Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. *Proceedings of ICPhS*, *99*, 1221–1224.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. *Proceedings of 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–24.

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
Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11–13), 1173–1198.
Burbach, L., Halbach, P., Plettenberg, N., Nakayama, J., Ziefle, M., & Valdez, A. C. (2019). " Hey, Siri"," Ok, Google"," Alexa". Acceptance-Relevant Factors of Virtual Voice-Assistants. *2019 IEEE International Professional Communication Conference (ProComm)*, 101–111.
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, 14-16 December 2010, Melbourne, Australia*, 13–17.
Clark, H. H. (1996). *Using language*. Cambridge university press.
Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287–299). Elsevier.
Clopper, C. G., Mitsch, J. F., & Tamati, T. N. (2017). Effects of phonetic reduction and regional dialect on vowel production. *Journal of Phonetics*, *60*, 38–59.
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.
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.
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.
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*. https://doi.org/10.3389/fcomm.2021.671429
Cohn, M., Sarian, M., Predeck, K., & Zellou, G. (2020). Individual variation in language attitudes toward voice-AI: The role of listeners' autistic-like traits. *Proc. Interspeech 2020*, 1813–1817.
Cohn, M., & Zellou, G. (2019). Expressiveness influences human vocal alignment toward voice-AI. *Proc. Interspeech 2019*, 41–45.
Cohn, M., & Zellou, G. (2020). Perception of concatenative vs. Neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. *Proceedings of Interspeech*, 1733–1737. http://dx.doi.org/10.21437/Interspeech.2020-1336
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 Jong, N. H., Wempe, T., Quené, H., & Persoon, I. (2017). *Praat script speech rate v2*. https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2
DiCanio, C. (2007). *Extract Pitch Averages*. https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_pitch.praat
Fischer, K., Niebuhr, O., Jensen, L. C., & Bodenhagen, L. (2019). Speech melody matters—How robots profit from using charismatic speech. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(1), 1–21.
Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: extending the

computers are social actors paradigm. *Human-Machine Communication*, *1*(1), 5.
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. In *Forms of talk* (pp. 78–122). University of Pennsylvania Press.
Habler, F., Schwind, V., & Henze, N. (2019). Effects of Smart Virtual Assistants' Gender and Language. In *Proceedings of Mensch und Computer 2019* (pp. 469–473).
Hall-Lew, L. (2011). The completion of a sound change in California English. *ICPhS*, 807–810.
Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139–2152. https://doi.org/10.1121/1.3623753
Hazan, V. L., Uther, M., & Granlund, S. (2015). How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles? *Proceedings of the 18th International Congress of Phonetic Sciences*.
Hoffmann, L., Krämer, N. C., Lam-Chi, A., & Kopp, S. (2009). Media equation revisited: Do users show polite reactions towards an embodied agent? *International Workshop on Intelligent Virtual Agents*, 159–165.
Huang, C.-W., Maas, R., Mallidi, S. H., & Hoffmeister, B. (2019). A Study for Improving Device-Directed Speech Detection toward Frictionless Human-Machine Interaction. *Proc. Interspeech 2019*, 3342–3346.
Jing-Schmidt, Z. (2007). *Negativity bias in language: A cognitive-affective model of emotive intensifiers*. *18*(3), 417–443. https://doi.org/10.1515/COG.2007.023
Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524.
Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, *20*(1–2), 13–22.
Knoll, M. A., Johnstone, M., & Blakely, C. (2015). Can you hear me? Acoustic modifications in speech directed to foreigners and hearing-impaired people. *Sixteenth Annual Conference of the International Speech Communication Association*.
Lee, K. M. (2008). Media Equation Theory. In *The International Encyclopedia of Communication*. https://doi.org/10.1002/9781405186407.wbiecm035
Liao, W., Bazarova, N. N., & Yuan, Y. C. (2018). Expertise judgment and communication accommodation in linguistic styles in computer-mediated and face-to-face groups. *Communication Research*, *45*(8), 1122–1145.
Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.
Loideain, N. N., & Adams, R. (2020). From Alexa to Siri and the GDPR: the gendering of virtual personal assistants and the role of data protection impact assessments. *Computer Law & Security Review*, *36*, 105366.
Lopatovska, I. (2020). Personality dimensions of intelligent personal assistants. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 333–337.
Lunsford, R., Oviatt, S., & Arthur, A. M. (2006). Toward open-microphone engagement for multiparty interactions. *Proceedings of the 8th International Conference on Multimodal Interfaces*, 273–280.
Mallidi, S. H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., & Hoffmeister, B. (2018). Device-directed utterance detection. *Interspeech 2018*. ISCA, Hyderabad, India.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962–3973.
Marcoux, K. P., & Ernestus, M. T. C. (2019). *Differences between native and non-native Lombard speech in terms of pitch range*.
Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Thirteenth Annual Conference of the International Speech Communication Association*, 1706–1709.
Mesquita, B., & Markus, H. R. (2004). Culture and emotion. *Feelings and Emotions: The Amsterdam Symposium*, 341.
Metcalf, K., Theobald, B.-J., Weinberg, G., Lee, R., Jonsson, M., Webb, R., & Apostoloff, N. (2019). Mirroring to Build Trust in Digital Assistants. *Proc. Interspeech 2019*, 4000–4004.
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
Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels (Indiana University Linguistics Club, Bloomington, IN)*.
Niebuhr, O., & Michalsky, J. (2019). Computer-generated speaker charisma and its effects on human actions in a car-navigation system experiment-or how Steve Jobs' tone of voice can take you anywhere. *International Conference on Computational Science and Its Applications*, 375–390.
Oviatt, S., Levow, G.-A., Moreton, E., & MacEachern, M. (1998). Modeling global and focal hyperarticulation during human–computer error resolution. *The Journal of the Acoustical Society of America*, *104*(5), 3080–3098.
Oviatt, S., MacEachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, *24*(2), 87–110.
Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *28*(1), 96–103.
Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569.
Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–2859. https://doi.org/10.1145/3027063.3053246
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., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) Program Suite v1. 2.2. *DOI: Https://Doi. Org/10.5281/Zenodo*, *22281*.
Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study

