Journal of Phonetics 90 (2022) 101123

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Acoustic-phonetic properties of Siri- and human-directed speech

Michelle Cohn [a,b,c,*], Bruno Ferenc Segedin [a], Georgia Zellou [a]

[a] Department of Linguistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
[b] Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
[c] Department of Psychology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 29 July 2020
Received in revised form 3 September 2021
Accepted 7 September 2021

Keywords:
Register adaptation
Voice-AI
Error correction
Human–computer interaction

ABSTRACT

Millions of people engage in spoken interactions with voice activated artificially intelligent (voice-AI) systems in their everyday lives. This study explores whether speakers have a voice-AI-specific register, relative to their speech toward an adult human. Furthermore, this study tests if speakers have targeted error correction strategies for voice-AI and human interlocutors. In a pseudo-interactive task with pre-recorded Siri and human voices, participants produced target words in sentences. In each turn, following an initial production and feedback from the interlocutor, participants repeated the sentence in one of three response types: after correct word identification, a coda error, or a vowel error made by the interlocutor. Across two studies, the rate of comprehension errors made by both interlocutors was varied (lower vs. higher error rate). Register differences are found: participants speak louder, with a lower mean f0, and with a smaller f0 range in Siri-DS. Many differences in Siri-DS emerged as dynamic adjustments over the course of the interaction. Additionally, error rate shapes how register differences are realized. One targeted error correction was observed: speakers produce more vowel hyperarticulation in coda repairs in Siri-DS. Taken together, these findings contribute to our understanding of speech register and the dynamic nature of talker-interlocutor interactions.

1. Introduction

With the rise of smartphones and smart speakers, millions of people now talk to a new type of interlocutor: voice-activated artificially intelligent (voice-AI) assistants (e.g., Apple's Siri, Amazon's Alexa, Google Assistant). Speech interaction with voice-AI has become a daily behavior for many people (Arnold et al., 2019). People of all ages use voice-AI for a variety of purposes, including to complete tasks (e.g., to play music, set timers, and other "internet of things" commands), request and search for information (e.g., "What's the weather?", "How many teaspoons are in a liter?") (Ammari et al., 2019; Bentley et al., 2018), as well as for playing games (e.g., "Tell me a joke") and engaging in conversational interaction (e.g., Amazon Alexa Prize chatbots in Ram et al., 2018). Unlike computers in the past, voice-AI is one of the only non-living entities that humans interact with using speech. Additionally, these systems are distinct in their cues of apparent humanity. For example, Apple's Siri has a name, apparent gender, a persona, a human-like voice, and improved speech recognition abilities compared to prior technology. Yet, whether humans have a systematically distinct way of talking to voice-AI assistants is an underexplored question.

Prior work has shown that speakers make systematic phonetic adjustments, known as register adaptations, when talking to different types of interlocutors. Speech directed toward typical adult interlocutors, adult-"directed speech" (DS)[1], has been shown to vary from speech directed toward other types of interlocutors, such as infant-DS (Fernald & Simon, 1984; Graf Estes & Hurley, 2013; Kuhl et al., 1997), non-native speaker-DS (Hwang et al., 2015; Lee & Baese-Berk, 2020; Uther et al., 2007), hearing-impaired individual-DS (Uchanski et al., 1996), pet-DS (Burnham et al., 1998; Burnham et al., 2002), and even computer-DS (Burnham et al., 2010; Oviatt et al., 1998; Stent et al., 2008). These patterns parallel some adjustments found in speech produced in noisy environments, which is louder and slower compared to speech produced in quiet conditions (e.g., "Lombard" speech

* Corresponding author.
E-mail addresses: mdcohn@ucdavis.edu (M. Cohn), bferencsegedin@ucdavis.edu (B.F. Segedin), gzellou@ucdavis.edu (G. Zellou).

[1] Note that "-DS" is applied to any type of register.

https://doi.org/10.1016/j.wocn.2021.101123
0095-4470/

in Brumm and Zollinger (2011)). At the same time, registers appear to vary depending on speakers' motives to improve intelligibility and/or convey emotional affect (Gergely et al., 2017; Hazan et al., 2015; Kitamura & Burnham, 2003; Trainor et al., 2000; Uther et al., 2007). For example, speech directed towards infants is slower, contains more vowel hyperarticulation than adult-DS, and has features associated with positive affect, including higher fundamental frequency (f0) and more f0 variation (Fernald, 2000; Kuhl et al., 1997). Speech directed toward hearing-impaired listeners is slower and contains less segmental reduction than conversational speech directed toward non-hearing impaired individuals (Picheny et al., 1986; Scarborough & Zellou, 2013). Still, the extent to which register adjustments improve communication varies. While some interlocutor-based adjustments are beneficial to the listener (Bradlow et al., 2003; Bradlow & Bent, 2002; Hargus Ferguson, 2004; Picheny et al., 1986), others are less advantageous (e.g., non-native speaker-DS is perceived negatively; for a review see Rothermich et al., 2019), suggesting that certain aspects of register adaptation can be attributed to speakers' *assumptions* about communicative barriers for certain types of interlocutors. Indeed, speech toward real versus imagined interlocutors contains different features (Scarborough et al., 2007; Scarborough & Zellou, 2013), supporting the notion that both presumed and authentic communicative difficulty leads to different types of speech adjustments.

Similarly, there is some evidence that speakers anticipate communicative difficulties when talking to various types of computer systems. There is a body of work characterizing the acoustic–phonetic patterns in computer-DS (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). However, few studies make a direct comparison between adult human-DS (henceforth "human-DS" in this paper) and speech towards computers/voice-AI; those that do vary in the acoustic features they measure and in their methodologies, as summarized in Table 1. For instance, Lunsford et al. (2006) observed that speech directed toward a computer system was perceived to be louder than that directed toward a real human interlocutor in a multi-party dialog, consisting of two humans and one computer. Other studies have found durational and hyperarticulation differences across computer- and human-DS. For instance, Burnham and colleagues (2010) recorded participants interacting with a human interlocutor (the experimenter through a computer screen) and a computer avatar in a scripted Wizard-of-Oz paradigm in which the interlocutors asked the participant questions about a narrative. They found that speech directed toward the computer avatar contained longer vowel durations and greater vowel space expansion than human-DS, yet they found no difference across the registers for f0. However, others have observed differences for f0. Mayo et al. (2012), for example, recorded a single talker reading sentences in five speech styles, plain, shouted, infant-, computer- and non-native speaker-DS, which were elicited through instructions (e.g., "Speak as though you were talking to a computer."). In computer-DS, they also found longer segment durations, but lower f0 range, relative to "plain" speech[2]. Still, not all studies find a general slowing of speech in computer-DS. Similar to the Burnham et al. (2010) study, Siegert et al. (2019) used a Wizard of Oz paradigm in a restaurant booking task, comparing spoken interactions with an apparent computer to that toward a human. They found that speakers produced shorter vowel durations in computer-DS, relative to human-DS.

While vowel hyperarticulation (or vowel space expansion) is well-studied in regards to infant-, non-native speaker-, and computer-DS registers (Burnham et al., 2002; Burnham et al., 2010; Uther et al., 2007), fewer studies have examined nasal coarticulation, or the degree of articulatory overlap between a nasal segment and a vowel (Chen, 1997). Indeed, none of the studies in Table 1 looked at coarticulatory patterns across computer- and human-DS. In prior work in human–human interaction, patterns of nasal coarticulation have been shown to vary across real and imagined listener conditions (Scarborough & Zellou, 2013), as well as in infant- and adult-DS (Zellou & Scarborough, 2015). Furthermore, many theoretical frameworks consider enhanced anticipatory coarticulation to be perceptually beneficial under certain circumstances (e.g., Beddor, 2009). Therefore, speakers might use coarticulation differently in computer- and human-DS to improve intelligibility. Taken together, these results highlight the necessity of examining multiple acoustic–phonetic features — at both the segmental and sentence-level — in classifying register adaptations.

1.1. Is there a voice-AI register?

The summary provided in Table 1 suggests that people vary in the way they speak to various computer systems or avatars. However, as mentioned above, voice-AI is a new type of system that is specifically designed to function via speech communication and the modern devices produce relatively naturalistic speech and language. There is some initial evidence of a distinctive *voice-AI* speech register. For example, speech directed toward humans versus voice-AI can be accurately classified via machine learning methods (e.g., Amazon's Alexa in Huang et al., 2019; Mallidi et al., 2018). Several studies have used the German Voice Assistant Conversation Corpus (VACC) (Siegert et al., 2018) to test differences in speech toward humans versus voice-AI (Raveh, Steiner, et al., 2019; Siegert & Krüger, 2021). In the corpus, German-speaking participants engage in unscripted conversations with an Alexa and a real human confederate to complete two tasks, setting an appointment on a calendar and completing a quiz. Both a physical Amazon Alexa Echo dot and human confederate were in the room. Both studies found that participants' speech was louder (i.e., had a higher intensity) in Alexa-DS, relative to human-DS (Raveh, Steiner, et al., 2019; Siegert & Krüger, 2021). Raveh et al. (2019) additionally reported differences in mean f0: participants showed higher pitch in speech addressed to Alexa. Here, they note that the gender of the human/Alexa voices may have contributed to the f0 differences, reflecting vocal alignment (in the VACC, the human

[2] The authors do not specify if this condition is read speech or speech directed toward an adult human interlocutor.

Table 1
Summary of technology-DS and human-DS studies. Note that "—" indicates that the feature was not measured. Differences are presented for speakers' productions after a misrecognition by their interlocutor.

Interlocutor	Study	Intensity/ Amplitude	Duration/rate	Segmental hyperarticulation	F0	Coarticulation
Computer vs. human	Lunsford et al. (2006)	Louder (perceived)	—	—	—	—
	Burnham et al. (2010)	—	Longer vowels	More vowel hyperartic.	No difference	—
	Mayo et al. (2012)	—	Longer segments	—	Smaller f0 range (no difference in median f0)	—
	(imagined) Siegert et al. (2019)	—	Shorter vowels	—	—	—
Voice-AI vs. human	Raveh et al. (2019) (VACC corpus)	Louder	No difference	—	Higher mean f0	—
	Siegert and Krüger, (2021) (VACC corpus)	Louder	No difference	Differs (No directionality reported)	Differs (No directionality reported)	—
	Cohn et al. (2021)	—	Slower rate	—	—	—

confederate is male, while the Alexa voice is female). Siegert & Krüger (2021), also using VACC, found differences in f0 and formant characteristics for Alexa- and human-DS, though no directionality was reported. Other work has examined more social interactions with voice-AI assistants. For example, in a user interaction study with an Amazon Alexa socialbot (where users can chitchat with Alexa about movies, food, animals, news, etc.), American English speakers produced significantly slower speech when talking to the system, relative to their baseline rate (assessed prior to the interaction) (Cohn et al., 2021).

While many prior studies use naturalistic interactions and similar tasks in human/voice-AI comparisons, it is possible that automatic speech recognition (ASR) errors could also contribute to participants' representations of the voice-AI communicative barriers. Since the conversations were spontaneous, the rate and type of errors made by the Alexa and human interlocutors could have varied. In the current study, we make a direct comparison between a set of target words in human- and voice-AI-DS (i.e., a female human vs. a female Apple Siri device interlocutor) elicited under identical conditions, and with the same rate and type of recognition "errors". A direct, controlled comparison is critical as the differences observed in other works may be task-related, or due to differences in linguistic or situational context, and not interlocutor-based per se. For example, Cohn and Zellou (2021) used a controlled experiment, where participants read target sentences from a screen and heard feedback from pre-recorded voices: an Alexa text-to-speech (TTS) voice and a recorded human voice. In 50% of trials, each interlocutor "heard" incorrectly, producing a vowel error (e.g., "bought" misheard as "bet"). Furthermore, in 50% of trials, the interlocutor used an emotionally expressive interjection congruent with the staged error (e.g., "Darn! I misunderstood…"). Overall, they found that speakers produced systematic differences in Alexa- and human-DS. Relative to their baseline productions (elicited prior to the interactions), speakers produced a slower speech rate, higher mean f0, and greater f0 variation in Alexa-DS. Meanwhile, no differences were observed in error repairs for the two interlocutors. Taken together, these differences suggest that speakers might have a voice-AI register, which may be due to different expectations for how likely voice-AI and humans are to understand them. Indeed, Siegert and Krüger (2018) found that participants rated their interactions with a human confederate to be "intuitive" and that they could "[speak] like [they] always do", while interactions with Alexa were reported as "more difficult" and "different to interacting with someone in the real world" (p. 6). The current study extends the controlled approach described in Cohn & Zellou (2021), varying the interaction by error rate for non-emotional interactions with voice-AI and human interlocutors.

1.2. Theoretical relevance for human/voice-AI comparison

Examining speech toward voice-AI and human interlocutors can serve as a lens into mechanisms of speech adaptation. On the one hand, if global speech patterns are distinct for the two interlocutors, this would support *listener-intelligibility accounts* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998). For example, Audience Design theory proposes that speakers make linguistic adjustments based on the (presumed) communicative needs of their interlocutor (Clark & Murphy, 1982). There is work which suggests that people presume that computer systems will exhibit greater intelligibility barriers, and less overall communicative competence, relative to humans (Cowan et al., 2015; Oviatt, MacEachern, et al., 1998). For example, in typed interactions, participants display greater syntactic and lexical alignment toward (apparent) computer interlocutors, relative to (apparent) humans, in situations where alignment will lead to greater communicative success in the interaction (Branigan et al., 2011, 2003). Support for this account comes from speech production as well: as summarized previously, computer-DS is often louder (Lunsford et al., 2006), slower (Burnham et al., 2010; Mayo et al., 2012), and more hyperarticulated (Burnham et al., 2010) than human-DS, which suggest that the user assumes the computer has greater perceptual difficulties.

On the other hand, if speakers do not have a different register for voice-AI and human interlocutors, this would support *technology equivalence accounts* (Nass et al., 1997, 1994; Reeves & Nass, 1996). For example, the Media Equation Theory (Chiasson & Gutwin, 2005; Hoffmann et al., 2009; Reeves & Nass, 1996) proposes that people engage with media (e.g., technology) the same way they do with people. This argument is based on the observation that people engaging with technology have a sense of physical or social presence and,

consequently, respond with the same behaviors as they do for humans (for review, see Lee, 2008). Another *technology equivalence account*, the Computers are Social Actors (CASA) theory (Nass et al., 1997, 1994), proposes that people subconsciously — and categorically — apply social behaviors from human-human interaction to those with a computer, given cues of "humanity" in that computer. In voice-AI, there are ample cues of humanity, including a relatively naturalistic voice, advanced speech recognition, and the device talker having a name (e.g., "Alexa", "Siri"). Indeed, there is some support for *technology equivalence accounts* for linguistic behavior toward voice-AI. For instance, several recent studies have shown that people vocally align toward both voice-AI and human interlocutors (Cohn et al., 2019; Raveh, Steiner, et al., 2019; Snyder et al., 2019; Zellou, Cohn, & Ferenc Segedin, 2021; Zellou, Cohn, & Kline, 2021), and even display similar gender-based speech asymmetries (such as aligning more to male, than female, TTS and human voices in Cohn et al., 2019). Hence, an alternative prediction in the current study, based on *technology equivalence accounts*, is that speech patterns to voice-AI and human interlocutors will not differ.

