frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 21 May 2024
DOI 10.3389/fcomp.2024.1384252

Linguistic analysis of human-computer interaction

Georgia Zellou[1]* and Nicole Holliday[2]

[1]Department of Linguistics, University of California, Davis, Davis, CA, United States, [2]Department of Linguistics, University of California, Berkeley, Berkeley, CA, United States

OPEN ACCESS

EDITED BY
Rocco Pietrini,
Marche Polytechnic University, Italy

REVIEWED BY
Marshall A. Taylor,
New Mexico State University, United States
Matthias Wölfel,
Karlsruhe University of Applied Sciences,
Germany

*CORRESPONDENCE
Georgia Zellou
✉ gzellou@ucdavis.edu

RECEIVED 09 February 2024
ACCEPTED 24 April 2024
PUBLISHED 21 May 2024

CITATION
Zellou G and Holliday N (2024) Linguistic analysis of human-computer interaction. *Front. Comput. Sci.* 6:1384252. doi: 10.3389/fcomp.2024.1384252

This article reviews recent literature investigating speech variation in production and comprehension during spoken language communication between humans and devices. Human speech patterns toward voice-AI presents a test to our scientific understanding about speech communication and language use. First, work exploring how human-AI interactions are similar to, or different from, human-human interactions in the realm of speech variation is reviewed. In particular, we focus on studies examining how users adapt their speech when resolving linguistic misunderstandings by computers and when accommodating their speech toward devices. Next, we consider work that investigates how top-down factors in the interaction can influence users' linguistic interpretations of speech produced by technological agents and how the ways in which speech is generated (via text-to-speech synthesis, TTS) and recognized (using automatic speech recognition technology, ASR) has an effect on communication. Throughout this review, we aim to bridge both HCI frameworks and theoretical linguistic models accounting for variation in human speech. We also highlight findings in this growing area that can provide insight to the cognitive and social representations underlying linguistic communication more broadly. Additionally, we touch on the implications of this line of work for addressing major societal issues in speech technology.

KEYWORDS

speech variation, human-computer interaction, speech production, speech perception, sociolinguistics

1 Introduction

It is a new digital era: People now regularly communicate with voice-activated artificially intelligent (AI) systems, such as Siri, Google Assistant, Alexa, and ChatGPT-enabled devices, that spontaneously and naturalistically produce interactive speech. Computers have long served as mediators of communication. Yet, with the rise of voice-enabled technologies, the amount of spoken language conversations where the interactants are devices is steadily growing for many individuals who use them to complete a variety of everyday tasks (e.g., complete a shopping list, get the weather report, compose a text message, query information) (De Renesse, 2017; Ammari et al., 2019), and in some cases even for social interactions (e.g., play a game, engage in chit chat with "socialbots") (Ram et al., 2018; Perkins Booker et al., 2024). Voice-enabled technologies can also be used for applications such as speech translation (Nakamura, 2009) and "emergency media" that are used to connect users to emergency service providers (Ellcessor, 2022).

Speech patterns during conversational interactions between humans and voice-AI present a test to our scientific understanding about speech communication and language use. The speech patterns people use when talking to devices can reveal the underlying mental representations people use when producing and perceiving language, as well as the role of AI in our society, which can inform both linguistic theory and models of human-technology interaction. We argue that interpreting language patterns during human-computer interaction using both HCI frameworks and models accounting for variation in human speech can provide insight to the underlying cognitive and social representations used for linguistic communication more broadly. We also touch on the implications of this line of work for addressing social issues in speech technology. This review is informed by our stance as linguists: we believe that the research questions, tools, approach, and knowledge from linguistics to can be used to investigate language variation during HCI in order to understand how people behave toward machines, as well as to investigate the social and functional factors that govern speech communication better, in general.

In section 2, we review recent literature investigating speech variation in production and comprehension during human-computer interactions. We also summarize work exploring how human-AI interactions are similar to, or different from, human-human interactions in the realm of speech variation, focusing on resolving misunderstandings or when accommodating an interlocutor. We additionally consider how interactions with voice-AI can influence human language patterns, both for a single individual or potentially leading to change across speech communities over time. In section 3, we consider the *machine* side of human-computer spoken language interactions. We argue that applying a sociolinguistic approach to examining spoken language use with machines can shed light on factors shaping communicative success as well as the impact of human-computer interaction on user language patterns. We also highlight the need to investigate and address issues of social inequality and bias in speech technology.

2 How do humans vary their speech when interacting with devices?

2.1 Theoretical setting

There is enormous variability in how a single word is pronounced across speakers and contexts. Much theoretical work in phonetic theory is concerned with accounting for the articulatory, social, and cognitive factors that give rise to systematic variation in speech. For instance, some influential models of speech production propose that a large amount of phonetic variation during a conversation is a result of the communicative demands made on the individuals in the interaction: speakers produce more hyper-articulated words when there are cues that the listener is likely to misunderstand them (e.g., Lindblom, 1990). Another model of speech production proposes that phonetic variation during conversations has social motivations; more specifically, that people adopt the pronunciation patterns of their interlocutor as a way of conveying social closeness (or, in contrast, diverge from them to signal social distance) [i.e., communication accommodation theory (Giles, 1973; Giles et al., 1973)]. Thus, there has been much progress in linguistics in the development of models for explaining and accounting for variation during spoken language communication.

In parallel, decades of studies in the field of human-computer interaction (HCI) have been aimed at understanding how humans approach and complete tasks that involve technology. For instance, much theoretical work in HCI explores users' "mental models" for technology, i.e., what people know and believe about the devices they use (Carroll and Olson, 1988; Payne, 2007). Much like how linguists use language behavior to make deductions about the processes underlying the production and comprehension of speech, mental models in HCI work is also observed indirectly: theoretical constructs about them are built by observing, for example, differences in user behavior toward technology across tasks/systems, or comparisons of how behavior changes over time through experience with a device, or user patterns when given different types of information about the system (for review of mental models, see Staggers and Norcio, 1993; for recent work on mental models of conversational agents, see Grimes et al., 2021). HCI research is broad in scope: topics include, e.g., examining user conceptualization and behavior when using computer software, how household devices like AC units are operated, interaction with others using social media, designing and testing optimal user interfaces, interactions between people and humanoid robots, etc. Yet, a subfield of HCI is focused on *linguistic communication* during interactions with "digital interlocutors including embodied machine communicators, virtual and artificially intelligent agents (e.g., spoken dialog systems), and technologically augmented persons, either in real or virtual and augmented environments" (Edwards and Edwards, 2017: 487).

Some work in this subfield examines communication in order to understand people's mental model of the linguistic and social competence of machines (Spence, 2019). A major theoretical framework in this area was launched by the work of Nass who synthesized HCI studies with methods from social and cognitive psychology. Nass' "computers as social actors" framework (also known as 'CASA') explores the extent to which users treat technological entities as social actors during interactions (Nass and Moon, 2000; see also "Media Equation Theory"; Lee, 2008). This was investigated across a wide range of studies. For instance, Nass et al. (1999) found that, after a brief tutoring session with a computer, participants were more likely to give higher performance evaluations of a computer-tutor when that same computer was in the room, compared to when they gave the evaluation on a different computer in another room. In human-human interaction, people tend to be more positive in describing another person when that individual is present or the one asking, compared to if they are asked by another individual (e.g., Finkel et al., 1991). The Nass et al. (1999) finding was interpreted as demonstrating the transfer of 'politeness norms' to computers. Recent work has replicated this effect with smartphones (Carolus et al., 2018) and explored use of politeness terms (e.g., using "please" and "thank you") when interacting with voice-AI devices (Lopatovska and Williams, 2018). (See Ribino, 2023 for a review of work examining politeness in HCI.)

The CASA premise is that people view technological agents as social actors and this mediates their behavior toward them. Moreover, they argued that when computers use *language*, this provides even stronger cues to users that they are social beings (Nass et al., 1994). Indeed, media that use language via text (Clark, 1999) or voice (Nass and Steuer, 1993) are rated as having a strong social presence [*cf.* Social Presence Theory which explores the extent to which users conceptualize an intelligent social interactor when using technology (Biocca et al., 2003; Lee, 2004)]. Spoken language, in particular, is a socially rich type of modality for communication. Therefore, there is much potential to extend or transform theoretical understanding of people's mental models of computers by exploring speech variation specifically. Applying Media Equivalence theories to speech variation with *spoken* language technology, it is predicted that people will be prone to use the same social-structured representations and patterns of behavior from human-human interactions to interactions with technology when machines use spoken language.