of real and imagined foreigner-directed speech. *Proceedings of the International Congress of Phonetic Sciences*, 2165–2168.
Scarborough, R., & Zellou, G. (2013). Clarity in communication:"Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, *134*(5), 3793–3807.
Shaver, P. R., Wu, S., & Schwartz, J. C. (1992). *Cross-cultural similarities and differences in emotion and its representation.*
Shulevitz, J. (2018). Alexa, should we trust you. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/
Siegert, I., & Krüger, J. (2020). "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* (pp. 65–95). Springer.
Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., & Lotz, A. (2018, May). Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA. *N Proc. of the 11th LREC*.
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.
Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, *3*(1), 236–264.
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.
Stent, A. J., Huffman, M. K., & Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, *50*(3), 163–178. https://doi.org/10.1016/j.specom.2007.07.005
Swerts, M., Litman, D., & Hirschberg, J. (2000). Corrections in spoken dialogue systems. *Sixth International Conference on Spoken Language Processing*.
Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, *27*(1), 24–27.
Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, *96*(4), 2101–2107. https://doi.org/10.1121/1.410151
Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*(3), 188–195.
Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication*, *49*(1), 2–7.
Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *SSW*, 125.
Van Hemert, D. A., Poortinga, Y. H., & van de Vijver, F. J. (2007). Emotion and culture: A meta-analysis. *Cognition and Emotion*, *21*(5), 913–943.
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.
Vertanen, K. (2006). Speech and speech recognition during dictation corrections. *Ninth International Conference on Spoken Language Processing*, 1890–1893.
Voss, W. G. (2016). European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. *The Business Lawyer*, *72*(1), 221–234.
Wade, E., Shriberg, E., & Price, P. (1992). User behaviors affecting speech recognition. *Second International Conference on Spoken Language Processing*.
Zellou, G., & Cohn, M. (2020). Social and functional pressures in vocal alignment: Differences for human and voice-AI interlocutors. *Proc. Interspeech 2020*, 1634–1638. http://dx.doi.org/10.21437/Interspeech.2020-1335
Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age- and Gender-Related Differences in Speech Alignment Toward Humans and Voice-AI. *Frontiers in Communication*, *5*, 1–11. https://doi.org/10.3389/fcomm.2020.600361
Zellou, G., & Scarborough, R. (2019). Neighborhood-conditioned phonetic enhancement of an allophonic vowel split. *The Journal of the Acoustical Society of America*, *145*(6), 3675–3685.