1.3. Different error correction strategies for interlocutors?

Given that one of our competing predictions is that global features of human- and voice-AI-DS might overlap, we also examine interlocutor-driven differences in response to local communicative pressure: when the interlocutor mishears them. While under-explored, differences in error correction strategies might be a key component of an interlocutor-based register, which can inform theoretical accounts of speech adaptation. *Targeted adaptation accounts* (Buz et al., 2016; Lindblom, 1990) propose that speakers dynamically adjust their output in order to address local communicative demands. For example, the Hyper- and Hypo-articulation (H&H) model proposes that speakers adjust their pronunciation to produce clear speech ("hyperspeech") when they see that the interaction is more demanding for the listener; otherwise, speakers conserve articulatory effort and produce more casual speech ("hypospeech") (Lindblom, 1990). Repeated, second mention, and more predictable words are produced with greater reduction (hypospeech) (Fowler & Housum, 1987; Lieberman, 1963). Meanwhile, explicit misunderstanding has been shown to trigger hyperspeech. For example, speech produced to improve intelligibility, has been shown to have higher intensity, longer segment durations, slower speech rate, and larger f0 range than casual or conversational speech (Bradlow et al., 2003; Picheny et al., 1986). Clear speech has also been shown to contain more vowel hyperarticulation and less segmental coarticulation than in connected, casual speech (Moon & Lindblom, 1994). The Adaptive Speaker Framework (Buz et al., 2016), another *targeted adaptation account*, similarly proposes a trade-off based on real-time difficulties, but it proposes that adjustments are segmentally targeted to the phonological source of confusion. Indeed, targeted hyperarticulation in response to specific phonological confusions has been found in other studies (Baese-Berk & Goldrick, 2009; Oviatt, MacEachern, et al., 1998; Schertz, 2013), suggesting that speakers have representations for how to adapt their speech in real-time in order to be best understood.

There is some evidence that speakers target different acoustic–phonetic adjustments for different interlocutors, lending support for a hybrid account of *listener-intelligibility* and *targeted adaptation accounts*. In human–human interaction, speakers show targeted adjustments on words for adults and infants based on specific lexical properties. For example, in adult-DS, degree of hyperarticulation and nasal coarticulation appear to be driven by how many lexical competitors the word has, and hence a high potential for confusability based on an adult lexicon measure (phonological neighborhood density) (Scarborough, 2013). For infant-DS, these same phonetic adjustments were made for words with a later age-of-acquisition (AoA) (Zellou & Scarborough, 2015). Assuming that AoA is a better metric of lexical difficulty for infants and likewise neighborhood density for adults, these findings suggest more generally that speakers respond to the presumed needs of their interlocutor by making precise, acoustic–phonetic enhancements for specific phonological confusions by that listener. Will speakers differ in how they correct phonological errors made by a voice-AI versus a human interlocutor?

Table 2 summarizes studies with evidence of targeted error corrections in spoken interaction with computer systems. Together, they show some similarities to strategies used in human–human interaction (Bell & Gustafson, 1999; Maniwa et al., 2009; Ohala, 1994; Oviatt, Levow, et al., 1998; Oviatt, MacEachern, et al., 1998; Schertz, 2013; Stent et al., 2008; Swerts et al., 2000; Vertanen, 2006). For instance, Bell & Gustafson (1999) analyzed a corpus of people's spontaneous interactions with a spoken dialog system, selecting utterances that were lexically identical in the first production and a subsequent repetition in response to an ASR error. In response to an ASR error, speakers produced louder, slower, and more hyperarticulated speech. Similarly, Schertz (2013) found that speakers repair errors made by an apparent speech recognition system (e.g., "pit" misheard as "bit") by increasing the duration of voice onset time (VOT) on the voiceless stop, enhancing a property that differentiates /p/ and /b/ in English. In a study where participants spoke to a dialog system, Stent et al. (2008) found that speakers displayed hyperarticulation by slowing their speech rate and producing more canonical segmental forms of /t/ and /d/ in target words, in responses produced after the computer recognition errors. Consistent with *targeted adaptation accounts*, hyperarticulation lingered after the error correction, but speakers eventually reverted to their original speaking style. Yet, not all adjustments in human–computer interactions are consistent with work in human–human interaction. For example, Maniwa et al. (2009) found that speakers responded to a computer error to their VCV syllables by using some of the same clear speech adaptations observed in human-DS (e.g., increasing duration), but with a surprising difference: error repair trials were produced with *less* intensity.

Fewer prior studies have directly compared error correction to human and to computer interlocutors. Burnham et al. (2010) found no differences in error correction between human- and computer avatar-DS, with staged errors occurring in 33% of trials for both interlocutors. Similarly, Cohn and Zellou (2021) found differences in prosodic characteristics of Alexa- and human-DS, but no targeted differences to staged misrecognitions (occurring in 50% of trials). In both human- and Alexa-DS, they observed greater vowel backing in response to a

Table 2
Summary of error correction in computer-DS. Note that "—" indicates that the feature was not measured. Differences are for error correction relative to non-correction productions.

Interlocutor	Study	Intensity/ Amplitude	Duration	Segment hyperarticulation	F0	Coarticulation
Computer only	Stent et al. (2008)	—	Slower rate	Consonant hyperartic. & front vowels (more fronted)	—	—
	Oviatt, MacEachern, & Levow (1998)	No difference	Longer overall; longer segments, longer pauses; slower rate	Consonant hyperartic.	Lower f0 (in lower error rate)	—
	Oviatt, Levow, Moreton, MacEachern (1998)	Louder	Longer overall; longer segments, more pauses; longer pauses;	Consonant hyperartic.	F0 max. and f0 range (N.S.)	—
	Ohala (1994)	—	Longer	Increased consonant hyperartic. (VOT)	—	—
	Bell & Gustafson (1999)	Louder	Slower rate	More vowel hyperartic.		—
	Maniwa et al. (2009)	Quieter	Longer	Higher frequency spectral peaks	Higher f0	—
	Vertanen (2006)	Quieter	Slower rate, more pauses	Increased formant frequencies	Greater f0 range	—
	Swerts et al. (2000)	Louder	Longer, longer pauses, less internal silence	—	Larger f0 variation	—
	Schertz (2013)	—	Longer vowels (for phonemic contrasts e.g., /ɪ/-/i/)	Vowels N.S. Increased consonant hyperartic. (VOT)	—	—
Computer vs. human	Burnham et al. (2010)	—	Longer segments (computer-/human-DS N.S.)	Vowel space expansion (computer-/human-DS N.S.)	Larger F0 range (computer-/human-DS N.S.)	—
Voice-AI vs. human	Cohn & Zellou (2021)	Louder (Alexa-/ human-DS N.S.)	Slower rate (Alexa-/human-DS N.S.)	F2 hyperartic. (Alexa-/human-DS N.S.)	Higher f0, larger f0 range (Alexa-/human-DS N.S.)	—

vowel error, as well as louder, slower rate, higher mean f0, and greater f0 variation. Still, both studies had relatively high error rates (33% and 50%), which might have led to more similar error adaptation strategies for the interlocutors.

1.4. Current study

The current study tests whether there is an overall voice-AI-DS register, compared to human-DS, and whether there are differences in targeted error-repair strategies across interlocutors. The production study consists of a pseudo-interactive task with two types of interlocutors in a laboratory setting: a voice-AI system (here, Apple's Siri[3]) and a native English speaking adult. The task was carefully controlled, holding the nature of the interaction, as well as rate and type of errors, constant across the two interlocutor types. Building off of related work examining productions toward an apparent speech recognition system (e.g., Schertz, 2013), participants produced target words and received visual feedback as to what the interlocutor "heard". Additionally, participants heard audio recordings of the interlocutor during each trial to create a four-turn interaction (see General Procedure in Section 2.2. and a trial schematic in Fig. 2). Participants produced the target words in an initial production (Original productions) and then responded to the interlocutor across three feedback conditions: repeat after the interlocutor indicated the correct target word (Correct Repeat), repeat following a coda consonant error (Coda Error: e.g., "Be**n**" misheard as "be**d**"), and repeat following a vowel error (Vowel Error, e.g., "b**e**t" misheard as "b**oa**t"). We measured features commonly investigated in computer-DS registers, including sentence intensity, mean f0, f0 range, as well as acoustic–phonetic properties of the target word vowels, including vowel duration, coarticulatory nasalization, and hyperarticulation.

Additionally, the current study tests whether differences for the Siri and human interlocutors (if present) change over the course of the interaction. Speakers' a priori expectations for how well a voice-AI (here, Siri) versus a human can understand them (i.e., "presumed competence") could be a factor in their speech behavior, and might also change as the speaker accumulates experience with that interlocutor. Cowan et al. (2015) showed that listeners have different presumptions of communicative competence for naturally-produced versus synthesized voices. In particular, people believe that a computer interlocutor is a less competent conversational partner than a human interlocutor, based on hearing their voice alone. Thus, speakers' beliefs about the communicative competence of Siri versus a human interlocutor might explain differences in their global register and targeted error correction patterns, especially at the beginning of the interaction. However, as the interaction unfolds, participants accumulate real-time evidence about the interlocutor's actual competence based on how well the interlocutor is "understanding" target words and the types and rates of errors they make. By varying the overall rate of errors occurring across trials (lower vs. higher), the effect of the computer interlocutor's "actual competence" might lead to differences in clear speech adjustments. For example, there are different effects for "real" versus "imagined" interlocutors observed in prior work (Scarborough et al., 2007; Scarborough & Zellou, 2013). In a

[3] The current study examines speech toward Siri and generalizes this to "voice-AI" directed speech, however, we acknowledge that voice-AI register patterns could vary across systems. This is a question for future work.

Table 3
Acoustic properties of the Human and Siri interlocutor productions used in this study.

	Human	Siri	Pairwise comparison
Speech rate	2.71 syll/sec (0.22)	2.39 syll/sec (0.26)	$t(21.44) = -0.91$, $p = 0.37$
F0 mean	17.78 ST (0.35)	13.63 ST (0.23)	$t(22.59) = -7.56$, $p < 0.001$
F0 range	2.77 ST (0.42)	3.13 ST (0.30)	$t(23.09) = 0.70$, $p = 0.49$

lower error rate condition, participants may be more reliant on a priori beliefs about the interlocutors' competence, which might lead to more distinct patterns across Siri- and human-DS. On the other hand, in a higher error rate condition, it is possible that speakers might show more similar adjustments in response to feedback from the two types of interlocutors as seen in some recent work (e.g., 50% error rate in Cohn & Zellou, 2021). Thus, in the current study, participants completed one of the two error rate conditions of the paradigm: a lower error rate (Experiment 1, Section 3) or a higher error rate (Experiment 2, Section 4).

1.4.1. Predictions

We can set up several predicted outcomes for the present study, based on the different theoretical accounts of speech adaptation. In terms of speech register, observing differences across Siri- and human-DS is consistent with *listener-intelligibility accounts*. Conversely, if speech patterns are the same, this supports *technology equivalence accounts*. For error correction responses, the extent to which we see different strategies to repair the vowel and coda errors can speak to *targeted adaptation accounts*. Here, we examine nasal coarticulation as a way speakers could enhance cues to the final nasal coda on the vowel when interlocutors mistake the final consonant (e.g., "be**d**" vs. "Be**n**"). Meanwhile, if interlocutors make vowel-phoneme errors, we predict this would be more likely to trigger vowel lengthening and vowel hyperarticulation (e.g., Bell & Gustafson, 1999). Furthermore, if targeted adjustments differ for the Siri and human interlocutors, this would support a hybrid *listener-intelligibility* and *targeted adaptation account.*

2. General methods

2.1. Stimuli. Interlocutor recordings

To create a pseudo-interactive task, pre-recorded productions by the human and Siri interlocutors were played during the experiment. At the start of each interlocutor block, the respective interlocutor introduced themself (see Appendix A), and provided voice-over instructions for the task. During the experimental trials, interlocutors provided responses to the initial production by the participant ("Did you say this word?" "Is this correct"? "Is this right?" "Is this the word?"), as well as a final response to the participants' second production ("Good", "Got it", "Great", "I think I get it now", "Okay, got it."). The TTS output for the Siri voice was generated using the command line on an Apple computer (OSX 10.13.6) with the "Samantha" voice (American female)[4]. For the human interlocutor, a female native California English speaker produced the recordings in a sound attenuated booth wearing a head-mounted microphone (Shure WH20 XLR). Recordings from Siri and the human female were amplitude normalized (60 dB). A summary of the acoustic properties of the interlocutors' productions is provided in Table 3, as well as the results of a t-test run comparing each feature across the Siri and human speakers. As seen, the human and Siri talkers did not vary in terms of overall speech rate or f0 range (shown in semitones, ST, relative to 75 Hz). However, overall, the human voice had a higher average f0 than the Siri voice.

2.1.1. Perceived human-likeness and communicative competence ratings

To assess differences in perceived human-likeness and communicative competence of the human and Siri voices, we conducted a perceptual ratings study of the interlocutors' voices.

2.1.1.1. Stimuli, participants, & procedure. Stimuli consisted of recordings produced by the Siri and human interlocutors (see Appendix A for full list), which were used in both Experiments 1 & 2 as the voice-over overview ($n = 4$ utterances), immediate follow-up responses ($n = 4$ utterances), and closing productions ($n = 5$ utterances) (the procedure for the speech production experiments is described in Section 2.2.2.). We did not present listeners with the initial introduction by the interlocutors as they differed for the two talkers ("Hi! [I'm Siri. I'm a digital assistant on Apple products.] | [I'm Melissa. I work here in the Phonetics Lab.]"). In total, there were 26 utterances evaluated (13 utterances * 2 interlocutors). Stimuli were amplitude normalized to 65 dB. This was louder than the in-lab experiments (Experiments 1 & 2) as the audio were presented online and participants would complete the experiment from home.