Yet, recent HCI work has noted that observations that people behave as if computers are social actors does not necessarily mean that they are deemed as socially equivalent to humans. Theoretical extensions of CASA, for instance, postulate that people create technology-specific behaviors based on the particular contexts, uses, and routines in which they interact with them (i.e., "Routinized" HCI scripts; Gambino et al., 2020; Gambino and Liu, 2022). Also, some have pointed out that the nature of modern human-computer interaction is dynamic and interactive, which can change the nature of communication in different ways across contexts and systems (i.e., Theory of Interactive Media Effects (TIME); Sundar et al., 2015). These more contemporary HCI frameworks are consistent with the idea that people's "mental models" for computers can be shaped through the nature of the interaction, changing experience with technology and/or other types of knowledge users might acquire or be told about how devices work (e.g., see also Pal et al., 2023 for discussion of the factors of conversational agent and chatbot design that contribute to the perception of apparent "personality traits" in voice-AI agents.). A "routinized" account of HCI is also an apt theoretical starting point for bridging work this line of work with tools and methods from linguistics since much phonetic variation can be attributed to the particular social grounding, communicative goals, or experience-based knowledge/expectations speaker-listeners bring to a conversational interaction.

There is also recent work investigating how qualitative differences in experience with devices over the lifespan, as well as developmental factors, influence individual variation in conceptualization and behavior toward technology. For instance, researchers have noted generational shifts in behavior toward technology. Prensky (2001) defined "digital natives" as individuals who are exposed to and interact with new technologies since childhood, while "digital immigrants" are people raised without being immersed in technology. Some researchers have postulated that these developmental differences in exposure to technology result in qualitative differences in how users interact with devices (Helsper and Eynon, 2010; Kesharwani, 2020). For example, digital natives display "fluency" in using devices, are easily able to operate new technologies, as well as develop novel ways of using media effectively; in contrast, digital immigrants see new technologies as novelties and tend to utilize only learned functions (Dingli and Seychell, 2015). Not all digital immigrants are older: since not everyone is raised while being immersed in technology, there are many children in the world that can be classified as digital immigrants (Helsper and Eynon, 2010; Kincl and Štrach, 2021). Yet, beyond experience with devices, other developmental and cognitive factors might affect how people view computers as social actors. Waytz et al. (2010) found individual differences in the extent to which people anthropomorphize non-human entities, such as computers. And recent work has shown that children tend to anthropomorphize voice-AI devices more than adults (Festerling and Siraj, 2022). There is also work showing that children are more likely to engage socially during conversational interactions with voice-AI devices, by asking personal questions to understand and relate to the voice agent, compared to adults (Lovato and Piper, 2015; Lovato et al., 2019). Differences across digital natives and digital immigrants in human-computer interaction could also be expected for language use and communication behavior. For instance, digital natives see devices as tools for communication, i.e., as a means for sharing content and interacting with other individuals (Dingli and Seychell, 2015). Therefore, differences in "routinization" across digital natives and digital immigrants could vary, and impact speech and language behavior during HCI.

With this interdisciplinary theoretical landscape in mind, the rest of this paper reviews recent empirical work that can speak to issues at the intersection of phonetic variation and linguistically-mediated human-computer interaction. Spoken language is simultaneously functional and social. And, indeed, when people interact with technology there are both functional (e.g., complete a task) and social (users are projecting some amount of sociality onto computers) factors involved. Spoken language simultaneously conveys both functional (e.g., expression of lexical meanings) and social (e.g., socio-indexical features) properties (Labov, 2015) and they are present in linguistic interactions with computers, as well. Do users simply transfer their speech and language behavior from human-human interaction to communication events with technology? Or, do people develop technology-specific linguistic behaviors which reflect the unique functional and/or social roles that voice-enabled machines play in their lives? How does this vary with the type of device, type of task, or type of user? And will this change over the lifespan and across generations as technology (and people's experience) evolves?

The next section reviews work that begins to touch on these questions. We also argue that linguistic theory can advance by integrating models of phonetic variation and use with HCI frameworks. Interactions between humans and computers when spoken language is the modality introduce new avenues for synthesizing phonetic and HCI theories and empirical observations can inform both fields. We come to this new area from the perspective of academic linguists. Therefore, we focus on research at the intersection of speech production/comprehension during spoken interactions between humans and technology that can speak to fundamental questions about the cognitive and social structures underlying language variation and use.

2.2 User speech variation in production during human-computer interaction

2.2.1 Intelligibility-motivated phonetic variation when talking to technology

One of a speaker's major goals when communicating is to make their speech understood by a listener. Lindblom's (1990) hyper- and hypo- articulation (H&H) model postulates that the speaker is dynamically monitoring the likelihood for communicative success of an interaction and adjusting their acoustic-articulatory output accordingly. When the conditions are deemed to be optimal for intelligibility, speakers conserve articulatory effort by adjusting toward more hypo-articulated, reduced speech variants; yet, when speakers sense that a listener might have some difficulty comprehending for some reason, they may exert more effort to produce hyper-articulated speech forms. Recent extensions of H&H model, such as targeted adaptation accounts (Baese-Berk and Goldrick, 2009; Schertz, 2013; Buz et al., 2016), propose that hyperarticulation can be focused on the acoustic features that enhance the source of a particular misunderstanding. Indeed, decades of empirical work on "clear speech" demonstrates that speakers produce slower speech with more extreme phonetic variants of words in conditions where they believe there is a communicative barrier for a listener (Picheny et al., 1986; Krause and Braida, 2004; Smiljanić and Bradlow, 2005; Uchanski, 2005), supporting the view that speech variation is adaptive and, to a large extent, reflects the real-time communicative pressures at play during a spoken interaction between individuals.

However, recent empirical work shows that intelligibility-motivated phonetic variation is multivariate and complex. For one, while greater clear speech adjustments are found for listeners who speakers might assume have a communicative barrier [i.e., speech toward hearing impaired individuals (Picheny et al., 1986) or non-native listeners (Uther et al., 2007)], there are systematic differences in phonetic enhancements observed in clear speech across real vs. imagined interlocutors (Scarborough et al., 2007; Scarborough and Zellou, 2013), as well across other types of imagined interlocutors (Aoki and Zellou, 2024). Moreover, real listener-directed clear speech is better perceived by human comprehenders (Scarborough and Zellou, 2013), suggesting that the presence of an authentic, embodied human affects speakers' ability to recruit the most optimal mental model for the type of speech that will indeed be most intelligible in that context.

At the intersection of speech production and HCI, researchers have asked questions such as: do people have a specific device-directed speech register, or adapt their speech in response to communicative difficulty in different ways for human vs. device interlocutors? Such findings are revealing as to the mental models users have about the spoken language comprehension capabilities of machines, and, more broadly, how people establish and adapt their mental models for what speech adjustments are appropriate for different types of interlocutors. Several studies that have looked at acoustic adjustments made by speakers when talking to technology, with or without a human-directed speech comparisons, have found that device-directed speech contains more hyperarticulated phonetic variants such as louder and slower speech (Mayo et al., 2012; Siegert and Krüger, 2021). Some have also found segmental hyperarticulation in technology-directed speech, such as more extreme vowel articulations (Burnham et al., 2010). (See Cohn et al., 2022 for review of device-DS findings). Greater articulatory effort when talking to a device indicates that speakers have an assumption that there is a larger communicative barrier to overcome in HCI, relative to with human listeners (Branigan et al., 2011; Cowan et al., 2015). Thus, device-directed speech patterns suggest that people conceptualize technology as a less communicatively competent spoken language comprehender than human listeners (Cohn and Zellou, 2021; Cohn et al., 2022).

Is this the same across all users? While exploring generational, or even individual, differences in clear speech is under-studied, there is some work by Cohn et al. (2019) comparing adults' and school-age children's device- vs. human-DS that reports even greater hyperarticulation by children toward Alexa. It is hypothesized that since kids are misunderstood by ASR at a higher rate than adults (Russell and D'Arcy, 2007), they have an even greater expectation of communicative difficulties when talking to technology and therefore produce even more effortful speech toward technology.

At the same time, there is evidence that the assumption of "communicative incompetence" that people appear to project onto devices is flexible and can change over the course of an interaction depending on the nature and amount of misunderstandings made by a machine. For instance, Cohn et al. (2022) compared participants' production of words to Apple's Siri digital assistant and a human interlocutor before and after feedback (in some trials the interlocutor correctly understood the target word; in others, the interlocutor misunderstood) across studies where there was a high and low rate of listener comprehension errors. They found that overall participants spoke slower and more loudly when speaking to Siri, compared to the human, consistent with prior work and an assumption of greater comprehension difficulty for the device. However, these acoustic differences mainly emerged over the course of the interaction: in particular, people got even louder when talking to Siri over the course of the experiment. Moreover, they found greater vowel hyperarticulation following comprehension errors by Siri in the lower error rate study, not in the higher error rate study. In other words, while prosodic-level hyperarticulation was increased for Siri in all cases, targeted phoneme-level hyperarticulation was greater for Siri after an occasional comprehension error; but equivalent when both Siri and the human misunderstood most of the time.