Appendix A. Model outputs for sentence measurements

Table A1. Intensity

	Coef	*SE*	*df*	*t*	*p*	
Intercept	2.63	0.45	58.12	5.86	<0.001	***
Interlocutor(Alexa)	0.07	0.04	6415.3	1.88	0.06	
MisunderstandingCond(misrecognized)	0.20	0.04	6415.31	5.13	<0.001	***
ExpressivenessCond(expressive)	-1.4e-03	0.04	6415.2	-0.03	0.97	
Int(Alexa)*Misunderstanding(misrec.)	0.04	0.04	6415.19	1.02	0.31	
Int(Alexa)*Expr(expressive)	-0.06	0.04	6415.17	-1.64	0.10	
Misunderstanding(misrec.)*Expr(express.)	-0.01	0.04	6415.2	-0.14	0.89	
Int(Alexa)* Misunderstanding(misrec.)* Expr(express.)	-0.01	0.04	6415.22	-0.15	0.88	

Num. observations =6,490, Num. subjects=53, *Sentences*=16

Table A2. Speech rate

	Coef	*SE*	*df*	*t*	*p*	
Intercept	0.03	0.05	59.71	0.53	0.60	
Interlocutor(Alexa)	-0.03	0.01	6416.32	-2.87	<0.01	**
MisunderstandingCond(misrecognized)	-0.02	0.01	6416.34	-1.96	<0.05	*
ExpressivenessCond(expressive)	0.02	0.01	6415.74	1.81	0.07	
Int(Alexa)*Misunderstanding(misrec.)	1.6e-03	0.01	6415.79	0.15	0.88	
Int(Alexa)*Expr(expressive)	0.01	0.01	6415.59	0.90	0.37	
Misunderstanding(misrec.)*Expr(express.)	-0.02	0.01	6415.67	-1.81	0.07	
Int(Alexa)* Misunderstanding(misrec.)* Expr(express.)	0.01	0.01	6415.73	1.2	0.23	

Num. observations =6,490, Num. subjects=53, *Sentences*=16

Table A3. Mean f0

	Coef	*SE*	*df*	*t*	*p*	
Intercept	0.83	0.15	53.69	5.65	<0.001	***
Interlocutor(Alexa)	0.03	0.01	6382.35	2.40	0.02	*
MisunderstandingCond(misrecognized)	0.06	0.01	6382.41	5.04	<0.001	***
ExpressivenessCond(expressive)	0.03	0.01	6382.23	2.49	0.01	*
Int(Alexa)*Misunderstanding(misrec.)	-6.9e-04	0.01	6382.15	-0.05	0.96	
Int(Alexa)*Expr(expressive)	2.6e-04	0.01	6382.11	0.02	0.98	
Misunderstanding(misrec.)*Expr(express.)	-0.01	0.01	6382.26	-0.92	0.36	
Int(Alexa)* Misunderstanding(misrec.)* Expr(express.)	0.01	0.01	6382.25	1.12	0.26	

Num. observations =6,457 Num. subjects=53, *Sentences*=16

Table A4. F0 variation

	Coef	*SE*	*df*	*t*	*p*	
Intercept	0.34	0.07	54.45	4.94	<0.001	***
Interlocutor(Alexa)	0.02	0.01	6382.74	2.79	<0.01	**
Misunderstanding(misrec.)	0.01	0.01	6382.79	1.98	<0.05	*
ExpressivenessCond(expressive)	-1.1e-03	0.01	6382.55	-0.14	0.89	
Int(Alexa)* Misunderstanding(misrec.)	-4.1e-04	0.01	6382.47	-0.05	0.96	
Int(Alexa)*Expr(expressive)	-4.9e-03	0.01	6382.41	-0.65	0.52	
Misunderstanding(misrec.)*Expr(express.)	4.8e-03	0.01	6382.56	0.63	0.53	
Int(Alexa)* Misunderstanding(misrec.)* Expr(express.)	0.01	0.01	6382.56	0.74	0.46	

Num. observations =6,457, Num. subjects=53, *Sentences*=16

Appendix B. Model outputs for vowel measurements
Table B1. F1 (Vowel height)