Twenty-three UC Davis undergraduates (mean age = 19.6 years; 19 female, 4 male), none of whom participated in the production studies, completed the rating study and received course credit for their participation. Participants completed the study online, using the Qualtrics experiment platform. Participants first completed an audio calibration check, where they could play a sentence (up to five times) and were asked to identify which target word they heard from three phonologically related options ("Bill heard we asked about the coast", "Bill heard we asked about the host", "Bill heard we asked about the toast"). Then, participants completed the human-likeness and communicative competence ratings blocks. The order of blocks was randomly selected for each participant. In the human-likeness block, participants heard each audio recording by the two interlocutors (randomly presented, one at a time) and were asked to rate "How human-like does the voice sound" on a sliding scale (0 = not human-like, 100 = extremely human-like). In the competence block, participants heard each item (randomly presented, one at a time) and rated "How well do you think this speaker would understand you?" on a sliding scale (0 = not well, 100 = extremely well).

[4] Since the time of this study, the Siri female voice has been updated by Apple. Both the "original" and "updated" voices are still available on Mac OS 11.0.1, both listed as "Siri Female".

2.1.1.2. Analysis & results. A mixed-effects linear regression model was fit separately for each of the two ratings, human-likeness and communicative competence. Both models contained a fixed effect for Interlocutor (Human vs. Siri), as well as by-Listener random intercepts and by-Listener random slopes for Interlocutor. Estimates for F-statistics, and p-values were computed using Satterthwaite approximation in the *lmerTest* package (Kuznetsova et al., 2017). Contrasts were treatment coded.

Fig. 1 provides the mean human-like and competence ratings for the Siri and Human voices. In the human-like ratings model, Interlocutor was a significant predictor, with the Siri productions rated as significantly less human-like than the Human productions [*Coef* = −86.04, *t* = −30.00, *p* < 0.001]. The communicative competence ratings model also revealed a difference by Interlocutor: participants rated that the Siri talker would be much less likely to understand them, compared to the Human voice [*Coef* = −39.67, *t* = −6.68, *p* < 0.001].

The results of the ratings task demonstrate that the Siri and Human voices are distinct in their perceived human-likeness as well as their perceived communicative competence. This manipulation check confirms that from the voice properties alone, the Human and Siri interlocutors have distinct apparent humanness and perceived communicative capabilities. In the experiments that follow, we test whether participants speaking to these two voices produce different register adaptations (voice-AI- vs. human-DS) and error correction strategies.

2.2. General methods. Production studies

2.2.1. Stimuli. Target words

Target words (*n* = 55) consisted of monosyllabic English words. To assess participants' vowel space, 12 CVC words consisting of corner vowels /i/ and /a/ (see Appendix B) (e.g., "BEAD") were selected. We additionally selected 44 words: 22 contained a nasal coda (CVN) and 22 contained an oral coda (CVC)[5], all of which have both a different-vowel minimal pair, contrasting in vowel backness with the target word, and a different-coda minimal pair, contrasting in coda nasality (target words provided in Appendix B). These minimal pairs were used for the two misrecognition conditions, in order to compare participant productions following incorrect vowel and incorrect coda errors made by the interlocutors. Target words were chosen to have a non-high vowel, since the acoustic nasality measurements made for this study are more accurate when F1 does not overlap with the first or second harmonic (acoustic measurements described in Section 2.3.1.).

2.2.2. General procedure

Participants in the production studies (Experiments 1 & 2) completed the experiment in a sound attenuated booth, seated in front of a Dell computer monitor and E-Prime button box (SRBOX) while wearing headphones (Sennheiser Pro) and a head-mounted microphone (Shure WH20 XLR) positioned to the right of their mouth. Recording levels were set with a protocol that was identical for all participants, with a +20 dB gain (ART Tube MP Preamplifier) and no additional gain when passed through the Focusrite Scarlett Mixer[6]. Once set up, participants were given some background information about the study. Participants were told they would be talking to a real human, "Melissa", and a digital device, Apple's Siri. Interactions with a given interlocutor were presented in a single block and order of interlocutor block (Human first or Siri first) was counterbalanced across participants. In both blocks, participants heard an introduction from the interlocutor along with an image, either an iPhone showing the "How can I help you?" screen (Siri block) or a stock image of a human female (Human block). Next, they were presented with an example trial with a voiceover (either Siri or Human voice) (all pre-scripted dialogue is provided in Appendix A).

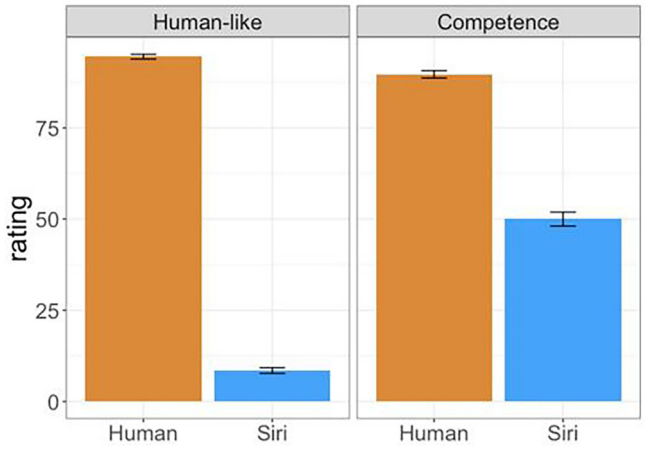

Fig. 1. Mean ratings of human-likeness and communicative competence for the Siri and Human voices used in the production study. Error bars show the standard error of the mean.

Each experimental trial consisted of four parts, schematized in Fig. 2: (1) Participants read the sentence they saw on the screen (e.g., "The word is Todd"), their "original" production. The slide was shown for 4000 ms with an output sound function that automatically generated the 4000 ms audio recording. (2) Participants saw a screen with a word written in red as they heard the interlocutor's voice asking for feedback (e.g., "Is this correct?" "Is this right?" etc.). If the word was correct (e.g., TODD), participants pressed "YES" on a labeled E-Prime button box. If the word was incorrect (e.g., TAD), participants responded "NO". The next screen would advance when participants made a button press. (3) Participants then repeated the sentence again (e.g., "The word is Todd") as their second production (correct repeat or error correction), with the slide shown for 4000 ms. (4) Finally, the interlocutor (Siri or female Human) gave positive feedback (e.g., "Great", "I think I get it now", "Got it", etc.). Interlocutor verbal responses were pre-recorded. Feedback and final responses were randomly selected from a set of options throughout the experiment (see Appendix A). These responses were designed to make the trials seem more interactive. There was an intertrial interval of 1000 ms.

In each interlocutor block, participants began with 12 vowel space trials consisting of CVC words with the corner vowels /i/ and /a/ (order randomized). Following the 12 initial trials, partic-

[5] One word, "Dodd" was used in both the vowel space and CVC error word lists.

[6] Once the gain was set in the pre-experimental set-up, it was not changed again during an entire participant recording session.

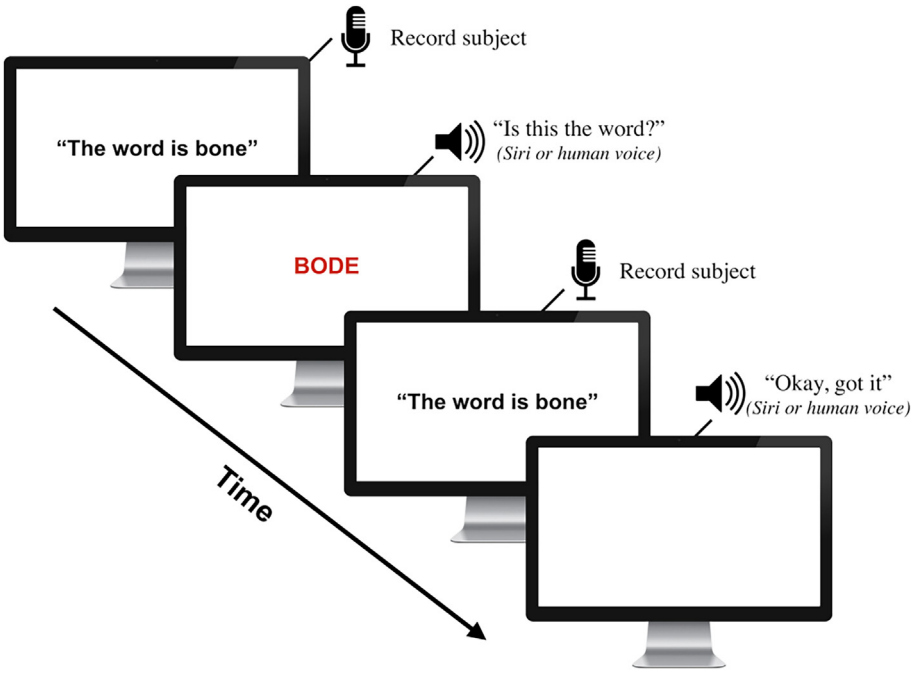

Fig. 2. Schematic of an experimental trial. The first slide showed the target for the original production. The second slide either showed the correct response (e.g., BONE) or the word with an incorrect coda (e.g., BODE) or vowel (e.g., BANE). The third slide showed the target for the second production in the trial (either repeat correct, or error correction).

ipants continued seamlessly to the experimental trials, consisting of 22 CVC and 22 CVN trials (alternating between CVC and CVN on each trial; target words randomly selected from each word type). The target word lists and corresponding errors are provided in Appendix B.

Staged recognition errors were matched, both in type of error and rate of error for the human and Siri interlocutors. For the vowel space trials, errors consisted of incorrect coda voicing (e.g., "BOD" for "bot"). For experimental trials, there were two types of errors: incorrect vowels and incorrect codas. Incorrect vowel errors consisted of the opposite vowel backness (e.g., "TAD" for "Todd"). Incorrect coda errors consisted of oral consonant codas in place of the nasal consonant codas at the same place of articulation (e.g., "BODE" for "bone"). In "correct" trials, the interlocutor (Siri/human) correctly heard the intended target word. Correspondence of interlocutor response (correct, vowel error, coda error) to target words were randomized in each interlocutor block. In total, participants completed 56 trials for each interlocutor, for a total of 112 trials (56 trials * 2 Interlocutors). The experiment lasted approximately 35 minutes in total. After the study, participants completed a background questionnaire about their demographics, language background, and their voice-AI usage.

Rate of errors was varied across experiments: at a lower error rate (Experiment 1) and a higher error rate (Experiment 2). More details on the specific procedure for each experiment are provided in Sections 3.1.3. and 4.1.3.

2.3. Analyses

2.3.1. Acoustic measurements

The second author listened to each of the recordings, ensuring that (1) the speaker indeed responded, and (2) they were not doing anything particularly marked or out of the ordinary. We excluded trials where there was any artifact (e.g., coughing, yawning). There was only one speaker who did not say the target sentences for Siri in the vowel space trials. Consequently, we excluded all of their data from analysis.

Four prosodic measurements were made at the sentence level. Mean intensity over the sentence was measured in decibels (dB) with a Praat script (using the "Get Intensity (dB)" function). Mean f0 measurements were calculated at 15 equidistant intervals over the sentence with a script that uses Praat's default autocorrelation method (that we adapted from DiCanio, 2007). We filtered f0 values for plausible maxima and minima by speaker gender (males: 78–150 Hz, females: 150–350 Hz) to remove spurious values and those resulting from vocal creak[7]. The mean of these values (converted to semitones, ST) served as the overall "mean f0" of the sentence. Based on the f0 measurements taken over 15 intervals, we calculated the utterance maximum and minimum f0 (in ST). Their difference (maximum – minimum) was the f0 range. We additionally measured speech rate (syllables per second) for each sentence with a Praat script (De Jong et al., 2017) to use as a predictor in vowel duration models.

Participants' utterances were force-aligned using FAVE (Rosenfelder et al., 2014). Segmentations for target word vowels were manually hand-corrected by trained phoneticians. Hand-correction focused on vowel-consonant boundaries in the target words. The boundary between the vowel and the adjacent consonants was determined to be where there was an abrupt change in amplitude in the waveform and an abrupt change in amplitude of the higher formant frequencies in the spectrogram. Note that speakers produced the sentence-final target words with creaky voice in some cases; following Pycha and Dahan (2016), we confirmed all still had visible glot-

[7] In the lower error rate study (Experiment 1), this resulted in removing 1 production above the plausible maxima and 56 below the plausible minima for speaker genders. In the higher error rate study (Experiment 2), this resulted in removing 20 productions above the plausible maxima and 49 below the plausible minima for speaker genders.

tal pulses[8]. Following hand-correction, vowel duration and formant frequency values (F1 and F2) were measured with FAVE extract (Rosenfelder et al., 2014).

For CVN targets, coarticulatory nasalization was measured acoustically within-speaker as A1–P0, a spectral measure of relative degree of vowel nasalization using a Praat script (Styler, 2017, 2018). A1–P0 is quantified as the difference in amplitudes between the F1 peak and a low frequency nasal peak (P0) in the spectrum (Chen, 1997). As nasalization is introduced, the relative amplitude (in dB) of the nasal formant peaks increases while the relative amplitudes of oral formant peaks (e.g., F1) decreases. The difference in amplitude between the nasal formants and the oral formants is a measure of relative nasalization; A1–P0 decreases as nasality increases. A1–P0 measurements for all CVN words were made automatically via script in Praat, taken at vowel midpoint following Scarborough & Zellou (2013).

To assess the degree of vowel hyperarticulation, we calculated degree of F1/F2 Euclidean distance from each speaker's vowel space center. First, we log transformed (log-base 10) the first and second formant frequencies (F1 and F2) to scale from Hertz (Hz). Next, we calculated each subject's average (logged) F1 values and (logged) F2 values from the corner vowels /i/ and /a/ taken from the first 12 trials of each block (following Bradlow et al., 1996). We used these subject-means for F1 and F2 to center the observations for each speaker, resulting in "log mean normalized" F1 and F2 values (Nearey, 1978). The mid central vowel, /ʌ/, was excluded from the vowel hyperarticulation analyses (as was done in Wedel, Nelson, & Sharp, 2018). Euclidean distance from each speakers' vowel-space center was calculated from each participants' F1 and F2 at 35% of the total vowel duration (Bradlow et al., 1996; Smiljanić & Bradlow, 2005). Examining vowel features at roughly ⅓ of the vowel portion is a common approach taken in sociophonetics (e.g., Fridland et al., 2014), which additionally addresses dynamic formant movement observed for both monophthongs and diphthongs in American English (Fox & Jacewicz, 2009; Nearey, 2013). In monophthongs, the initial half of the vowel is characterized as the most stable (Fox & Jacewicz, 2009; Hagiwara, 2005). The initial, onglide portion of diphthongs is also observed to be the most critical and stable portion (Gottfried & Triesch, 1993; Nearey & Assmann, 1986). Therefore, we assess differences at 35% of vowel duration in order to have the most stable and consistent formant measure across monophthongs and diphthongs.