Finally, it is important to note that the assumption of communicative incompetence can be mediated by properties of the device voices, beyond simply conceptualization of the interlocutor as a "device" vs. "human." In particular, stereotyping individuals as having certain psychological traits based on their socio-indexical features is ubiquitous in human-human interaction; for instance, women are judged to possess less communicative competence than men when reading identical political speeches (e.g., Aalberg and Jenssen, 2007). This has been shown to apply to voice-AI as well: users perceive male voice assistants as more competent than female voice assistants (Ernst and Herm-Stapelberg, 2020). Since voice-based stereotyping also occurs based on the racial and age-based cues present in talkers' speech (e.g., Kurinec and Weaver, 2021 for race; e.g., Hummert et al., 2004 for age), we predict that similar biases in judgments of communicative competence vary based on apparent ethnicity and age of device voices [see discussion of Holliday (2023) and related work in section 3]. Whether these factors influence patterns and extent of pronunciation adjustments present in device-DS is a ripe question for future work.

Taken together, the work investigating device-directed speech variation provides evidence that speakers adapt their speech production in real-time in response to the assumed and real communicative needs of a computer interlocutor. We can use speech variation toward devices, across contexts and across individuals, to reveal fine-grained changes in the mental models about what will be most intelligible to a particular listener, explore how both social and functional factors affect speech variation, and observe how speech production targets are dynamically updated as an interaction unfolds.

2.3 Vocal alignment toward speech technology

Other approaches to speech variation seek to understand how the properties in an interlocutor's speech might influence how a speaker's pronunciation changes over the course of an interaction. In particular, speakers have been shown to adopt the acoustic-phonetic properties of their interlocutor - this is known as vocal accommodation, phonetic imitation, or speech entrainment. Speech accommodation can be revealing about the nature of representations used during speech production: e.g., that they are dynamically updated based on the specific sensory information that a speaker experiences (Shockley et al., 2004). Thus, phonetic imitation is often cited as evidence supporting exemplar-based models of speech representations, which are built from stored experiences during conversational interactions (Goldinger, 1998; Goldinger and Azuma, 2004).

What can users' phonetic imitation of device speech contribute to theoretical models of speech representations? For one, the speech produced by devices is synthetically derived in some way. Synthetic speech often contains less prosodic and segmental variation compared to naturally produced speech (Németh et al., 2007; O'Mahony et al., 2021; Zellou et al., 2021) and increasing the perceived prosodic naturalness of synthetic speech does not always lead to increases in intelligibility (Cohn and Zellou, 2020). There is also evidence that synthetic speech is remembered less well than naturally-produced speech (Paris et al., 2000). One fundamental question is whether people align less toward synthetic speech, compared to naturally-produced speech. Since it contains less variation and less well remembered, it could be stored with less robust memory traces. However, a recent study compared automatic imitation of naturally-produced and computer-generated syllables (e.g., "ba," "da") and found equivalent imitative responses across speech types (Wilt et al., 2022). Also, Gessinger et al. (2021) compared phonetic imitation of prosodic and segmental patterns across natural and synthesized speech during interactions with a spoken dialog system and likewise found similar patterns of imitation across these conditions.

Moreover, speech imitation is a highly socially-mediated behavior. Communication Accommodation Theory (CAT), for instance, views people's motivation to accommodate toward their interlocutor's linguistic patterns as a function of socio-affective outcomes (Giles, 1973; Giles et al., 1973). For instance, there is much work showing that speakers align toward the speech patterns of social groups that they identify with (e.g., jocks vs. burnouts in Eckert, 1989; ethnic/nationalist identity in Mendoza-Denton, 1997). And, in conversational interactions, speakers often adopt the speech patterns of the interlocutors that they evaluate as more attractive (Babel, 2012), or who they feel a closer affinity toward (Pardo et al., 2012), or are simply more alike them (Kim et al., 2011).

Several recent studies have asked what predictions might Communication Accommodation Theory make for accommodation during human-computer interaction. For instance, Cohn et al. (2019) compared patterns of phonetic imitation by young adults shadowing words produced by Apple's Siri voices and human speakers, while also viewing images corresponding to these interlocutor types. They found overall less imitation toward the Siri voices than toward the human voices, consistent with the hypothesis that people will align to a lesser extent toward devices since they are less socially alike. Yet, there were *similar* socially-mediated patterns across voice types: people imitated male voices (both human and Siri) more than female voices. Such interlocutor gender-mediated behavior across human and computer talkers is found in prior work, too: male voiced-computers are rated as more knowledgeable on topics such as technology, whereas female voiced-computers are rated as more knowledgeable on topics such as love and relationships (Nass et al., 1997). Thus, even though there is less alignment toward device interlocutors, suggesting that device interlocutors are viewed as socially distinct from humans, people still apply gender stereotypes to technological agents based on the properties of the voice alone. More recent work finds similar biases in evaluation of robots, smart speakers, and voice assistants based on social-indexical properties of the voices (Ernst and Herm-Stapelberg, 2020; Holliday, 2023; and see Sutton et al., 2019 for discussion of biases and speech-based attitudes and discrimination as relevant for voice-AI design). The question of how such biases play out in vocal alignment behavior toward voice-AI is an open question for future work.

Another study found that the apparent *age* of the voice was an additional social variable that mediated people's alignment toward device interlocutors. Zellou et al. (2021) compared younger adults' (aged 19–39 years old) older adults' (aged 53–81) vocal alignment toward Siri and human voices and found that participants showed the largest alignment toward voices that sounded closest to them in age: older adults aligned most toward the voice rated as the oldest-sounding, which happened to be the female Siri voice; meanwhile, younger adults aligned most toward the youngest-rated voice - the male human talker. The interpretation of these cross-generational differences in alignment is that they reflect age-based socially mediated accommodation across voices: individuals of different age-identities align more strongly toward model talkers of similar apparent ages, in human-human interaction, there is even evidence of under-accommodation by older adults away from younger adult interlocutors (Giles et al., 1992) which supports socially-mediated accommodation theories. Moreover, several studies compared accommodation toward a variety of different TTS voices showing that people's rated affinity and positive attitudes of individual voices correlates with stronger degree of vocal alignment toward those voices (Cohn et al., 2023; Dodd et al., 2023). Taken together, the differential patterns of imitation across both TTS and human voices in these studies suggest that there is socially-mediated accommodation of device interlocutors based on the apparent social properties in their speech.

As soon as people interact with a device that generates spoken language, this presents an opportunity for technology to influence the speech production of the user. But, speech variation is highly socially-structured. Vocal alignment toward devices is pro-social: when the voice-AI system displays human-based social characteristics, human shadowers apply the similar patterns of phonetic imitation from human-human interaction, using decreases in acoustic distance to signal social closeness. In other words, spoken interactions with voice-AI influence human speech patterns in socially-meaningful ways. This derives from social properties apparent from the voice (gender, age, likeability). Indeed, people do display distinct social attitudes and affinities for technology, and that has been shown to influence accommodative behavior. Moreover, *users'* social characteristics (their age, gender, experience) also shape their attitudes and accommodative behavior toward machines. This supports the proposal that mental models for technology include complex, human-based social structures.

Yet, HCI frameworks propose that people develop distinct routines for behavior during interactions with technology (Gambino et al., 2020). This perspective opens avenues for future research. For instance, people most often use voice-AI technology in functional ways, such as to make a shopping list, set a timer, operate internet-of-things devices, or request information. Does vocal alignment behavior differ when people are interacting with technology while performing the most common types of tasks for these systems (*cf.* Zellou et al., 2021)? As voice-AI technological advancements introduce more diverse voices and socially-relevant contexts in which we use devices, will people align more toward these systems?