	Coef	*SE*	*df*	*t*	*p*	
Intercept	1.2e-03	0.01	37.12	-0.10	0.92	
Interlocutor(Alexa)	1.1e-03	3.6e-03	5576.60	0.31	0.76	
MisunderstandingCond(misrecognized)	-3.0e-04	3.6e-03	5576.39	-0.08	0.93	
ExpressivenessCond(expressive)	1.7e-03	3.6e-03	5575.11	0.46	0.65	
VowelCategory: HighorLow(high)	-2.0e-03	0.01	10.89	-0.27	0.79	
Duration	2.1e-04	7.8e-05	4823.13	2.62	<0.01	**
Int(Alexa)* Misunderstanding(misrec.)	4.8e-04	3.6e-03	5575.16	0.13	0.89	
Int(Alexa)*Expr(expressive)	2.7e-05	3.6e-03	5575.00	0.01	0.99	
Misunderstanding(misrec.)*Expr(expressive)	3.8e-03	3.6e-03	5574.24	1.05	0.30	
Int(Alexa)*Vowel(high)	-2.3e-03	3.6e-03	5573.96	-0.64	0.52	
Misunderstanding(misrec.)*Vowel(high)	1.0e-03	3.6e-03	5573.61	0.28	0.78	
Expr(express.)*Vowel(high)	-1.7e-03	3.6e-03	5573.93	-0.47	0.64	
Int(Alexa)* Misunderstanding(misrec.)Expr(express.)	-1.4e-03	3.6e-03	5574.67	-0.39	0.70	
Int(Alexa)* Misunderstanding(misrec.)Vowel(high)	-1.6e-03	3.6e-03	5573.59	-0.44	0.66	
Int(Alexa)*Expr(express.)*Vowel(high)	-7.2e-04	3.6e-03	5573.60	-0.20	0.84	
Misunderstanding(misrec.)*Expr(express.)* Vowel(high)	-2.0e-03	3.6e-03	5573.29	-0.55	0.59	
Int(Alexa)* Misunderstanding(misrec.)* Expr(express.)*Vowel(high)	5.9e-04	3.6e-03	5573.48	0.16	0.87	

Num. observations =5,656, Num. subjects=53, *Words*=13

Table B2. F2 (Vowel backness)

	Coef	*SE*	*df*	*t*	*p*	
Intercept	0.01	0.01	30.83	1.06	0.30	
Interlocutor(Alexa)	-2.2e-03	1.5e-03	5576.13	-1.43	0.15	
MisunderstandingCond(misrecognized)	-7.1e-04	1.5e-03	5575.87	-0.46	0.65	
ExpressivenessCond(expressive)	9.9e-04	1.5e-03	5575	0.64	0.52	
VowelCategory:FrontorBack(back)	2.0e-03	4.0e-03	10.9	0.51	0.62	
Duration	-1.8e-04	3.4e-05	5024.68	-5.41	<0.001	***
Int(Alexa)* Misunderstanding(misrec.)	-8.5e-04	1.5e-03	5574.95	-0.55	0.59	
Int(Alexa)*Expr(expressive)	-2.7e-04	1.5e-03	5575	-0.17	0.86	
Misunderstanding(misrec.)*Expr(expressive)	2.9e-03	1.5e-03	5574.26	1.9	0.06	
Int(Alexa)*Vowel(back)	3.0e-03	1.5e-03	5573.58	1.94	0.05	
Misunderstanding(misrec.)*Vowel(back)	-0.01	1.5e-03	5573.59	-3.46	<0.001	***
Expr(express.)*Vowel(back)	-3.0e-03	1.5e-03	5573.54	-1.92	0.06	
Int(Alexa)*Misunderstanding(misrec.)* Expr(express.)	-1.5e-03	1.5e-03	5574.67	-0.10	0.92	
Int(Alexa)*Misunderstanding(misrec.)* Vowel(back)	1.4e-03	1.5e-03	5573.53	0.90	0.37	
Int(Alexa)*Expr(express.)*Vowel(back)	-1.1e-03	1.5e-03	5573.6	-0.69	0.49	
Misunderstanding(misrec.)*Expr(express.)* Vowel(back)	2.4e-04	1.5e-03	5573.48	0.16	0.88	
Int(Alexa)*Misunderstanding(misrec.)* Expr(express.)*Vowel(back)	1.1e-03	1.5e-03	5573.49	0.68	0.50	

Num. observations =5,656, Num. subjects=53, *Words*=13