2.3.2. Statistical analyses

Each of the acoustic properties of interest (intensity, f0 mean & range, vowel duration, vowel hyperarticulation, acoustic vowel nasality) were analyzed using separate linear mixed effects models with the *lme4* R package (Bates et al., 2015).

Fixed effects included Interlocutor (2 levels: Human, Siri), which was sum coded such that differences reflect changes in Siri- and human-DS from the grand mean. We also included the fixed effect of Production Type (4 levels): (1) original: their first production of the sentence (prior to direct feedback from the interlocutor), (2) correct repeat: their repetition of the sentence following correct feedback, (3) incorrect coda: their repetition following an incorrect coda, and (4) incorrect vowel: their repetition following an incorrect vowel. As the number of observations for the Production Type levels varied (with more observations for "original" than the others), we used weighted effect coding for the Production Type factor using the *wec* R package (Nieuwenhuis, te Grotenhuis, Pelzer, et al., 2017). Similar to sum coding, weighted effect coding is used to determine if factor levels differ from the grand mean; but it additionally weighs estimated means (both grand mean and factor mean) based on the number of observations of each level of a factor (Nieuwenhuis, 2016; Nieuwenhuis, te Grotenhuis, & Pelzer, 2017). We also included the fixed effect of Trial Number (standardized), and two-way interactions between Interlocutor and Production Type, and Interlocutor and Trial Number. Models included a maximal random effects structure, with by-Word and by-Participant random intercepts, and by-Participant random slopes for Interlocutor and Production Type (*lmer* syntax provided in Eq. (1)).

$$\begin{aligned}(feature \sim \text{Interlocutor} * (\text{Production Type} + \text{TrialNumber}) \\ + (\text{Interlocutor} * \text{ProductionType}|\text{Participant}) \\ + (1|\text{Word})\end{aligned} \quad (1)$$

In order to account for the effect of temporal factors on segmental phonetic effects, we included the fixed effect of Vowel Duration (centered) in the acoustic nasality and vowel hyperarticulation models to account for duration-based effects (e.g., Zellou & Scarborough, 2019). In the vowel duration model, we additionally included a fixed effect of Speech Rate (centered). The acoustic nasality models were only run on target CVN words (n = 22 in each block).

For all acoustic feature models, we first attempted to fit a complex random effects structure (with by-Participant random slopes for Interlocutor, Production Type, and their interaction) to account for inter-subject variability (Barr et al., 2013). We defined a systematic approach (adapted from Barr et al., 2013) for simplifying random effects structure in response to a singularity or convergence error, which can indicate that the model has been "overfit".

1) Examine random effect variance. If a single predictor results in a value close to "0", try removing it. If it does not improve fit, keep it in.
2) Remove by-Participant random slopes for interaction (Interlocutor * ProductionType|Participant) → (Interlocutor + ProductionType|Participant).
3) Compare results when selectively removing each random slope in two models, retaining random slopes for Interlocutor in one and random slopes Production Type in another.
4) If removing neither of the random slopes improves convergence, remove both — leaving only by-Participant and by-Word random intercepts.

We fit each model first with the same Production Type factor leveling (omitted level = "Correct Repeat"), such that the model output would show effects for "original", "incorrect coda", and "incorrect vowel" productions. In order to uncover the effect of "correct repeat" productions, we re-ran each model, holding the structure constant, with the releveled factor for Production Type (omitted level = "Original") (contrasts are provided in Supplementary Data 1 for Experiment 1 and Supplementary Data 2 for Experiment 2) (Schad et al., 2020). For each acoustic feature, we provide both versions of the model outputs in the

Experiment 1 and Experiment 2 Supplementary Data 1 and 2 (Tables 10–21).

3. Experiment 1. Lower error rate

3.1. Methods

3.1.1. Participants

Participants (n = 30) consisted of native English speaking adults recruited from the UC Davis Psychology Pool (mean age = 19.7 ± 1.6 years, age range = 18–24 years, 25 females, 5 males). All reported having had experience with voice-AI systems and having had no hearing impairments. All participants gave informed consent to participate, in pursuance with the UC Davis Institutional Review Board.

3.1.2. Stimuli

Described in Section 2.2.1.

3.1.3. Procedure (lower error rate study)

The general procedure is described in Section 2.2.2. The overall error rate for Experiment 1 was 14.3% (errors on 8/56 trials; 48 correct trials). In the 12 vowel space trials, Siri/Human always showed the correct word. In the experimental trials (22 CVC and 22 CVN target words), the interlocutors made eight staged recognition errors: four coda errors and four vowel errors (order and correspondence to word were randomized). In the remaining trials, the interlocutor heard "correctly", showing the correct target word to the participant.

3.1.4. Analysis

See general methods for acoustic and statistical analysis in Section 2.3.

3.2. Results

Model outputs are provided in the Supplementary Data 1 (Tables 10–21); sentence-level results are shown in Fig. 3, while vowel-level results are displayed in Fig. 4.

3.2.1. Sentence-level results

Mean intensity is plotted in Fig. 3A. The models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant) + (1|Word)) showed an effect of Interlocutor, demonstrating that speakers talk louder to Siri than to the human interlocutor [*Coef* = 1.63, t = 2.24, p < 0.05]. There was also an effect of Production Type: participants produce louder original productions [*Coef* = 0.74, t = 7.34, p < 0.001], but quieter coda repair productions [*Coef* = −2.51, t = −3.29, p < 0.01] and vowel repair [*Coef* = −2.40, t = −3.15, p < 0.01], relative to the weighted grand mean. The model with Production Type releveled (omitted level = "Original") also showed lower intensity in correct repeat productions [*Coef* = −0.61, t = −5.60, p < 0.001]. Interlocutor also interacted with Production Type: in Siri-DS, participants produce even louder original productions [*Coef* = 0.51, t = 5.08, p < 0.001] yet less of an increase for coda repair [*Coef* = −1.73, t = −2.30, p < 0.05], vowel repair [*Coef* = −1.81, t = −2.41, p < 0.01], and correct repeat [*Coef* = −0.42, t = −3.84, p < 0.01] productions. We also observed changes by Trial Number, wherein participants' productions become louder over time within each block [*Coef* = 0.36, t = 2.56, p < 0.05]. The way intensity changes over time is also shaped by Interlocutor. The interaction between Interlocutor and Trial Number, depicted in Fig. 3A, demonstrates that participants become louder in Siri-DS over time [*Coef* = 0.64, t = 6.31, p < 0.001]. Note that throughout the paper, to better visualize the over-time effects, the plots present the data across the "First" and "Second" portions of each block.

Mean f0 is plotted in Fig. 3B. The mean f0 models including by-Participant random intercepts for Production Type resulted in singularity errors, while by-Word random intercepts resulted in non-convergence. The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant)) showed no differences by Interlocutor. There was an effect of Production Type, with higher mean f0 in original productions [*Coef* = 0.16, t = 15.36, p < 0.001], and lower mean f0 in vowel repair productions [*Coef* = −0.17, t = −2.23, p < 0.05]. There was no difference in mean f0 for coda repairs, relative to the weighted grand mean. When Production Type was releveled (omitted level = "Original"), correct repeat productions had lower mean f0 overall [*Coef* = −0.17, t = −14.82, p < 0.001]. No other predictors or interactions were significant in the f0 model. As there were no effects of Trial Number, the plot is not faceted by "first" and "second" half.

Fig. 3C. summarizes the values for f0 range. The models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant) + (1|Word)) showed an effect of Interlocutor, indicating that speakers produce a smaller f0 range when talking to Siri than to the human interlocutor [*Coef* = −0.12, t = −2.27, p < 0.05]. There were also effects of Production Type wherein participants produce a smaller f0 range in original productions [*Coef* = −0.06, t = −2.57, p < 0.05], relative to the weighted grand mean. No difference was observed for coda or vowel repairs. When Production Type was releveled (omitted level = "Original"), the model showed that correct repeat productions had larger f0 range [*Coef* = 0.05, t = 2.23, p < 0.05]. Trial Number also predicted f0 range: over the course of each block, participants' f0 range increased [*Coef* = 0.15, t = 5.58, p < 0.001]. No other predictors or interactions were significant in the f0 model. Note that while in Fig. 3C. there is a numerical difference for Siri- and human-DS for coda repairs (in the first half of the experiment), no differences for vowel or coda repairs were seen overall — or by Interlocutor — when taking into account the random effects structure (recall also that there were only four coda repair trials for each interlocutor).

3.2.2. Vowel-level results

Mean vowel duration in target words is plotted in Fig. 4A. The models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained

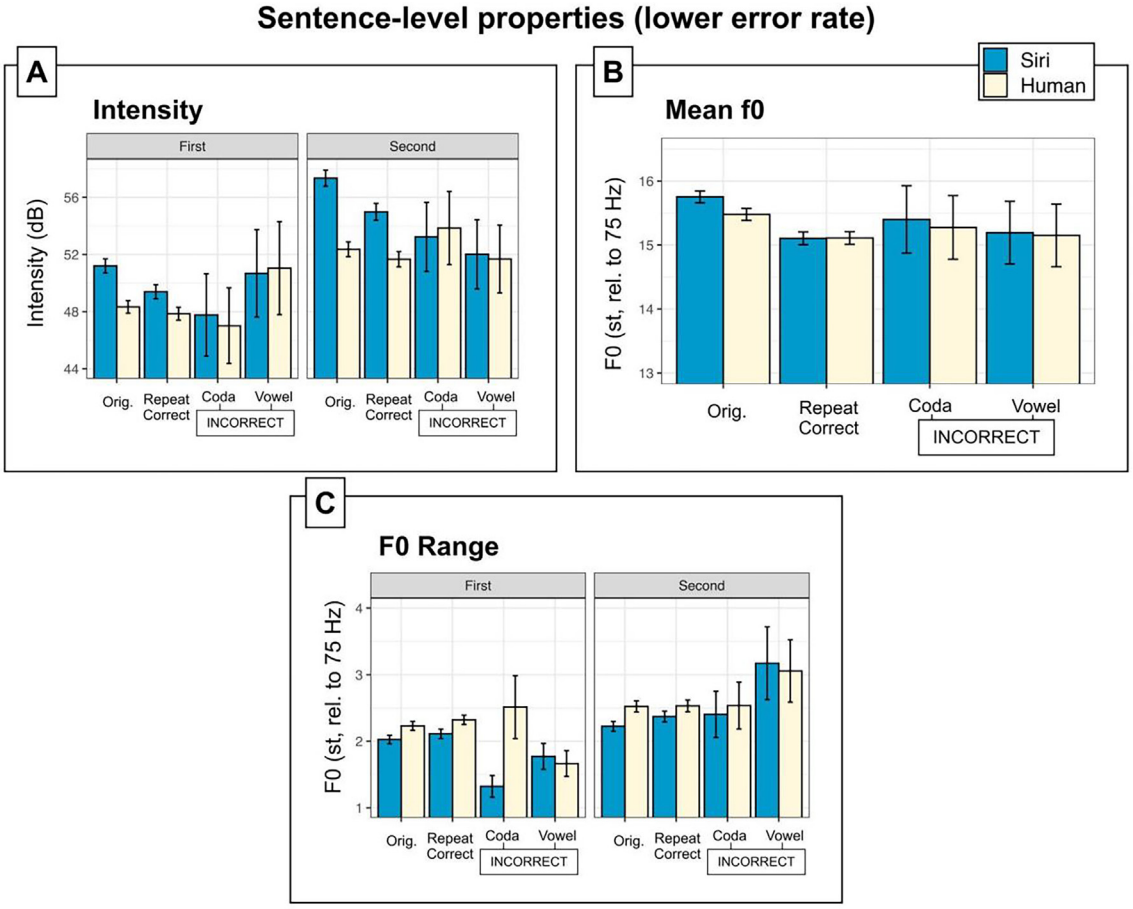

Fig. 3. Experiment 1. Mean sentence-level acoustic values by Portion of the Block (First Half, Second Half), Production Type (Original, Repeat Correct, Incorrect Coda, and Incorrect Vowel), and Interlocutor (Siri = dark blue, Human = light yellow) for (A) intensity, (B) mean f0, and (C) f0 range. Error bars show standard error of the mean. (Note that if no differences by Trial Number were observed, the plot was not faceted by Portion of the Block).

vowel duration model (Interlocutor * ProductionType + Interlocutor * Trial + SpeechRate + (1 + Interlocutor|Participant) + (1|Word)) showed no effect of Interlocutor. However, we did see effects of Production Type, above and beyond any variation explained by speech rate, which was not a significant predictor of target word vowel duration in the model. Specifically, vowel duration was longer in original productions [*Coef* = 0.004, t = 11.75, $p < 0.001$] and even longer in vowel repair productions [*Coef* = 0.01, t = 2.43, $p < 0.05$], relative to the weighted grand mean. There was no difference in vowel duration in coda repairs. When the Production Type factor was releveled (omitting level = "Original"), the model revealed that vowel duration was shorter in correct repeat productions [*Coef* = −0.01 t = −12.55, $p < 0.001$]. Trial Number also predicted vowel duration in target words, such that participants produced slightly longer vowels over the course of each block [*Coef* = 0.001, t = 2.38, $p < 0.05$]. No other predictors or interactions were significant in the vowel duration model.

Acoustic nasality is plotted in Fig. 4B. (only CVN target words). The acoustic nasality models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + Vowel Duration + (1 + Interlocutor|Participant) + (1|Word)) revealed that vowel nasality did not differ by Interlocutor. However, it did vary by Production Type. Releveling the Production Type factor (omitted level = "Original") showed that in correct repeat productions, speakers produced greater coarticulatory nasalization in vowels in CVN words (lower A1–P0 values) [*Coef* = −0.24, t = −2.10, $p < 0.05$]. No other effects or interactions were observed. While it may appear in Fig. 4B. that there is a numerical difference in vowel nasality in correct repeat productions in Siri-DS, this is not significant in the model. Here, it is important to note that the hierarchical random effects structure is not taken into account in the figures — rather, figures show the mean and standard error.