3 Listener perception of speech during human-computer interaction

3.1 Factors related to speech generation and TTS variation

Having addressed issues related to how humans produce language when interacting with non-human interlocutors, we now turn to how humans *perceive* language as produced by such technological actors. The speech produced by modern voice-AI is typically generated via a process known as text-to-speech (TTS). The speech, while derived from voice actors' productions of lots of recorded utterances, is artificially machine-synthesized following one of several waveform generation methods (see Kaur and Singh, 2023 for an in-depth review of TTS generation and methods). One waveform generation method is concatenation whereby individual acoustic chunks are selected from a database and re-stitched together via unit selection, in addition to application of prosodic-smoothing algorithms (like, pitch synchronous overlap add; PSOLA) to increase the prosodic cohesion and naturalness of a concatenated utterance. The original Siri and Alexa voices are generated via unit selection. Another speech generation method is statistical parametric speech synthesis which extracts acoustic parameters from a database and builds waveforms using a generative model (Zen et al., 2009). Parametric speech synthesis using autoregressive deep learning models trained on speaker datasets to synthesize high fidelity and highly naturalistic speech (van den Oord et al., 2016). Such neural TTS approaches are rapidly being adopted industry-wide.

Recent studies on speech synthesis have focused on questions related to how TTS generation methods affect the perceived naturalness and intelligibility of the waveform. Parametric speech synthesis methods generate speech that is evaluated as more naturalistic and human-sounding than concatenative TTS (van den Oord et al., 2016). Yet, recent work has shown that, while neural TTS is more natural sounding, it is less intelligible in a speech-in-noise transcription task than concatenative speech generated from the same speaker datasets (Cohn and Zellou, 2020). This is potentially due to the presence of more phonetic reduction and acoustic overlap present in neural TTS; while increasing phonetic reduction has the effect of creating more naturalistic sounding speech, it can also make the acoustic cues to lexical contrast less distinctive. However, with the development of more advanced techniques integrated into neural TTS methods, the loss to intelligibility can be ameliorated. New methods have been introduced that can be used to generate different types of speech variation, such as emotional prosody (Yamagishi et al., 2004), style shifting like newscaster and bedtime story register (Wood and Merritt, 2018), and even accented speech (Liu and Mak, 2020) that is not present in the original speaker dataset. For instance, the "newscaster" speech style introduced by Amazon in 2018, generated by augmented existing style-neutral TTS voices using a separate data set of newscaster-style recordings, is more intelligible than the original default neural TTS speech (Aoki et al., 2022). Moreover, the introduction of emotionally expressive interjections into TTS leads to higher social ratings of socialbot conversations by users (Cohn et al., 2019).

Speech technology firms are consistently expanding the types of voices offered in TTS systems, at least in part as a response to user demand for more diverse voices. For example, Apple's Siri Voice Assistant has expanded from offering only one American English voice option in 2010, to offering five as of Fall 2023. In a press release in February 2022, Apple stated: "We're excited to introduce a new Siri voice for English speakers, giving users more options to choose a voice that speaks to them" (Axon, 2022). Of particular interest is the fact that the new voices introduced by Apple expanded in their range of both perceived and espoused social identities. After 2010's original "American English female" Siri, the second voice to debut was "American English male," in 2013. In Spring 2021, Apple released two additional voices and revamped the original two. While the 2021 voices were in beta testing, online users began to speculate about the voices' "races" and "genders" (Waddell, 2021). Holliday (2023) found that indeed, the four Siri voices released in 2021 were evaluated differently from one another in terms of gender, age, race, and regional background, demonstrating that listeners did have differing social perceptions of them. In 2022, Apple expanded upon this pattern of introducing new, more diverse voices when it added a fifth Siri voice, "Voice 5″, also known publicly as "Quinn" (Porter, 2022). This voice represented a major shift in Apple's marketing of TTS voices, which had previously never been identified with a proper name or any demographic information about its voice actor. Apple named Quinn and publicly stated that the voice was recorded "by a member of the LGBTQ+ community" (*ibid*). In reference to the new voice, an Apple spokesperson said: "Millions of people around the world rely on Siri every day to help get things done, so we work to make the experience feel as personalized as possible" (Axon, 2022). Apple's public statements about its expansion of the Siri voice offerings indicate that they believe there is demand for voices that reflect the identities of their users.

While companies such as Apple expand their TTS offerings to contain a wider array of voices with different social identities, these strategies are not without cause for concern. Holliday (2023) observes that while listeners attach different demographic traits to the different Siri voices, they also attach negative stereotypes about those traits. Her study found that Siri Voice 3, the voice most likely to be categorized as Black, male, and young, was also judged as less competent and less professional than the other voices. This evaluation mirrors well-worn stereotypes of Black male speakers in the United States, indicating that TTS systems have the potential to reinforce and potentially reproduce negative social biases.

3.2 Top-down factors

Spoken word comprehension is a complex process. There is much work demonstrating that explicit social information provides 'top-down' influences on how an acoustic signal is perceived (e.g., Niedzielski, 1999; Hay et al., 2006; Hay and Drager, 2010). How might listeners' expectations, biases, or social knowledge shape how they perceive speech when it is produced by a device? To address this question, several recent studies have explored how people's perceptions change on the basis of different top-down information that the speech is generated by a machine or by another person. For instance, Aoki et al. (2022) compared the intelligibility of speech-in-noise when listeners were shown a picture of a device vs. when they saw a picture of a person (and were told the picture depicted the talker). They found that intelligibility of both TTS and naturally-produced speech decreased when listeners were told the speech generated by a device. In this case, it is possible that the *expectation* that machines produce less intelligible speech led to the decrease in accuracy, paralleling prior work in human-human communication that when listeners hear speech from a talker they think might have a foreign accent (i.e., a photo of an East Asian face), they show reduced comprehension (Rubin, 1992). Thus, when people *think* they will have a hard time understanding a speaker, they subsequently show worse comprehension.

At the same time, expectation that a speaker uses a non-native variety can improve comprehension if the speech is accented: McGowan (2015) had a study with a similar design as Rubin (1992), except the speech was produced by a Mandarin-accented talker and he found that an image of an Asian face *improved* comprehension. An open question is whether a similar boost for top-down knowledge that the speaker is a device could be found in contexts where speech is highly degraded (e.g., very robotic or choppy). This is an open avenue for future work.

Beyond intelligibility, top-down guise manipulations that the speaker is human vs. device have been shown to influence listeners' perception of speech in other ways, too. Zellou et al. (2023) investigated whether learning of a vowel shift differs if the listener thinks the speaker is a device or human. They exposed listeners to a voice that produced a 'dialect' of English consisting of a vowel lowering, e.g., 'beb' [bɛb] as an instance of the word *bib*, while given information that the talker was either a human or a device. After exposure, they tested if listeners' vowel category boundary had shifted for that talker, as well as whether it generalized to new talkers either in the same or different guise as the exposure talker. While learning the shift was equivalent for device and human guises, listeners showed the greatest generalization of learning from a device exposure talker to new device talkers. In other words, people appear more likely to assume that different device voices share a common "accent," than different human voices. This is further evidence that the mental models users generate about the language capabilities and patterns of device interlocutors are distinct from those for human interlocutors, and this impacts people's linguistic behavior during human-computer interaction. Here, the expectation that devices will produce speech patterns that are more homogenous and uniform across voices perhaps stems from the particular experiences that people have with device speech - that it is less variable and contains less diversity than speech across human speech communities.

3.3 Automatic speech recognition factors: machine comprehension of speech variation

If speech generation systems are machines imitating the human faculty for speech production, then speech recognition systems are machines imitating the human faculty for spoken word comprehension. Automatic speech recognition (ASR) is the technology that transforms a speech signal into corresponding text via computational algorithms. It is a critical component of voice-enabled technologies that facilitates spoken human-computer communication (See O'Shaughnessy (2023) for an in-depth review of ASR technology and developments). Much HCI work examining ASR technology has focused on how it deals with the variation present in human speech, across and within users, as well as biases stemming from ASR architecture or training that has major societal consequences.

While ASR technology has improved exponentially over the last few decades, its accuracy on non-noisy speech signals in non-ideal acoustic conditions remains far below human comprehension ability (Spille et al., 2018). Recently, researchers have explored issues related to degraded performance of ASR systems for speakers who use "non-standard" varieties of English, including marginalized varieties of United States English as well as L2 varieties (see for review Ngueajio and Washington, 2022). One of the first major papers to examine this issue is Koenecke et al. (2020) who examined word error rates across systems and dialects. They found that speakers of African American English are misrecognized at higher rates than speakers of "Mainstream" American English. The authors remark that the asymmetry in recognition accuracies "arise primarily from a performance gap in the acoustic models, suggesting that the systems are confused by the phonological, phonetic, or prosodic characteristics of African American Vernacular English rather than the grammatical or lexical characteristics" (Koenecke et al., 2020, p. 7687). Work such as this highlights a major bias in the underlying ASR training methods used by commercial speech technology systems: they simply underperform for speakers of marginalized and "non-standard" varieties (see also Wassink et al., 2022).

Another emerging issue is that biases against speakers from marginalized backgrounds can be especially problematic when ASR systems are used to give feedback to users about their language and speech patterns. Holliday and Reed (2022) examine one of the first widely-available commercial devices designed to provide feedback about a user's language practices, the Amazon Halo. The fitness tracker Halo was released in Summer 2020 and was designed as a health and wellness device. Unlike other devices in this space, the Halo contained a unique "tone" feature, which marketing by Amazon (Press Center) described in this way:

> "The globally accepted definition of health includes not just physical but also social and emotional well-being. The innovative Tone feature uses machine learning to analyze energy and positivity in a customer's voice so they can better understand how they may sound to others, helping improve their communication and relationships. For example, Tone results may reveal that a difficult work call leads to less positivity in communication with a customer's family, an indication of the impact of stress on emotional well-being".

In public-facing materials like this, Amazon claimed that the device was designed to improve the user's communication skills, but this is a fraught task due to the complexity of contextual and interpersonal factors involved in sociopragmatic interpretation as well as basic issues of processing sociolinguistic variation. In short, such a device would likely need rich social and sociolinguistic information to follow through on its claims.

In a recent study, Holliday and Reed (2022) examined how the Halo evaluated speakers of different races and genders, as well as how it responded to differences in voice quality properties. The Halo device can be activated to listen to specific speech samples that the user chooses and then to provide energy and positivity scores out of 100, as well as qualitative feedback in the form of an adjective list for each sample. Holliday and Reed found a number of concerning relationships between Halo's ratings for energy and positivity, and the gender and race, and some voice quality features, of users. First, in a task where all speakers read the same passage, the Halo demonstrated no differences in positivity ratings between speakers, indicating that it is likely not evaluating speech at all but rather using a speech-to-text model that employs sentiment analysis. In this way, the Halo is not evaluating "tone of voice" at all, but rather attaching positivity scores to lexical items. With respect to how the Halo evaluates energy, the authors find that the Halo has a strong preference for less "gender normative" voices. That is, it penalizes voices for being "too high" in F0 if the user is male, and "too low" in F0 if the voice is female. It also gives lower scores for energy to female speakers and Black speakers, reproducing biases seen in other ASR systems. Overall, users relying on the Halo for feedback to "better understand how they sound to others" are likely to receive biased results if they are women or people of color. One major issue for devices like the Halo that would claim to evaluate social and communicative well-being is that there are few reliable mechanisms for preventing bias in training data, an issue also raised by Koenecke et al. (2020). The findings of Holliday and Reed (2022) demonstrate the potential damage if such devices are not trained on a diverse set of voices, and not designed to consider existing social biases against speakers who come from sociolinguistically marginalized groups.

In general, ASR systems have a number of unique difficulties related to their ability to manage dialect diversity as well as individual speaker factors. Human listeners are able to adjust their expectations of a speaker utilizing social information to improve their word recognition. For example, a number of studies (Creel, 2018; Dossey et al., 2020) have found effects such that listener intelligibility of speakers of unfamiliar regional varieties improves with additional input. In theory, machine learning algorithms should be able to do the same, and there is evidence of training effects for a number of digital assistant systems as well. For example, Apple's Siri does utilize training data from the phone's user to improve recognition over time (Hu et al., 2019). However, voice assistants and similar technology are not able to compensate for misunderstandings using social information because they do not have access to the wealth of social and contextual information that human listeners can utilize to disambiguate signals.

Finally, ASR systems face significant challenges at the intersection of social information and the quality of the speech signal itself. Holliday (2021) compared human perception of different intonational contours in an experiment where listeners were exposed to low-pass filtered stimuli as well as original, unmanipulated stimuli, and found differences between how listeners rated the ethnicity of the speakers. Essentially, when listeners are presented with degraded stimuli, human perception of sociolinguistic variation may be altered such that they make different judgments about a speaker's race. Degraded stimuli therefore alter human ability to use social information to do on-line language processing. In theory then, ASR systems that rely on speaker recognition may be subject to the similar issues when presented with degraded stimuli. This is a particular challenge because voice assistants designed for everyday use can be presented with stimuli of varying quality, impairing a system's ability to perform speaker dialect classification and/or identification. ASR systems may perform differently in a quiet home environment as compared to a loud coffee shop, or a street with significant traffic, or when a speaker is talking farther away from the device (Wölfel and McDonough, 2009). So attempts to provide the systems with necessary input to accommodate speakers who use different dialects must also consider the real-world conditions in which the devices are likely to be used, and how noise may result in especially degraded performance for some groups, even if systems are trained on a variety of dialects.

3.4 Inequality, social justice implications, and effects on language use

In addition to concerns about how humans interact with devices, as well as inequality in both how TTS and ASR systems are designed and utilized, there are larger issues of algorithmic bias and social justice. In particular, researchers across fields have been increasingly concerned about the risk of the amplification of various types of social inequality due to increasing reliance on devices. These issues fall broadly into 3 main concerns: device accessibility, bias in access and evaluation and device impacts on user language, each of which are discussed in turn.

3.4.1 Access to devices

Perhaps the most obvious issue for a world in which speech technology devices are necessary for ever more daily tasks is the question of who has access to them in the first place. According to a 2021 analysis by Strategy Analytics, nearly half of the world's population has access to a smartphone. However, there are massive differences with respect to the quality of the devices, access to Wi-Fi, mobile, and even electricity across the world. In the United States, a nation with advanced wireless and cellular infrastructure, nearly 5% of the population has no access to broadband internet according to the FCC (Fourteenth Broadband Deployment Report). Even where broadband is available, the FCC estimates that 100 million people, or nearly 25% of the United States population, does not subscribe. These individuals are disproportionately likely to reside on tribal lands and/or in rural areas, representing significant inequality that locks entire communities out of the economic benefits of new technology. These problems are obviously much more stark in the developing world. For example, the World Economic Forum reports that 50% of people in India, or 685 million people, have no access to the internet (Ang, 2020).

As systems are developed that require internet and device access for basic functions such as banking, healthcare, education, and transportation, disconnected individuals far even farther behind. There are also immense inequalities in access to technology due to the limitations on languages that they are designed to support. There are approximately 7,000 languages spoken in the world present-day, but there are only commercially available TTS in, generously, about 50 languages. Users want to use technology in their home language (Markl and Lai, 2021). These asymmetries and gaps in language technology can lead to even larger economic and social inequalities throughout the world.

3.4.2 Bias in speech evaluation and access

There are a number of striking cases showing dramatic systematic biases in speech technology even for varieties spoken within the

United States. In particular, devices can fail to function for speakers of all types of "non-standard" varieties of English, including and especially varieties of L2 English. So far, our discussion has focused on issues that have arisen for English speakers, with implications for speakers of all languages as speech technology spreads. However, a discussion of issues related to speech technology and language variation would not be complete without an acknowledgment of the fact that many of the problems discussed above are compounded for both multilingual individuals and multilingual societies.

A number of studies, including Wu et al. (2020), Choe et al. (2022), and Dubois et al. (2024), report that popular transcription systems fail at an unacceptable rate for L2 speakers of English. Using a corpus of formal speech created from TED (Technology, Entertainment, and Design) talks, Dubois et al. (2024) tested several videoconferencing and social media platforms and revealed that the error rate for L2 speakers of English is more than double that for L1 English speakers. This represents systematic discrimination against such speakers, but also shows the difficulty that different types of automated systems have with L2-English speakers. In particular, users who rely on captions because they are deaf or hard of hearing are forced to rely on degraded output, compounding issues of accessibility for such users.

With respect to bilingual speakers, Cihan et al. (2022) observe that speakers who engage in code-switching or language mixing frequently report the failures of such technologies to recognize their speech. This generally leads to either users abandoning the technology, or being forced to adapt their language to the systems. As multilingualism is widespread across the world, these limitations affect a significant number of speakers. As Cihan et al. (2022) note, most humans are multilingual, but most voice assistants assume monolingualism. Technologies that cannot adapt to the ways that human beings use language in society are either not optimal, or they impose the restrictions of their designs on the users themselves. Monolingually-biased speech technologies which are integral to the use of cars, appliances, and phones may reinforce a United States-centric monolingual standard (Lippi-Green, 2011). Human-centric speech technology systems should consider code-switching and language mixing in the design of such systems in order for them to be both more fair and more functional for users across the world. Notably, however, advances in ASR, such as OpenAI's Whisper, do support speech recognition for more than one language at a time (e.g., Lyu et al., 2024). So, recent developments are overcoming this limitation.