Vowel space expansion (hyperarticulation) values are plotted in Fig. 4C. The vowel space expansion models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + VowelDuration + (1 + Interlocutor|Participant) + (1|Word)) showed no overall effect of Interlocutor. Yet, there was an effect of Production Type. Speakers produced greater vowel space expansion in original productions [*Coef* = 0.01, t = 4.07, $p < 0.001$]. Additionally, an interaction between Interlocutor and Production Type indicated that speakers produce even greater vowel hyperarticulation in incorrect coda repairs

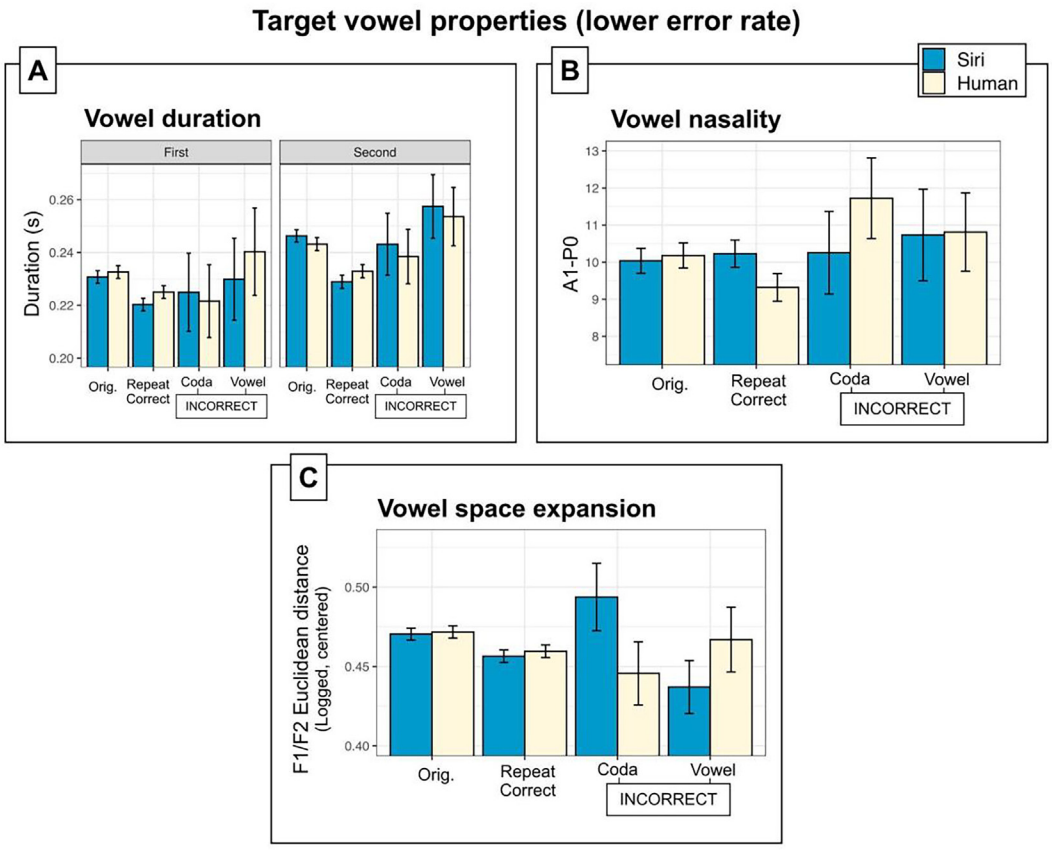

Fig. 4. Experiment 1. Mean acoustic values for (A) target vowel duration, (B) vowel nasality, and (C) vowel space expansion (hyperarticulation) by Portion of the Block (First Half, Second Half), by Production Type (Original, Repeat Correct, Incorrect Coda, and Incorrect Vowel), and Interlocutor (Siri = dark blue, Human = light yellow). Error bars show standard error of the mean. Only CVN tokens are plotted for acoustic vowel nasality. (Note that if no differences by Trial Number were observed, the plot was not faceted by Portion of the Block).

in Siri-DS [*Coef* = 0.02, t = 2.10, $p < 0.05$], seen in Fig. 4C. Note that no difference was observed in Siri-DS or overall in vowel repairs. Vowel Duration was also an effect: with greater expansion for longer vowels [*Coef* = 0.09, t = 2.35, $p < 0.05$]. No other effects or interactions were significant in the model.

3.3. Interim discussion

Experiment 1 compared Siri- and human-DS in pseudo-interactive dialogs. Table 4 provides a descriptive summary of the findings for all the acoustic features by Interlocutor and Production Type, as well as any interactions with Trial. Comparison of Siri- and human-DS revealed both prosodic and segmental differences in the lower error rate experiment.

First, consistent with work showing increased amplitude in computer- and Alexa-DS (Lunsford et al., 2006; Raveh, Steiner, et al., 2019), speakers produce louder utterances in Siri-DS. This increased loudness in Siri-DS suggests that speakers expect Siri to misunderstand them, even as the explicit communicative barriers were identical in rate and nature for the human and Siri interlocutors. Further, speakers increase their

Table 4
Summary of acoustic variation patterns observed in Experiment 1 (low error rate).

	Interlocutor	Production Type
Intensity	• Increased in Siri-DS • Louder in original productions for Siri-DS • Quieter in correct repeats, coda repairs, and vowel repairs in Siri-DS • Increases over time in Siri-DS	• Louder in original productions • Quieter in correct repeats, coda repairs, and vowel repairs
Mean f0	• No difference	• Increases in original productions • Decreases in correct repeats and vowel repairs
F0 range	• Decreases in Siri-DS	• Decreases in originals • Increases in correct repeats
Vowel duration	• No difference	• Longer in original productions and vowel repairs • Shorter in correct repeats
Acoustic nasality (only CVNs)	• No difference	• Increased nasalization in correct repeats
Vowel hyperarticulation	• Increased expansion in coda repairs in Siri-DS	• Increased expansion in original productions • Contracted in correct repeats

intensity in Siri-DS over time. These findings support *listener-intelligibility accounts* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998) in that, for Siri, participants produce louder speech, consistent with greater articulatory effort. These findings counter *technology equivalence accounts* (Nass et al., 1997, 1994; Reeves & Nass, 1996) which argue that people interact similarity with computer and human interlocutors. Indeed, Siri was rated both as sounding less "human-like" and less "communicatively competent" in our independent rating study (described in Section 2.1.1.), consistent with related work showing that people rate TTS voices as having less communicative competence (Cowan et al., 2015).

Second, we observe an overall smaller f0 range in Siri-DS. One possible explanation for this pattern is that f0 range reflects differences in conveying emotional *affect*. Speech directed toward computers, in particular, has been hypothesized to contain less emotional affect, relative to speech towards other types of interlocutors (Burnham et al., 2010), similar to findings of less emotional affect in non-native speaker-DS relative to native speaker-DS (Uther et al., 2007). While flattening f0 has been shown to reduce intelligibility for *human* listeners (Laures & Weismer, 1999), another possibility is that speakers adapt their speech to "sound" more like Siri's (i.e., more robotic). Indeed, Mayo et al. (2012) found a reduced f0 range in "imagined" computer-DS and there is a body of work suggesting that speakers align their speech to improve intelligibility (for a discussion, see Pickering & Garrod, 2006). While there was no difference in f0 range between the human and Siri voices used in the studies, it is possible that participants perceived the voice as sounding more "monotone" (i.e., having smaller f0 variation), consistent with what is reported in related work (Siegert & Krüger, 2021). At the same time, another possibility is that participants were *increasing* their f0 range more in human-DS. This would be consistent with increased vocal effort towards humans in response to greater apparent communicative barriers. Yet, the effect of increasing f0 variation on enhancing intelligibility has produced equivocal results in other work. For example, Miller and colleagues (2010) found that exaggerating f0 variation actually leads to reduced intelligibility in speech-in-noise.

Furthermore, we saw other differences in Siri-DS that were mediated by local intelligibility strategies, lending support for a hybrid *targeted adaptation* and *listener-intelligibility* account. For example, speakers produce greater vowel space expansion when repairing coda errors in Siri-DS. This increased vowel hyperarticulation aligns with related work showing vowel space expansion and targeted segmental hyperarticulation (e.g., in consonants) in computer-DS (Burnham et al., 2010; Stent et al., 2008). Furthermore, speakers are louder in their original productions in Siri-DS, consistent with first-mention hyperarticulation (Fowler & Housum, 1987). At the same time, speakers are quieter in Siri-DS in correct repeat and error repairs. Together, the targeted "original" adjustments appear to be a strategy to improve intelligibility in first mention for an interlocutor the speaker presumes might misunderstand them.

At the same time, we still see similar adjustments in Siri- and human-DS in some communicative contexts. For example, in response to vowel errors, speakers systematically produce longer vowels. Additionally, during the original sentence (i.e., "first mention"), speakers produce "clear" speech adjustments for both interlocutors: increased intensity, a higher mean f0, longer vowel duration, and more vowel space expansion. Yet, when communication goes smoothly (correct repeat productions) speakers produce more hypospeech: decreased intensity, lower mean f0, shorter vowel duration, less vowel space expansion, and greater coarticulatory nasalization. These patterns of coarticulatory nasalization align with prior work, where speakers increase nasal coarticulation in casual speech but decrease coarticulation in speech addressed to (imagined) hard-of-hearing addressees (Scarborough & Zellou, 2013), suggesting that coarticulatory vowel nasalization is part of targeted intelligibility adjustments.

While we find that speakers adjust vowel properties (duration, coarticulatory nasalization) across original and correct repeat productions, we critically observe only one difference in segmental features in Siri- and human-DS. This raises a question: why do Siri- and human-DS vary primarily in terms of *prosodic* features in this study? For instance, prior work has found evidence for segmental lengthening in computer-DS (Burnham et al., 2010) and targeted nasalization in words adults or infants might misunderstand (Scarborough & Zellou, 2013; Zellou & Scarborough, 2015). One potential explanation for the observed overlap in duration and coarticulation in Siri- and human-DS is that it occurred due to the rate of interlocutor errors. In Experiment 1, comprehension mistakes occurred in just eight out of 56 trials for both the human and Siri interlocutors. It is possible that there were too few error repair trials to detect other differences in targeted acoustic–phonetic adjustments across the response types (weighted sum coding takes into account the number of observations in each category). Another possibility is that there were too few errors to elicit additional interlocutor-specific adjustments by the participants. Prior work examining computer-DS, without a human comparison, has observed increased segmental hyperarticulation when the rate of errors is increased (Oviatt, MacEachern, et al., 1998; Stent et al., 2008). Therefore, in Experiment 2, an independent set of speakers participated in a study with the same design as that in Experiment 1, but with a higher rate of vowel and coda errors.

4. Experiment 2. Higher error rate

Experiment 2 tests whether a higher error rate influences speech adjustment patterns in Siri- and Human-DS. We use the same design as in Experiment 1, but with errors occurring in 60.7% of trials (compared to 14.3% errors in Experiment 1).

4.1. Methods

4.1.1. Participants

Participants consisted of 31 adults (mean age 20.5 ± 2.3 years, age range 18–30 years; 23 females, 8 males). They were recruited from the UC Davis Psychology Pool and received course credit for participation. All participants were either monolingual English speakers or bilingual (e.g., Spanish-English), with English as their dominant language. All 31 participants reported experience using voice-AI systems. None of the participants reported hearing impairment. All participants gave informed consent to participate, in pursuance with the UC Davis Institutional Review Board.

4.1.2. Stimuli

Described in Section 2.2.1.

4.1.3. Procedure (higher error rate study)

The general procedure is described in Section 2.2.2. The overall error rate for Experiment 2 was 60.7% (recognition errors in 34/56 trials; 22 correct). In the 12 vowel space trials, Siri/human produced staged misrecognition errors on 6 trials (randomly selected). In the experimental trials for each interlocutor block (22 CVN, 22 CVC), Siri/human produced a coda error in 14 trials and a vowel error in 14 trials. In the remaining trials, the interlocutor heard "correctly", showing the correct target word to the participant.

4.1.4. Analysis

General methods for acoustic and statistical analysis are described in Section 2.3.

4.2. Results

Model outputs are provided in the Supplementary Data 2 (Tables 10–21) (with the retained model structure indicated for each). Mean values for Interlocutor and Production Type at the sentence-level are plotted in Fig. 5 and vowel-level measurements are displayed in Fig. 6.

4.2.1. Sentence-level results

Mean utterance intensity is plotted in Fig. 5A. The intensity models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant) + (1|Word)) showed effects of Production Type, as well as differences in Siri-DS that emerged over time. First, as seen in Fig. 5A, participants spoke louder in their original productions [*Coef* = 0.33, t = 13.29, $p < 0.001$] and quieter in coda repair [*Coef* = −0.23, t = −4.20, $p < 0.001$] and vowel repair productions [*Coef* = −0.16, t = −2.38, $p < 0.05$]. The releveled model (omitted level = "Original") revealed that speakers are also quieter in correct repeat productions [*Coef* = −0.54, t = −10.60, $p < 0.001$]. There was also an interaction between Interlocutor and Trial Number. As seen in Fig. 5A, participants are louder in Siri-DS later in the block [*Coef* = 0.15, t = 5.80, $p < 0.001$]. No other effects or interactions were observed.

Mean f0 is plotted in Fig. 5B. The mean f0 models including by-Participant random slopes for Production Type resulted in singularity errors. The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant) + (1|Word)) showed effects of Production Type, and differences in Siri-DS that emerged over time. First, as seen in

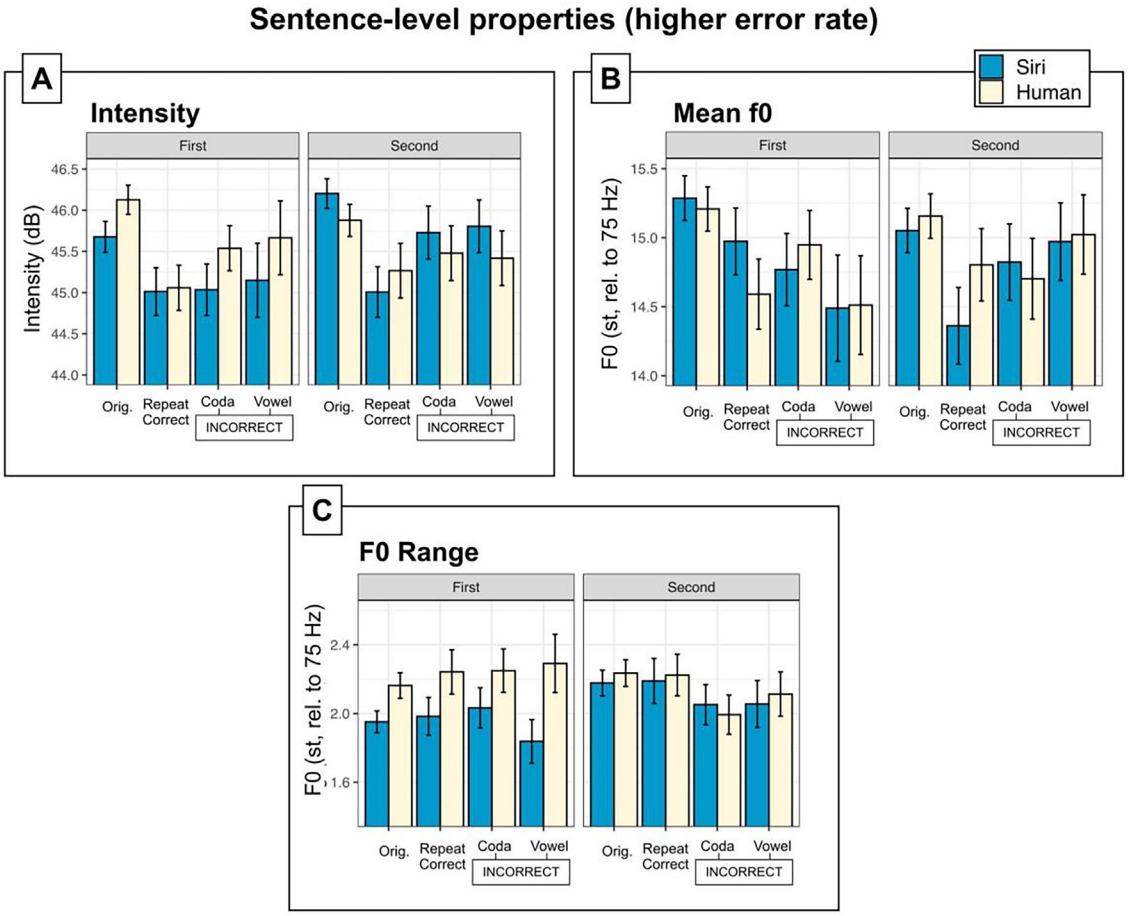

Fig. 5. Experiment 2. Mean sentence-level acoustic values by Portion of the Block (First Half, Second Half), Production Type (Original, Repeat Correct, Incorrect Coda, and Incorrect Vowel), and Interlocutor (Siri = dark blue, Human = light yellow) for (A) intensity, (B) mean f0, and (C) f0 range. Error bars show standard error of the mean. (Note that if no differences by Trial Number were observed, the plot was not faceted by Portion of the Block).