Relatedly, there are significant challenges in the area of commercial translation systems, which frequently do not account for linguistic variation or the challenges of casual speech, and thus can be extremely ineffective. Such systems have exploded in popularity over the last few decades because they are often more accessible and affordable alternatives to human interpreters and translators, but an overreliance on such systems and overconfidence in their accuracy can create significant challenges, especially for lesser-resourced languages. For example, Habash (2010) discusses the challenges of machine translation systems for different dialects of Arabic, and finds poorer performance and fewer resources for local dialects than for Modern Standard Arabic (MSA). This means that users with a stronger command of MSA would receive better translation output than ones who use "less standard" dialects that the system is not trained to recognize. When translation technology is increasingly used across domains such as tourism, government, and even medicine, this has the potential to lead to systematically worse outcomes for speakers who are already disenfranchised in both linguistic and non-linguistic domains.

3.4.3 User experience and device impacts on user language

Linguists have become interested in the effects of interacting with devices on people's language use. For instance, during the COVID-19 pandemic, a number of studies found that users were making adjustments to their speech as a result of having their conversations with other people mediated by devices or software such as Zoom or Facetime (e.g., Bleaman et al., 2022).

The extent to which using speech technology leads users to change their linguistic patterns will also vary greatly across contexts and across individuals based on the variety of a language they speak. This can occur due to explicit feedback from the device, e.g., in the case of applications like "Halo," as described above, that give users feedback on their speech and language use. It also happens in implicit ways, based on the underlying design properties of the speech technologies. As outlined in several sections above, both the TTS and ASR systems underlying speech technology are trained on "standard" varieties of a language. Higher rates of comprehension failures occur disproportionately with speakers of "non-standard" varieties (Koenecke et al., 2020; Wassink et al., 2022; see also Zellou and Lahrouchi, 2024 for an examination of linguistic disparities in cross-language ASR transfer). And, in turn, this results in qualitatively different experiences for users who speak these varieties. For instance, Mengesha et al.'s (2021) diary study of Black users' experiences with voice assistants found African Americans have to accommodate their speech in order to be better understood by the speech technology. Harrington et al. (2022) also report that Black Americans experienced frustration and pressure to code-switch due to misunderstandings when interacting with a Google Home device.

In the long term, such experiences have the potential to influence language usage, such that speakers of "non-standard" varieties either implicitly or explicitly change their linguistic patterns to be understood by technology that was not designed to accommodate them. As a result, "standard" varieties of English and other languages gain additional social power because speech technologies that are necessary for everyday tasks require a command of specific varieties in order to function effectively. Users who do not or cannot conform to the speech styles that the devices were trained on may then be functionally excluded from new technologies.

One can also consider the role of voice-AI usage on child language development and use. In contrast to previous generations, many children are currently acquiring their language with non-zero input and experiences from voice-enabled technologies. What effect might this have on their language acquisition and use? This is an empirical question for future work and a ripe direction to explore what effect experience with voice-AI might have on language use and linguistic change.

4 General discussion

In this paper, we have focused on factors related both to how humans adjust their speech when interacting with machine interlocutors, and how they perceive speech from voice-enabled devices.

Section 2 focused on studies examining how speakers adapt their speech production either (1) in response to real or apparent communicative difficulties by a voice-AI interlocutor, (2) to adopt the speech patterns of the voice-AI, or (3) due to social dynamics of the interaction. Across studies, it was observed that speakers systematically change their speech during interactions with voice-AI agents. One broad generalization we can distill from this review is that users do tend to have distinct expectations and conceptualizations of the functional capabilities and social perceptions of machines. Though precisely how that affects user speech behavior varies based on the type and nature of the task. Exploring "machine" as a social category as distinct from, or similar to, humans, as well as how human-based social biases or norms are applied to technology is an area ripe for future work.

In section 3, we discussed recent issues and research related to how humans perceive the speech of voice-AI interlocutors. In particular, we focused on research showing that humans attribute social identities and stereotypes to machine interlocutors, utilizing social information from their experience with humans to do so. We also examined the use of new technologies that aim to evaluate the speech of human interlocutors, and their potential for social bias. Finally, we discussed issues related to access and inequality in a world that increasingly relies on HCI for the completion of everyday tasks.

Our review also considered how human-computer interaction work can be bridged with linguistic analysis to make interdisciplinary theoretical advancements. An example we highlighted is that the concept of mental models can be useful when applied to theoretical linguistic constructs. For instance, speakers have a conceptualization for how to adapt their speech to be best perceived by a listener, based on certain apparent social qualities (i.e., they are a non-native speaker or they have a hearing impairment) and this can be dynamically updated in response to real-time feedback about whether the interlocutor has understood an utterance or not.

At the intersection of linguistics and human-machine interaction, there is growing evidence of enormous individual or group-level variation in behavior. But, our review revealed large gaps in studies examining what factors might predict differences across users in how they approach communication with devices. Future work exploring how the cognitive, social, and experiential properties of users influencing their speech patterns toward devices can vastly expand our scientific understanding of linguistic variation during human-computer interaction.

One observation we can make from our review is that there is a considerable increase in research in these areas in the past several years alone, particularly as speech technology becomes an increasingly common and prevalent part of everyday lives. Another important aspect of HCI work based on our review is that technology is rapidly evolving. How people change their speech and language behavior in the face of different types of technology opens many empirical questions that can inform the questions raised here. Moreover, the collective experience that a society has with spoken language technology will change over generations. Thus, there is opportunity to examine real- and apparent-time differences in human-computer interaction which can further illuminate the nature of speech and language variation.

Another generalization from our review of this work is that, as many of the studies illustrate, speech is inherently social and humans use many social and contextual cues present to adapt and perceive language. However, speech technology systems do not have the ability to do this in the same way as humans. (Socio-)Linguistic analysis and insights have the potential to facilitate a wave of innovation and improvements for engineering speech technology. An open direction for future theoretical and applied work is to examine how speech technology systems can be developed to use multi-layered social information to improve communication.

Finally, a major issue that underlies much of the research in this area is the presence of bias and inequality in many speech technology systems. Exploring these inequities further is a ripe direction for future work. For instance, the majority of HCI work studying user speech production patterns (reviewed in section 2) has largely focused on white "Mainstream" American English speakers. In light of the vastly different experiences that speakers of "mainstream" and "non-standard" varieties of a language experience, investigating how users of a wide range of language varieties adapt and change their speech when interacting with devices is critical for a comprehensive understanding of HCI. It is also necessary to ensure that new technologies do not become the exclusive domain of those with linguistic and other types of social power, as these technologies become increasingly important for everyday functions.

As our review demonstrates, human-computer linguistic communication is a rich phenomenon that provides numerous avenues to test theoretical questions and concerns across disciplines. There is enormous potential for future work examining linguistic variation during HCI to enrich and elaborate linguistic theory, as well as potential for linguists to collaborate with other researchers to improve both the function and fairness of these technologies.

Author contributions

GZ: Conceptualization, Writing – original draft. NH: Conceptualization, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aalberg, T., and Jenssen, A. T. (2007). Gender stereotyping of political candidates. *Nordicom Rev.* 28, 17–32. doi: 10.1515/nor-2017-0198

Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. (2019). Music, search, and IoT: how people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact.* 26, 1–28. doi: 10.1145/3311956

Ang, C. (2020). *These are the countries where internet access is lowest*. World Economic Forum. Available at: https://www.weforum.org/agenda/2020/08/internet-users-usage-countries-change-demographics/.

Aoki, N. B., Cohn, M., and Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: effects of speaking style and visual guise. *JASA Express Lett.* 2:045204. doi: 10.1121/10.0010274

Aoki, N. B., and Zellou, G. (2024). Being clear about clear speech: intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners. *J. Phon.* 104:101328. doi: 10.1016/j.wocn.2024.101328

Axon, S. (2022) *Apple will add fifth US English Siri voice in IOS 15.4*. Ars Technica. Available at: http://arstechnica.com/gadgets/2022/02/apple-will-add-fifth-us-english-siri-voice-in-ios-15-4/.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon.* 40, 177–189. doi: 10.1016/j.wocn.2011.09.001

Baese-Berk, M., and Goldrick, M. (2009). Mechanisms of interaction in speech production. *Lang. Cogn. Proc.* 24, 527–554. doi: 10.1080/01690960802299378

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence* 12, 456–480. doi: 10.1162/105474603322761270

Bleaman, I. L., Cugno, K., and Helms, A. (2022). Medium-shifting and intraspeaker variation in conversational interviews. *Lang. Var. Chang.* 34, 305–329. doi: 10.1017/S0954394522000151

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition* 121, 41–57. doi: 10.1016/j.cognition.2011.05.011

Burnham, D., Joeffry, S., and Rice, L. (2010). Computer-and human-directed speech before and after correction. *Spaceflight* 6, 13–17,

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: feedback from interlocutors affects speakers' subsequent pronunciations. *J. Mem. Lang.* 89, 68–86. doi: 10.1016/j.jml.2015.12.009

Carolus, A., Schmidt, C., Schneider, F., Mayr, J., and Muench, R. (2018). Are people polite to smartphones? How evaluations of smartphones depend on who is asking. In: A. Carolus *Human-computer interaction. Interaction in context: 20th international conference, HCI international 2018, Las Vegas, NV, USA, July 15–20, 2018, proceedings, part II 20*. Berlin: Springer International Publishing, pp. 500–511

Carroll, J. M., and Olson, J. R. (1988). "Mental models in human-computer interaction" in *Handbook of Human-Computer Interaction*. eds. J. Vanderdonckt, P. Palanque and M. Winckler (Berlin: Springer), 45–65.