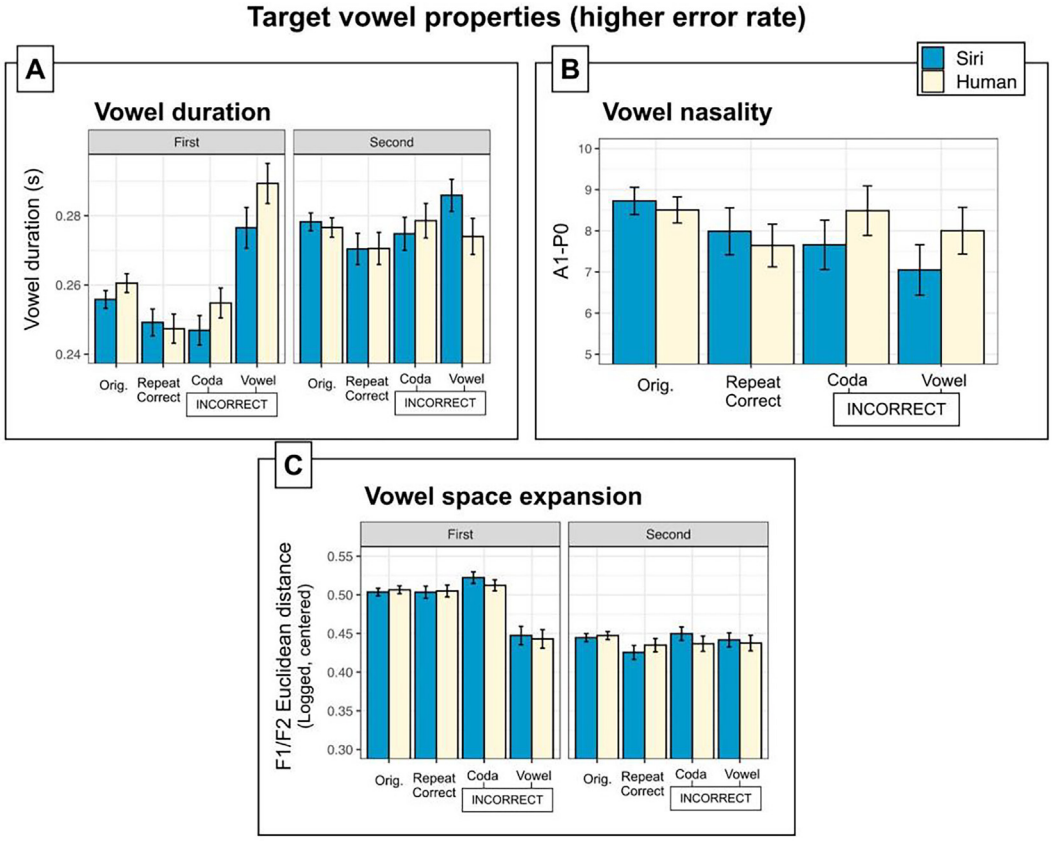

Fig. 6. Experiment 2. Mean acoustic values for (A) target vowel duration, (B) vowel nasality, and (C) vowel space expansion (hyperarticulation) by Portion of the Block (First Half, Second Half), by Production Type (Original, Repeat Correct, Incorrect Coda, and Incorrect Vowel), and Interlocutor (Siri = dark blue, Human = light yellow). Error bars show standard error of the mean. Only CVN tokens are plotted for acoustic vowel nasality. (Note that if no differences by Trial Number were observed, the plot was not faceted by Portion of the Block).

Fig. 5B, original productions have a higher mean f0 [*Coef* = 0.20, *t* = 17.46, *p* < 0.001], while both coda repairs [*Coef* = −0.15, *t* = −6.10, *p* < 0.001] and vowel repairs [*Coef* = −0.13, *t* = −4.31, *p* < 0.001] have a lower mean f0, compared to the weighted grand mean. The releveled model (omitted level = "Original") showed that speakers also produce a lower mean f0 in correct repeat productions as well [*Coef* = −0.29, *t* = −12.42, *p* < 0.001]. Mean f0 also varied by Trial Number wherein participants' mean f0 decreased over time [*Coef* = −0.04, *t* = −3.08, *p* < 0.001]. Additionally, there was a significant interaction between Interlocutor and Trial Number. Participants' mean f0 decreased in Siri-DS over the course of the block [*Coef* = −0.05, *t* = −4.29, *p* < 0.001]. No other effects or interactions were observed.

Values for f0 range are plotted in Fig. 5C. The f0 range models including by-Participant random slopes for Production Type resulted in singularity errors. The retained model (Interlocutor * ProductionType + Interlocutor * Trial + (1 + Interlocutor|Participant) + (1|Word)) showed no difference by Production Type, but an interaction between Interlocutor and Trial Number: f0 range increases in Siri-DS over the course of the block [*Coef* = 0.05, *t* = 2.06, *p* < 0.05], as seen in Fig. 5C. No other effects or interactions were observed.

4.2.2. Vowel-level results

The vowel duration results are plotted in Fig. 6A. The models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence errors (Interlocutor + ProductionType|Participant). The retained model (Interlocutor * ProductionType + Interlocutor * Trial + SpeechRate + (1 + Interlocutor|Participant) + (1|Word)) showed no effects of Interlocutor, but differences based on Production Type. Participants produce longer vowel duration in original [*Coef* = 0.0008, *t* = 2.0, *p* < 0.05] and incorrect vowel [*Coef* = 0.005, *t* = 4.05, *p* < 0.001] productions. The releveled model (omitted level = "Original") showed that speakers also produce shorter vowels in correct repeat trials [*Coef* = −0.01, *t* = −7.24, *p* < 0.001], relative to the weighted grand mean. There was an effect of Trial Number indicating that over the course of the block, speakers produce longer vowel duration [*Coef* = 0.004, *t* = 6.49, *p* < 0.001], as seen in Fig. 6A. There was no difference based on Speech Rate, indicating that these vowel duration adjustments are local adaptations, not a result of changing utterance speed. No other effects or interactions were significant in the model.

Acoustic vowel nasality for CVN tokens is plotted in Fig. 6B. The nasality models including by-Participant random slopes for Production Type resulted in singularity errors. In the retained model (Interlocutor * ProductionType + Interlocutor * Trial + Vowel Duration + (1 + Interlocutor|Participant) + (1|Word)), we find no effects or interactions involving Interlocutor. The releveled model (omitted level = "Original") showed that correct repetitions contained vowels with lower A1–P0 (indicating greater coarticulatory vowel nasality) [*Coef* = −0.58, *t* = −2.24, *p* < 0.05], compared to the weighted grand mean. There was also a significant effect of Vowel Duration on acous-

tic nasality wherein increasing vowel duration was associated with increased coarticulatory vowel nasalization (lower A1–P0 values) [*Coef* = −19.86, *t* = −5.59, *p* < 0.001]. No other effects or interactions were significant in the model.

Vowel space expansion is plotted in Fig. 6C. The vowel space expansion models including by-Participant random slopes for Production Type resulted in singularity (Interlocutor * ProductionType|Participant) or convergence (Interlocutor + ProductionType|Participant) errors. The retained model (Interlocutor * ProductionType + Interlocutor * Trial + VowelDuration + (1 + Interlocutor|Participant) + (1|Word)) showed no effects or interactions involving Interlocutor. There were effects of Production Type on vowel space expansion, where speakers produce greater vowel hyperarticulation in original productions [*Coef* = 0.004, *t* = 3.88, *p* < 0.001], but less hyperarticulation in correct repeat productions [*Coef* = −0.01, *t* = −4.11, *p* < 0.001]. There was an effect of Vowel Duration: as duration increased, vowel hyperarticulation increased [*Coef* = 0.17, *t* = 6.00, *p* < 0.001]. Additionally, there was an effect of Trial Number. As seen in Fig. 6C, vowel space expansion decreased over time [*Coef* = −0.003, *t* = −2.12, *p* < 0.05]. No other effects or interactions were observed. While there is a numerical difference in expansion in Siri-DS for coda repairs in the "first half" in Fig. 6C, this was not significant when accounting for both participant and word variability in the mixed effects models.

4.3. Interim discussion

Experiment 2 compared Siri- and human-DS in a design that employed a higher error rate than Experiment 1, with staged misrecognitions in 60.7% of trials for both interlocutors. Table 5 provides a descriptive summary of the findings for all the acoustic features by Interlocutor and Production Type.

In the higher error rate, we observed prosodic differences across Siri- and human-DS, but that only emerged over time. Over the course of the block, Siri-DS increases in intensity and f0 range, but decreases in mean f0. These adjustments align with prior studies on the acoustic features in speech directed toward modern voice-AI technology. For example, Alexa-DS is louder in the VACC corpus, relative to human-DS (Raveh, Steiner, et al., 2019; Siegert & Krüger, 2021). While Mayo et al. (2012) found reduced f0 range in computer-DS, Cohn & Zellou (2021) recently found greater f0 variation in Alexa-DS, as well as greater f0 variation in response to errors when they occurred in 50% of trials. This suggests that increasing f0 range might be part of a dynamic intelligibility strategy here where the error rate was 60.7%. Together, the prosodic differences in Siri- and human-DS support *listener intelligibility accounts* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998), in that speakers adapt their speech in distinct ways for Siri and human interlocutors, even when the communicative scenario —and rate and type of errors — is controlled across registers.

We additionally observed differences in mean f0 over time: speakers produce *lower* mean f0 in Siri-DS over time. This contrasts with work showing increasing mean f0 in more effortful speech (e.g., Lombard speech) and in Alexa-DS (Cohn & Zellou, 2021). One possibility is that the lower f0 in Siri-DS is a vocal alignment effect, where participants' productions became more similar to the interlocutors'. This parallels Raveh et al.'s (2019) interpretation of a difference in mean f0 for Alexa-DS, where they found higher mean f0 in Alexa-DS, as the Alexa had a female voice while the human was male. Indeed, recent work has examined and found that vocal alignment differs toward voice-AI and human interlocutors (Cohn et al., 2019; Zellou, Cohn, & Ferenc Segedin, 2021). Here, acoustic analysis of the interlocutors confirmed that the human speaker had a higher mean f0 than the Siri voice, lending support for this interpretation. Since the current study was not designed to examine phonetic alignment, this is a direction for future work.

While we raised the possibility that the error rate was too low in Experiment 1 to detect targeted adjustments in Siri-DS, we did not observe different error adaptations in Siri-DS at the higher error rate. For example, in response to a vowel error, speakers produced systematically longer target vowels in both Siri- and human-DS. The lack of vowel lengthening following coda repairs suggests that vowel duration was part of a targeted intelligibility strategy specifically for these types of errors, aligning with *targeted adaptation accounts* (e.g., Buz et al., 2016).

Additionally, we observe similar speech-adjustment tendencies across original and correct repeat productions in Siri- and human-DS. In first-mentions, speakers' productions are louder, have a higher mean f0, longer vowel duration, less coarticulatory nasalization, and greater vowel space expansion. These adjustments are consistent with clear speech adjustments (Lee & Baese-Berk, 2020; Smiljanić & Bradlow, 2005). Here, speakers appear to produce more effortful and intelligible speech in first mentions of words, aligning with prior work (Fowler & Housum, 1987). When the interlocutor (Siri or human) heard "correctly", speakers' productions are quieter,

Table 5
Summary of acoustic features observed in Experiment 2 (higher error rate).

	Interlocutor	Production Type
Intensity	• Increases over time in Siri-DS	• Increases in original productions • Decreases in correct repeats, coda repairs, and vowel repairs
Mean f0	• Decreases over time in Siri-DS	• Increases in original productions • Decreases in correct repeats, coda repairs, and vowel repairs
F0 range	• Increases over time in Siri-DS	• No difference
Vowel duration	• No difference	• Longer in original productions and vowel repairs • Shorter in correct repeats
Acoustic nasality (only CVNs)	• No difference	• Less nasalization in original productions • More nasalization in correct repeats
Vowel hyperarticulation	• No difference	• More expansion in originals • Contraction in correct repeats

have lower mean f0, shorter target vowel duration, and greater coarticulatory nasalization, and less vowel space expansion. These adjustments are consistent with the properties of less effortful speech (hypospeech, Lindblom, 1990).

5. General discussion

The current study compares participants' speech style registers for a human and voice-AI interlocutor (here, Siri). We conducted two experiments, systematically varying the rate of errors made by the interlocutors: a relatively lower error rate (14.3% error rate: Experiment 1) and a higher error rate (60.7% error rate: Experiment 2). Manipulating the rate of errors in computer-directed speech studies is one way to understand how people dynamically adjust to real-time evidence about the communicative competence of the system (Oviatt, MacEachern, et al., 1998; Stent et al., 2008). Here, we investigate how a priori expectations of a TTS voice interact with real-time evidence of the interlocutor's understanding to influence speakers' patterns of prosodic and segmental variation.

5.1. Evidence for a Siri-DS register

Across both studies, we observed global differences across Siri- and human-DS. Siri-DS tends to be louder, have a lower mean f0, and a smaller f0 range (but one that increases over time).

First, we find across both studies that productions are louder in Siri-DS. Increased intensity reflects more effortful speech. For instance, it is a commonly observed feature in speech produced in adverse listening situations (e.g., in background noise; Brumm & Zollinger, 2011), as well as in clear speech (Ferguson et al., 2010). That people speak more loudly toward Siri is consistent with observations of louder speech (in general) in computer-DS (Lunsford et al., 2006) and in recent findings for speech directed to voice-AI, versus a human confederate (Raveh, Steiner, et al., 2019; Siegert & Krüger, 2021). In the current study, we used a head-mounted microphone and amplitude normalized recordings of the human and Siri productions, minimizing differences across interlocutors. This suggests that the intensity differences indeed reflect interlocutor-specific adaptations, rather than artifacts of spatial separation from a microphone and/or interlocutor (Pelegrín-García et al., 2011).

Second, we observe differences in f0 range. In the lower error rate study, f0 range was smaller in Siri-DS overall. As mentioned in the Interim Discussion (Section 3.3), a possible explanation is that this reflects *less* affective-emotional responses in Siri-DS. Happy speech, for example, contains wider f0 variation (Abadjieva et al., 1993) and related work has shown that positive affect is perceived in infant- and pet-DS, but not in non-native speaker-DS (Burnham et al., 2002; Uther et al., 2007), suggesting less affect might be present in computer- and voice-AI-DS. In line with an emotional-affect interpretation, "imagined" computer-DS contains a smaller f0 range (Mayo et al., 2012). In contrast, a recent study observed *greater* f0 variation in Alexa-DS (Cohn & Zellou, 2021) — yet, in that study, the Alexa TTS produced both emotionally expressive and neutral speech, which might be a factor in adaptation. Future work investigating external listeners' affecting ratings of speech directed toward voice-AI/human interlocutors, as well as overall intelligibility of that speech, can tease apart this possible emotional contribution in the f0 range findings. Another interpretation we raised for this f0 range pattern is that participants might be aligning to the smaller f0 variation of a more monotone-sounding speaker. But critically, if the reduced f0 range indeed reflects alignment, it is based on the listeners' expectation of a more monotonous voice-AI talker, rather than true acoustic realization. Our analysis of the interlocutors' speech confirmed no significant difference in f0 range across the Siri and human talkers. The possibility that speakers are "converging-to-expectation" draws support from recent work showing convergence in dialectal features *assumed* to be heard, but that were never present in the stimuli (e.g., Southern American English in Wade, 2020). These possibilities open many new avenues of study. For instance, future work systematically varying f0 variation in the stimuli — for both TTS and naturally recorded human voices — can tease apart the possible contribution of "converging-to-expectation" in adaptation strategies.

The difference in f0 range for Siri-DS was also dynamic, increasing over time to approach the levels in human-DS. In the higher error rate study, the increasing f0 range over time in the Siri-DS block suggests that these changes might reflect adaptation strategies to improve intelligibility in response to the accumulation of many interlocutor errors (here, 60.7% error rate), as well as for an addressee they assume to be less communicatively competent. But, as previously mentioned, introducing greater f0 variation can result in *lower* intelligibility for human listeners (Miller et al., 2010), suggesting that these adjustments are based on assumptions rather than experience-based adaptation strategies. The extent to which increasing f0 range in Siri-DS might shape ASR accuracy remains another open avenue for future work.

Third, speakers decrease their mean f0 over time in Siri-DS. Critically, this was only observed in the higher error rate study. As mentioned earlier, one possibility is that the lower f0 in Siri-DS is a vocal alignment effect, where participants' productions become more similar to the interlocutors' (e.g., Raveh et al., 2019). Yet, we do not observe the same effect in the lower error rate. While the increased error rate increased communicative barriers, and might have possibly driven increased alignment as proposed by functional accounts of vocal alignment (e.g., Pickering & Garrod, 2006), an f0 alignment account of the present findings remains speculative. Future studies which 1) control f0 parameters for human/TTS voices, and 2) parametrically manipulate them can reveal how alignment and adaptation strategies might interact for f0 properties in human- and voice-AI-DS.

Together, the global prosodic differences in Siri- and human-DS counter predictions made by *technology equivalence accounts* (Nass et al., 1997, 1994; Reeves & Nass, 1996). Despite Siri having "human-like" qualities (e.g., name, gender, persona), people talk to voice-AI and humans in distinct ways. These findings provide support for *listener-intelligibility accounts* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998) which argue that speakers produce different speech adaptations for different interlocutors based on (apparent) communicative barriers. As mentioned, in the independent ratings experiment of the voices used in the present study, listeners perceived the Siri interlocutor both as less "human-like", as well as less likely to under-

stand their speech, compared to the human interlocutor — even without any interaction. Thus, the difference in the human versus Siri interlocutors' presumed communicative competence could explain the speech adjustments made in Siri-DS.

5.2. Different error correction in Siri-DS

While our first research question asks whether there is a distinct voice-AI register, our second question asks whether there are different error correction strategies in Siri- and human-DS. Here, we observed one targeted difference in the lower error rate study: speakers show greater vowel hyperarticulation in Siri-DS following a coda misrecognition. While vowel hyperarticulation is a commonly observed — and *stable* — feature of some registers (e.g., infant-DS in Kuhl et al., 1997; computer-DS in Burnham et al., 2010), we see that it is dynamically targeted by the local communicative context. Overall, observing different error correction strategies in Siri- and human-DS — even if only at the lower error rate — supports a hybrid account, combining *listener-intelligibility* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998) and *targeted adaptation accounts* (Buz et al., 2016; Lindblom, 1990).

That we did *not* see parallel adjustments in vowel expansion in the higher error rate study (60.7%) is consistent with other studies that had relatively higher error rates. For example, Cohn & Zellou (2021) found that speakers produce similar vowel adjustments in both Alexa- and human-DS when the interlocutors made staged misrecognitions in 50% of trials. Similarly, Burnham et al. (2010) found no differences in error correction for human- and computer avatar-DS when the rate of errors was 33%. Together, these findings suggest that at a higher error rate, vowel hyperarticulation strategies are more similar across computer-/voice-AI- and human-DS registers.

While we observed the difference in vowel space expansion, we did not observe differences in vowel duration or nasal coarticulation in Siri- and human-DS in response to errors. In particular, we predicted that vowel errors might trigger more vowel duration and hyperarticulation, while coda nasality errors might trigger more nasal coarticulation, and that these would be differentially tuned in Siri- and human-DS. As mentioned, we only observed increased vowel space expansion in Siri-DS for *coda* repairs. Here, one explanation is that speakers had other strategies for repairing vowels, namely increasing vowel duration. Indeed, vowel repairs had longer vowel duration in both experiments, though there were no differences across Siri- and human-DS. This contrasts from other work finding differential targeting of nasal coarticulation in adult- and infant-DS based on the lexical properties of words (Scarborough, 2013; Zellou & Scarborough, 2015). One possibility is that speakers do not target nasal coarticulation differently for computer/voice-AI interlocutors. Future work examining other acoustic features (e.g., voice onset time, /t/ release) and lexical properties can further shed light on segmental error correction strategies in voice-AI- and human-DS, if present.

5.3. Differences in "original" and "correct repeat" productions in Siri-DS

In the present study, each trial consisted of an original, "first mention" production of the target sentence, followed by a repetition in response to either the correctly identified target word, or an incorrect vowel/coda option. In addition to the difference observed in response to errors (discussed in Section 5.2.), speakers also produced distinct patterns in "original" and "correct repeat" productions in Siri-DS.

First, speakers produced louder original productions to Siri in the lower error rate. This adjustment could be a strategy to improve intelligibility in first mention for an interlocutor the speaker presumes might misunderstand them. Second, speakers show lower intensity in correct repeat productions in Siri-DS reflecting *less* effortful speech. Indeed, this decrease in intensity parallels more general adaptation strategies in correct repetitions to both interlocutors. Here, correct repeat productions showed patterns of "hypospeech" consistent with the idea that when communication is successful, speakers expend less effort (e.g., Lindblom, 1990). In the present study, that reduced effort is realized as quieter utterances, shorter vowel durations, lower f0, less vowel hyperarticulation, and greater nasal coarticulation as rapid connected speech can result in reduction and greater articulatory overlap (cf. Farnetani & Recasens, 2010).

Together, these differences in "original" and "correct repeat" productions for Siri-DS support a hybrid *targeted adaptation* (Buz et al., 2016; Lindblom, 1990) and *listener-intelligibility account* (Branigan et al., 2011; Clark & Murphy, 1982; Oviatt, MacEachern, et al., 1998). These differences can be interpreted along with prior findings that clear speech involves the targeted restructuring of phonetic patterns (Bradlow, 2002; Scarborough & Zellou, 2013). Specifically, while participants produced louder first mention utterances in Siri-DS, this did not occur with concomitant acoustic–phonetic differences (e.g. in F1/F2 hyperarticulation). Compared with other work where increasing vocal intensity observed in elicited clear speech co-occurs with vowel space expansion and longer durations (Lam et al., 2012), the current finding suggests that the similarities observed across Siri- and human-DS in these first mention utterances reflect active maintenance of acoustic–phonetic features in speech directed toward both interlocutor types. Thus, while loudness is a feature speakers increase in first mention productions to Siri, there appears to be active compensation for articulatory changes that might co-occur with increasing intensity to maintain acoustic–phonetic output of other features in this condition (e.g., F1/F2, duration, etc.).

5.4. Implications for voice technology

In addition to serving as a direct comparison between a human and voice-AI interlocutor, this study sheds light on some of the cognitive factors at play in human–computer interaction, which have real-world applications for voice technology. The adjustments we find in Siri-DS (e.g., increased intensity) reflect, in general, more effortful speech. Yet, many ASR systems have been trained on naturalistic, casual speech (e.g., Wade et al., 1992). Hyperarticulation in computer-DS can lead to a "cycle of misunderstanding" (Oviatt, MacEachern, et al., 1998; Stent et al., 2008): when speakers experience an ASR error, they produce more hyperarticulated speech, which in turn can lead to additional ASR errors. There is evidence in the current study for more vowel hyperarticulation following errors, suggesting that an "ASR error cycle" could be at play with modern voice-

AI. Yet, at the higher error rate, adjustments were largely parallel in human- and Siri-DS. Since we find only one difference in error correction strategies (see Section 5.2), this suggests that including training data from human-human interaction in how people correct errors might improve ASR models.

Furthermore, the current study provided participants with information indicating the source of an error in comprehension, using visual feedback of the interlocutor's "guess" of the word. While more common in screen-based voice-AI systems (e.g., Echo Show, Siri interface on the iPhone), confirming correct recognition or confusion about a target word could be elicited vocally in speech interactions. For example, Cohn & Zellou (2021) found targeted vowel formant adjustments (specifically *F*2) when the Alexa or human interlocutor made a vowel backing mistake ("I think I missed that. I think I heard boat or bet."). In the current study, we find that vowel duration is specifically targeted in vowel repair trials — after the participant has seen visual feedback that the interlocutor made a vowel mistake (but heard all other segments correctly). The extent to which speakers' adaptation strategies vary according to their interlocutors' feedback — for example, signaling more general intelligibility difficulties (e.g., "I missed that.") versus more targeted approaches (e.g., "I might have misunderstood. I heard bat.") — are avenues for future research.

5.5. Limitations and directions for future work

One limitation of the present study is that the participants engaged with the interlocutors via pre-recorded, disembodied voices. Prior work has shown differences in articulations for real and imagined speakers (e.g., Scarborough & Zellou, 2013). In some ways, the paradigm in the current study reflects a context that may be more ecologically valid for communicating with voice-AI (i.e., via a computer) than for speaking with a human. Future work using interactions with real humans and devices could address this limitation with more authentic interactions. Additionally, there is a growing body of work examining the role of embodiment in interactions with technology (Appel et al., 2012; Lee et al., 2006). Physical form, varying in terms of human-likeness, has been shown to more gradiently shape vocal alignment patterns (Cohn, Jonell, et al., 2020), suggesting that embodiment might also play a role in speech adaptation strategies as well. While prior work has compared human- and computer avatar-DS (Burnham et al., 2010) or between human- and computer-DS without a visual component (e.g., Oviatt et al., 1998), future work manipulating the presence of physical form (e.g., using virtual reality and/or physical embodiment differences, comparing to voice-only conditions) can shed further light on the source of register differences.

Furthermore, while the present study used only one human and device voice, future work should compare register adaptation for multiple types of interlocutors, such as voice-AI-DS with non-native speaker-DS and child-DS (cf. Cooke et al., 2014). At the same time, our findings suggest that learned associations with a given TTS voice might also shape the interaction. In the current study, the Siri voice is rated both as less human-like *and* less communicatively competent. One explanation is that prior experience — or cultural knowledge— of a voice-AI system having ASR difficulty might lead to more distinct register adaptation strategies from human-DS. Using novel TTS voices (to the participants) might be an avenue to circumvent these expectations and subsequent adaptations, but more research is needed in this area.

In addition to varying the voices, comparing interactions across voice-AI systems (e.g., Apple's Siri, Google Assistant, etc.) can speak to properties shared across all voice-AI-DS. While the nature of these interactions across voice-AI assistants is similar, with users asking for information and giving commands, there may be fundamental differences in the ASR, natural language understanding (NLU), natural language generation (NLG), and TTS voices which might shape how the participant views the system's competence and subsequently the way they talk to the system. Indeed, prior work has shown variation across ASR systems underlying common voice-AI assistants (Koenecke et al., 2020; Palanica et al., 2019).

Future work could also explore other types of situations which present obstacles to communication (e.g., speaking in noise or in multi-talker babble, cf., Hazan & Baker, 2011) to explore how that might influence talkers' interlocutor-based clear speech strategies across different communicatively challenging conditions. Furthermore, this study examined responses in a higher and lower error rate (errors on 60.7% and 14.3% trials, respectively), relative to each other. While similar to Stent et al. (2008) who had 50% and 8.3% errors for higher and lower error rates, it is possible that a larger magnitude of difference might lead to other changes over time. For instance, this could be the case if the conversation is more successful, or if the interlocutor gets it "wrong" in an even greater proportion of trials.

Furthermore, the extent to which individual variation by humans' social and cognitive characteristics shapes speech adaptation to voice-AI is a promising area for future research. Prior work has shown variation in how people perceive and personify technological agents, such as robots (Hinz et al., 2019) and voice-AI (Cohn, Raveh, et al., 2020; Etzrodt & Engesser, 2021). Recently, some work has revealed differences in speech alignment toward voice-AI by speaker age (e.g., older vs. college-age adults in Zellou, Cohn, & Ferenc Segedin, 2021) and cognitive processing style (e.g., autistic-like traits in Snyder et al., 2019), suggesting these differences could shape voice-AI speech adaptation as well.