Choe, J., Chen, Y., Chan, M. P. Y., Li, A., Gao, X., and Holliday, N. (2022). *Language-specific effects on automatic speech recognition errors for world Englishes*. In Proceedings of the 29th international conference on computational linguistics (pp. 7177–7186).

Cihan, H., Wu, Y., Peña, P., Edwards, J., and Cowan, B. (2022). *Bilingual by default: voice assistants and the role of code-switching in creating a bilingual user experience*. In: Proceedings of the 4th conference on conversational user interfaces, pp. 1–4.

Clark, H. H. (1999). *How do real people communicate with virtual partners*. In: Proceedings of 1999 AAAI fall symposium, psychological models of communication in collaborative systems, pp. 43–47.

Cohn, M., Chen, C. Y., and Yu, Z. (2019). *A large-scale user study of an Alexa prize chatbot: effect of TTS dynamism on perceived quality of social dialog*. In: Proceedings of the 20th annual SIGdial meeting on discourse and dialogue, pp. 293–306.

Cohn, M., Ferenc Segedin, B., and Zellou, G. (2019). *Imitating Siri: socially-mediated vocal alignment to device and human voices*. Proceedings of the 19th International Congress of Phonetic Sciences, pp. 1813–1817.

Cohn, M., Ferenc Segedin, B., and Zellou, G. (2022). Acoustic-phonetic properties of Siri-and human-directed speech. *J. Phon.* 90:101123. doi: 10.1016/j.wocn.2021.101123

Cohn, M., Keaton, A., Beskow, J., and Zellou, G. (2023). Vocal accommodation to technology: the role of physical form. *Lang. Sci.* 99:101567. doi: 10.1016/j.langsci.2023.101567

Cohn, M., and Zellou, G. (2020). *Perception of concatenative vs. neural text-to-speech (TTS): differences in intelligibility in noise and language attitudes*. In: Proceedings of Interspeech.

Cohn, M., and Zellou, G. (2021). Prosodic differences in human-and Alexa-directed speech, but similar local intelligibility adjustments. *Front. Commun.* 6:675704. doi: 10.3389/fcomm.2021.675704

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human− computer dialogue. *Int. J. Hum. Comput. Stud.* 83, 27–42. doi: 10.1016/j.ijhcs.2015.05.008

Creel, S. C. (2018). Accent detection and social cognition: evidence of protracted learning. *Dev. Sci.* 21:e12524. doi: 10.1111/desc.12524

De Renesse, R. (2017). *Virtual digital assistants to overtake world population by 2021*. Ovum, may, 17.

Dingli, A., and Seychell, D. (2015). *The new digital natives: Cutting the chord*. Berlin: Springer.

Dodd, N., Cohn, M., and Zellou, G. (2023). Comparing alignment toward American, British, and Indian English text-to-speech (TTS) voices: influence of social attitudes and talker guise. *Front. Comput. Sci.* 5:1204211. doi: 10.3389/fcomp.2023.1204211

Dossey, E., Clopper, C. G., and Wagner, L. (2020). The development of sociolinguistic competence across the lifespan: three domains of regional dialect perception. *Lang. Learn. Dev.* 16, 330–350. doi: 10.1080/15475441.2020.1784736

Dubois, D., Holliday, N., Choffnes, D., and Waddell, K. (2024). *Fair or fare? Understanding automated transcription error Bias in social media and videoconferencing platforms*. ICWSM 2024.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. New York: Teachers College Press.

Edwards, A., and Edwards, C. (2017). "Human-machine communication in the classroom" in *Handbook of instructional communication*. eds. M. L. Houser and A. Hosek (Abingdon: Routledge), 184–194.

Ellcessor, E. (2022). *In case of emergency: How technologies mediate crisis and normalize inequality*. New York: NYU Press.

Ernst, C. P. H., and Herm-Stapelberg, N. (2020). *The impact of gender stereotyping on the perceived likability of virtual assistants*. AMCIS.

Festerling, J., and Siraj, I. (2022). Anthropomorphizing technology: a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integr. Psychol. Behav. Sci.* 56, 709–738. doi: 10.1007/s12124-021-09668-y

Finkel, S. E., Guterbock, T. M., and Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll Virginia 1989. *Public Opin. Q.* 55, 313–330. doi: 10.1086/269264

Gambino, A., Fox, J., and Ratan, R. A. (2020). Building a stronger CASA: extending the computers are social actors paradigm. *Hum. Mach. Commun.* 1, 71–85. doi: 10.30658/hmc

Gambino, A., and Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: explicating differences in processes of communication and socialization with social technologies. *Hum. Mach. Commun.* 4, 111–130. doi: 10.30658/hmc.4.6

Gessinger, I., Raveh, E., Steiner, I., and Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: behavior of groups and individuals in speech shadowing. *Speech Comm.* 127, 43–63. doi: 10.1016/j.specom.2020.12.004

Giles, H. (1973). Accent mobility: a model and some data. *Anthropol. Linguist.* 152, 87–105,

Giles, H., Coupland, N., Coupland, J., Williams, A., and Nussbaum, J. (1992). Intergenerational talk and communication with older people. *Int. J. Aging Hum. Dev.* 34, 271–297. doi: 10.2190/TCMU-0U65-XTEH-B950

Giles, H., Taylor, D. M., and Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: some Canadian data 1. *Lang. Soc.* 2, 177–192. doi: 10.1017/S0047404500000701

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295X.105.2.251

Goldinger, S. D., and Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychon. Bull. Rev.* 11, 716–722. doi: 10.3758/BF03196625

Grimes, G. M., Schuetzler, R. M., and Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decis. Support. Syst.* 144:113515. doi: 10.1016/j.dss.2021.113515

Habash, N. Y. (2010). *Introduction to Arabic natural language processing* Morgan & Claypool Publishers.

Harrington, C. N., Garg, R., Woodward, A., and Williams, D. (2022). *It's kind of like code-switching*. Black older adults' experiences with a voice assistant for health information seeking. In proceedings of the 2022 CHI conference on human factors in computing systems, pp. 1–15.

Hay, J., and Drager, K. (2010). Stuffed toys and speech perception. *Linguistics* 48, 865–892. doi: 10.1515/ling.2010.027

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phon.* 34, 458–484. doi: 10.1016/j.wocn.2005.10.001

Helsper, E. J., and Eynon, R. (2010). Digital natives: where is the evidence? *Br. Educ. Res. J.* 36, 503–520. doi: 10.1080/01411920902989227

Holliday, N. R. (2021). Perception in black and white: effects of intonational variables and filtering conditions on sociolinguistic judgments with implications for ASR. *Front. Artif. Intell.* 4:642783. doi: 10.3389/frai.2021.642783

Holliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Front. Commun.* 8:1116955. doi: 10.3389/fcomm.2023.1116955

Holliday, N., and Reed, P. E. (2022). *Effects of race, gender, and voice quality on automated "tone of voice"*. Evaluation. Paper presented at sociolinguistic symposium 2022, Ghent, Belgium.

Hu, Q., Marchi, E., Winarsky, D., Stylianou, Y., Naik, D., and Kajarekar, S. (2019). *Neural text-to-speech adaptation from low quality public recordings*. In: Speech Synthesis Workshop, No. 10.

Hummert, M. L., Garstka, T. A., Ryan, E. B., and Bonnesen, J. L. (2004). The role of age stereotypes in interpersonal communication. In: J. F. Dovidio and S. L. Gaertner *Handbook of Communication and Aging Research. 2*. Mahwah: Erlbaum, pp. 91–114.