Finally, this study used acoustic–phonetic measures to characterize differences in Siri- and human-DS. In some cases, the differences between Siri- and human-DS were small (e.g., an increase in intensity of 0.64 dB over time in Siri-DS). The actual impact of these adjustments on intelligibility is an open question. A future direction for this line of work is to investigate how speech directed toward human and voice-AI interlocutors influences recognition by real listeners and ASR systems, respectively.

6. Conclusion

Overall, this study serves as a comprehensive and controlled examination of the acoustic–phonetic properties of Siri- and human-DS in a laboratory setting. Across two error rate experiments, we find differences in the registers in their phonetic changes over time, suggesting that speakers shape their productions based on both "presumed" and "actual" communicative barriers faced by different types of interlocutors.

Examining the dynamic register responses can serve as avenues for future investigations of additional interlocutor types and communicative contexts.

CRediT authorship contribution statement

Michelle Cohn: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Bruno Ferenc Segedin:** Conceptualization, Methodology, Data curation, Writing – original draft. **Georgia Zellou:** Conceptualization, Methodology, Writing – original draft, Supervision.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: 2018 Amazon Faculty Research Award (ARA) to Georgia Zellou, NSF SBE (#1911855) Postdoctoral fellowship to Michelle Cohn.

Acknowledgments

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC. This work was partially supported by an Amazon Faculty Research Award to GZ. Thank you to our undergraduate researchers who assisted with data collection for the project: Jazmina Chavez, Melina Sarian, Divine Otico, Patricia Sandoval, and Eleanor Lacaze.

Appendix A. Productions generated by the human and Siri voices.

Introduction
- Hi! [I'm Siri. I'm a digital assistant on Apple products. | Hi! I'm Melissa. I work here in the Phonetics Lab.] I will show sentences on the screen. Please read them aloud to me. They will always be in black and in quotations. I will give you feedback as we go along. Now, I will show you an example.

Voice-over instructions (4)
- Here's how this is going to work. First you read this sentence aloud. Then I will show you in red what I think it is.
- If my response is right, press "Yes" on the button box. If it is wrong, press "No".
- After that, another screen will come up. Read the whole sentence aloud to confirm or correct what I wrote.
- Let's begin.

Immediate feedback (4)
- Did you say this word?
- Is this correct?
- Is this right?
- Is this the word?

Closing (5)
- Good.
- Great.
- Got it.
- Okay, got it.
- I think I get it now.

Appendix B. Target word lists.

Target CiC ∼ *Incorrect Coda*	**Target CaC** ∼ *Incorrect Coda*
1. **deed** ∼ *deet*	2. **Dodd** ∼ *dot*
3. **teed** ∼ teet	4. **Todd** ∼ *tot*
5. **bead** ∼ *beat*	6. **bod** ∼ *bot*
7. **beep** ∼ *beam*	8. **bop** ∼ *bomb*
9. **beet** ∼ *bead*	10. **bot** ∼ *bod*
11. **peep** ∼ *peeb*	12. **pop** ∼ *pob*
Target CVN ∼ *Incorrect Coda* ∼ *Incorrect Vowel*	**Target CVC** ∼ *Incorrect Coda* ∼ *Incorrect Vowel*
1. **bone** ∼ *bode* ∼ *bane*	1. **bode** ∼ *bone* ∼ *bade*
2. **calm** ∼ *cob* ∼ *cam*	2. **cob** ∼ *calm* ∼ *cab*
3. **con** ∼ *cod* ∼ *can*	3. **cod** ∼ *con* ∼ *cad*
4. **come** ∼ *cub* ∼ *chem*	4. **cub** ∼ *come* ∼ *cab*
5. **dawn** ∼ *Dodd* ∼ *dan*	5. **Dodd** ∼ *dawn* ∼ *dad*
6. **dumb** ∼ *dub* ∼ *dem*	6. **dub** ∼ *dumb* ∼ *deb*
7. **game** ∼ *Gabe* ∼ *gum*	7. **Gabe** ∼ *game* ∼ *gob*
8. **gone** ∼ *god* ∼ *gain*	8. **god** ∼ *gone* ∼ *gad*
9. **lane** ∼ *laid* ∼ *loan*	9. **laid** ∼ *lane* ∼ *loud*
10. **lawn** ∼ *laud* ∼ *lan*	10. **laud** ∼ *lawn* ∼ *lad*
11. **line** ∼ *lied* ∼ *loon*	11. **lied** ∼ *line* ∼ *lewd*
12. **loan** ∼ *load* ∼ *lane*	12. **load** ∼ *loan* ∼ *laid*
13. **pain** ∼ *paid* ∼ *pine*	13. **paid** ∼ *paid* ∼ *pod*
14. **pen** ∼ *ped* ∼ *pun*	14. **ped** ∼ *pen* ∼ *pod*
15. **rain** ∼ *raid* ∼ *ron*	15. **raid** ∼ *rain* ∼ *road*
16. **rum** ∼ *rub* ∼ *ram*	16. **rub** ∼ *run* ∼ *rib*
17. **shine** ∼ *shied* ∼ *shane*	17. **shied** ∼ *shine* ∼ *shooed*
18. **sign** ∼ *side* ∼ *sane*	18. **side** ∼ *sign* ∼ *sued*
19. **son** ∼ *sud* ∼ sin	19. **sud** ∼ *son* ∼ *said*
20. **ten** ∼ *ted* ∼ *ton*	20. **ted** ∼ *ten* ∼ *Todd*
21. **tone** ∼ *toad* ∼ *ten*	21. **toad** ∼ *tone* ∼ tide
22. **wine** ∼ *wide* ∼ *won*	22. **wide** ∼ *wine* ∼ *wooed*

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wocn.2021.101123.

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*.

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.

Appel, J., von der Pütten, A., Krämer, N. C., & Gratch, J. (2012). Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction, 2012*, 13.

Arnold, R., Tas, S., Hildebrandt, C., & Schneider, A. (2019, September 20). *An empirical analysis of voice assistants' impact on consumer behavior and assessment of emerging policy challenges (July 25, 2019)*. TPRC47: Research Conference on Communications, Information and Internet Policy, Washington DC, United States.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes, 24*(4), 527–554. https://doi.org/10.1080/01690960802299378.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

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.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, 785–821.

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.

Bradlow, Ann (2002). Confluent talker-and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 241–274). New York: Mouton de Gruyter.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272–284.

Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities. *Journal of Speech, Language, and Hearing Research, 46*(1), 80–97.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*(3–4), 255–272.

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

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

Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour, 148*(11–13), 1173–1198.

Burnham, D., Francis, E., Vollmer-Conna, U., Kitamura, C., Averkiou, V., Olley, A., ... Paterson, C. (1998). Are you my little pussy-cat? Acoustic, phonetic and affective qualities of infant-and pet-directed speech. *Fifth International Conference on Spoken Language Processing, Paper 0916*.

Burnham, D. K., Joeffry, S., & Rice, L. (2010). Computer-and human-directed speech before and after correction. In *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (pp. 13–17).

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435–1435.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language, 89*, 68–86.

Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America, 102*(4), 2360–2370.

Chiasson, S., & Gutwin, C. (2005). Testing the media equation with children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 829–838).

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in psychology* (Vol. 9, pp. 287–299). Elsevier.

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. In *Proceedings of the Cognitive Science Society* (pp. 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*, 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. (2021). Prosodic differences in human- and alexa-directed speech, but similar local intelligibility adjustments. *Frontiers Communication, 6* (675704), 1:13. https://doi.org/10.3389/fcomm.2021.675704.

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language, 28*(2), 543–571.

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.

Etzrodt, K., & Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication, 2*(1), 3.

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech. In *The handbook of phonetic sciences*, pp. 316–352). Oxford: Blackwell.

Ferguson, S. H., Poore, M. A., Shrivastav, R., Kendrick, A., McGinnis, M., & Perigoe, C. (2010). Acoustic correlates of reported clear speech strategies. *Journal of the Academy of Rehabilitative Audiology, 43*, 45–64.

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica, 57*(2–4), 242–254.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20*(1), 104.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*(5), 489–504.

Fox, R. A., & Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America, 126*(5), 2603–2618.

Fridland, V., Kendall, T., & Farrington, C. (2014). Durational and spectral differences in American English vowels: Dialect variation within and across regions. *The Journal of the Acoustical Society of America, 136*(1), 341–349.

Gergely, A., Faragó, T., Galambos, Á., & Topál, J. (2017). Differential effects of speech situations on mothers' and fathers' infant-directed and dog-directed speech: An acoustic analysis. *Scientific Reports, 7*(1), 13739.

Gottfried, T. L., & Triesch, S. K. (1993). Influence of dynamic spectral information on rate-dependent vowel perception. *The Journal of the Acoustical Society of America*, *93* (4), 2423–2423.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy, 18*(5), 797–824.

Hagiwara, R. (2005). Revisiting the Canadian English vowel space. *The Journal of the Acoustical Society of America*, *117*(4), 2461–2461.

Hargus Ferguson, S. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America, 116*(4), 2365–2373.

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*.

Hinz, N.-A., Ciardo, F., & Wykowska, A. (2019). Individual differences in attitude toward robots predict behavior in human-robot interaction. *International Conference on Social Robotics*, 64–73.

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. https://doi.org/10.21437/Interspeech.2019-2840.

Hwang, J., Brennan, S. E., & Huffman, M. K. (2015). Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design. *Journal of Memory and Language, 81*, 72–90.

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy, 4*(1), 85–110.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research*.

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42*(5), 1148–1156. https://doi.org/10.1044/jslhr.4205.1148.

Lee, D.-Y., & Baese-Berk, M. M. (2020). The maintenance of clear speech in naturalistic conversations. *The Journal of the Acoustical Society of America, 147*(5), 3702–3711. https://doi.org/10.1121/10.0001315.

Lee, K. M. (2008). Media equation theory. In *The international encyclopedia of communication* (pp. 1–4). John Wiley & Sons, Ltd..

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies, 64*(10), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*(3), 172–187.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Lunsford, R., Oviatt, S., & Arthur, A. M. (2006). Toward open-microphone engagement for multiparty interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (pp. 273–280). https://doi.org/10.1145/1180995.1181049.

Mallidi, S. H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., & Hoffmeister, B. (2018). Device-directed utterance detection. *Interspeech 2018*. Hyderabad, India: ISCA.

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.

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 https://isca-speech.org/archive/archive_papers/interspeech_2012/i12_1708.pdf.

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America, 128*(1), 435–443. https://doi.org/10.1121/1.3397384.

Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America, 96*(1), 40–55.
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. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78). https://doi.org/10.1145/259963.260288.
Nearey, T. (2013). Vowel inherent spectral change in the vowels of North American English. *Vowel Inherent Spectral Change*, 49–85.
Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Bloomington, Indiana, USA: Indiana University Linguistics Club.
Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America, 80*(5), 1297–1308.
Nieuwenhuis, R. (2016). *When size matters: Advantages of weighted effect coding in observational studies*.
Nieuwenhuis, R., te Grotenhuis, H. F., & Pelzer, B. J. (2017). *Weighted effect coding for observational data with wec*.
Nieuwenhuis, R., te Grotenhuis, M., Pelzer, B., Schmidt, A., Konig, R., Eisinga, R., & Nieuwenhuis, M. R. (2017). *Package 'wec.'*
Ohala, J. J. (1994). Acoustic study of clear speech: A test of the contrastive hypothesis. In *Proceedings of the International Symposium on Prosody* (pp. 75–89).
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.
Palanica, A., Thommandram, A., Lee, A., Li, M., & Fossat, Y. (2019). Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *Npj Digital Medicine, 2*(1), 1–6. https://doi.org/10.1038/s41746-019-0133-x.
Pelegrín-García, D., Smits, B., Brunskog, J., & Jeong, C.-H. (2011). Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical Society of America, 129*(4), 1981–1990. https://doi.org/10.1121/1.3552881.
Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 29*(4), 434–446.
Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation, 4*(2–3), 203–228.
Pycha, A., & Dahan, D. (2016). Differences in coda voicing trigger changes in gestural timing: A test case from the American English diphthong/aɪ. *Journal of Phonetics, 56*, 15–37.
Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., ... Nagar, A. (2018). Conversational AI: The science behind the Alexa Prize. *ArXiv Preprint*. ArXiv:1801.03604.
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. In *Proc. Interspeech 2019* (pp. 4005–4009). https://doi.org/10.21437/Interspeech.2019-1825.
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.
Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
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. https://doi.org/10.5281/zenodo.9846.
Rothermich, K., Harris, H. L., Sewell, K., & Bobb, S. C. (2019). Listener impressions of foreigner-directed speech: A systematic review. *Speech Communication, 112*, 22–29.
Scarborough, R. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics, 41*(6), 491–508.
Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. In *Proceedings of the International Congress of Phonetic Sciences* (pp. 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.
Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*. https://doi.org/10.1016/j.jml.2019.104038 104038.
Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics, 41*(3–4), 249–263.
Siegert, I., & Krüger, J. (2018). How do we speak with alexa: Subjective and objective assessments of changes in speaking style between hc and hh conversations. *Kognitive Systeme, 2018*(1).
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., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., & Lotz, A. (2018). Voice assistant conversation corpus (VACC): A multi-scenario dataset for addressee detection in human-computer-interaction using Amazon's ALEXA. *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. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118*(3), 1677–1688.
Snyder, C., Cohn, M., & Zellou, G. (2019). Individual variation in cognitive processing style predicts differences in phonetic imitation of device and human voices. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 116–120).
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.
Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America, 142*(4), 2469–2482.
Styler, W. (2018). *Nasality Automeasure Script Package* [Praat]. https://github.com/stylerw/styler_praat_scripts/tree/master/nasality_automeasure.
Swerts, M., Litman, D., & Hirschberg, J. (2000). Corrections in spoken dialogue systems. *Sixth International Conference on Spoken Language Processing*.
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.
Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech, Language, and Hearing Research, 39*(3), 494–509.
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.
Vertanen, K. (2006). Speech and speech recognition during dictation corrections. *Ninth International Conference on Spoken Language Processing*, 1890–1893.
Wade, E., Shriberg, E., & Price, P. (1992). User behaviors affecting speech recognition. *Second international conference on spoken language processing*.
Wade, L. (2020). The linguistic and the social intertwined: Linguistic convergence toward southern speech. *Dissertation*.
Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language, 100*, 61–88. https://doi.org/10.1016/j.jml.2018.01.001.
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., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience*, 1–15. https://doi.org/10.1080/23273798.2021.1931372.
Zellou, G., & Scarborough, R. (2015). Lexically conditioned phonetic variation in motherese: Age-of-acquisition and other word-specific factors in infant-and adult-directed speech. *Laboratory Phonology, 6*(3–4), 305–336.
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.