Kaur, N., and Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: a review. *Artif. Intell. Rev.* 56, 5837–5880. doi: 10.1007/s10462-022-10315-0

Kesharwani, A. (2020). Do (how) digital natives adopt a new technology differently than digital immigrants? A longitudinal study. *Inf. Manag.* 57:103170. doi: 10.1016/j.im.2019.103170

Kim, M., Horton, W. S., and Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Lab. Phonol.* 2, 125–156. doi: 10.1515/labphon.2011.004

Kincl, T., and Štrach, P. (2021). Born digital: is there going to be a new culture of digital natives? *J. Glob. Scholars Market. Sci.* 31, 30–48. doi: 10.1080/21639159.2020.1808811

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2020). Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci.* 117, 7684–7689. doi: 10.1073/pnas.1915768117

Krause, J. C., and Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.* 115, 362–378. doi: 10.1121/1.1635842

Kurinec, C. A., and Weaver, C. A. (2021). "Sounding Black": speech Stereotypicality activates racial stereotypes and expectations about appearance. *Front. Psychol.* 12:785283. doi: 10.3389/fpsyg.2021.785283

Labov, W. (2015). "Linguistic change as a form of communication" in *Human communication*. ed. W. Labov (Abingdon: Routledge), 221–256.

Lee, K. M. (2004). Presence, explicated. *Commun. Theory* 14, 27–50. doi: 10.1111/j.1468-2885.2004.tb00302.x

Lee, K. M. (2008). *Media equation theory*. International Encyclopedia of Communication.

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory" in *Speech production and speech modelling*. ed. B. Lindblom (Dordrecht: Springer Netherlands), 403–439.

Lippi-Green, R. (2011). *English with an accent: Language, ideology and discrimination in the United States*. London: Routledge.

Liu, Z., and Mak, B. (2020). *Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment*. In: Proceeding Interspeech, pp. 2932–2936.

Lopatovska, I., and Williams, H. (2018). *Personification of the Amazon Alexa: BFF or a mindless companion*. In: Proceedings of the 2018 conference on Human Information Interaction and Retrieval, pp. 265–268.

Lovato, S., and Piper, A. M. (2015). *"Siri, is this you"? Understanding young children's interactions with voice input systems*. In: Proceedings of the 14th international conference on interaction design and children, pp. 335–338.

Lovato, S. B., Piper, A. M., and Wartella, E. A. (2019). *Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions*. In: Proceedings of the 18th ACM international conference on interaction design and children, pp. 301–313.

Lyu, K. M., Lyu, R. Y., and Chang, H. T. (2024). Real-time multilingual speech recognition and speaker diarization system based on whisper segmentation. *PeerJ Comput. Sci.* 10:e1973. doi: 10.7717/peerj-cs.1973

Markl, N., and Lai, C. (2021). *Context-sensitive evaluation of automatic speech recognition: considering user experience and language variation*. In: Proceedings of the first workshop on bridging human–computer interaction and natural language processing, pp. 34–40.

Mayo, C., Aubanel, V., and Cooke, M. (2012). *Effect of prosodic changes on speech intelligibility*. In: Thirteenth Annual Conference of the International Speech Communication Association.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Lang. Speech* 58, 502–521. doi: 10.1177/0023830914565191

Mendoza-Denton, N. C. (1997). *Chicana/Mexicana identity and linguistic variation: An ethnographic and sociolinguistic study of gang affiliation in an urban high school*. Stanford University.

Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., and Tuennerman, E. (2021). I don't think these devices are very culturally sensitive. Impact of automated speech recognition errors on African Americans. *Front. Artif. Intell.* 4:169. doi: 10.3389/frai.2021.725911

Nakamura, S. (2009). *Overcoming the language barrier with speech translation technology*. NISTEP Science and Technology Foresight Center.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nass, C., Moon, Y., and Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *J. Appl. Soc. Psychol.* 29, 1093–1109. doi: 10.1111/j.1559-1816.1999.tb00142.x

Nass, C., Moon, Y., and Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* 27, 864–876. doi: 10.1111/j.1559-1816.1997.tb00275.x

Nass, C., and Steuer, J. (1993). Voices, boxes, and sources of messages: computers and social actors. *Hum. Commun. Res.* 19, 504–527. doi: 10.1111/j.1468-2958.1993.tb00311.x

Nass, C., Steuer, J., and Tauber, E. R. (1994). *Computers are social actors*. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 72–78.

Németh, G., Fék, M., and Csapó, T. G. (2007). *Increasing prosodic variability of text-to-speech synthesizers*. In: Eighth Annual Conference of the International Speech Communication Association.

Ngueajio, M. K., and Washington, G. (2022). "Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review" in *International conference on human-computer interaction*. ed. M. K. Ngueajio (Cham: Springer Nature Switzerland), 421–440.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *J. Lang. Soc. Psychol.* 18, 62–85. doi: 10.1177/0261927X99018001005

O'Mahony, J., Oplustil-Gallegos, P., Lai, C., and King, S. (2021). *Factors affecting the evaluation of synthetic speech in context*. Proceesing SSW, pp. 148–153.

O'Shaughnessy, D. (2023). Understanding automatic speech recognition. *Comput. Speech Lang.* 83:101538. doi: 10.1016/j.csl.2023.101538

Pal, D., Vanijja, V., Thapliyal, H., and Zhang, X. (2023). What affects the usage of artificial conversational agents? An agent personality and love theory perspective. *Comput. Hum. Behav.* 145:107788. doi: 10.1016/j.chb.2023.107788

Pardo, J. S., Gibbons, R., Suppes, A., and Krauss, R. M. (2012). Phonetic convergence in college roommates. *J. Phon.* 40, 190–197. doi: 10.1016/j.wocn.2011.10.001

Paris, C. R., Thomas, M. H., Gilson, R. D., and Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Hum. Factors* 42, 421–431. doi: 10.1518/001872000779698132

Payne, S. J. (2007). Mental models in human-computer interaction. *Hum. Comput. Interact. Hand.* 17, 89–102. doi: 10.1201/9781410615862.ch3

Perkins Booker, N., Cohn, M., and Zellou, G. (2024). Linguistic patterning of laughter in human-Socialbot interactions. *Front. Commun.* 9:738. doi: 10.3389/fcomm.2024.1346738

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *J. Speech Lang. Hear. Res.* 29, 434–446. doi: 10.1044/jshr.2904.434

Porter, J. (2022). *Siri gets a new voice in iOS 15.4 beta*. The Verge. Available at: https://www.theverge.com/2022/2/23/22947150/ios-15-4-quinn-siri-voice-5-american.

Prensky, M. (2001). Digital natives, digital immigrants part 2: do they really think differently? *Horizon* 9, 1–6. doi: 10.1108/10748120110424843

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., et al. (2018). *Conversational AI: The science behind the alexa prize*. arXiv

Ribino, P. (2023). The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artif. Intell. Rev.* 56, 445–482. doi: 10.1007/s10462-023-10540-1

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Res. High. Educ.* 33, 511–531. doi: 10.1007/BF00973770

Russell, M., and D'Arcy, S. (2007). *Challenges for computer recognition of children's speech*. In: Workshop on Speech and Language Technology in Education.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., and Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *J. Acoust. Soc. Am.* 121:3044. doi: 10.1121/1.4781735

Scarborough, R., and Zellou, G. (2013). Clarity in communication:"clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J. Acoust. Soc. Am.* 134, 3793–3807. doi: 10.1121/1.4824120

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *J. Phon.* 41, 249–263. doi: 10.1016/j.wocn.2013.03.007

Shockley, K., Sabadini, L., and Fowler, C. A. (2004). Imitation in shadowing words. *Percept. Psychophys.* 66, 422–429. doi: 10.3758/BF03194890

Siegert, I., and Krüger, J. (2021). "Speech melody and speech content Didn't fit together"–differences in speech behavior for device directed and human directed interactions. *Adv. Data Sci.* 1, 65–95. doi: 10.1007/978-3-030-51870-7_4

Smiljanić, R., and Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J. Acoust. Soc. Am.* 118, 1677–1688. doi: 10.1121/1.2000788

Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: human-machine communication. *Comput. Hum. Behav.* 90, 285–287. doi: 10.1016/j.chb.2018.09.014

Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. *Comput. Speech Lang.* 48, 51–66. doi: 10.1016/j.csl.2017.10.004

Staggers, N., and Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *Int. J. Man Mach. Stud.* 38, 587–605. doi: 10.1006/imms.1993.1028

Sundar, S. S., Jia, H., Waddell, T. F., and Huang, Y. (2015). "Toward a theory of interactive media effects (TIME) four models for explaining how interface features affect user psychology" in *The Handbook of the Psychology of Communication Technology*. ed. S. S. Sundar (New York: John Wiley and Sons), 47–86.

Sutton, S. J., Foulkes, P., Kirk, D., and Lawson, S. (2019). *Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction*. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp. 1–14.

Uchanski, R. M. (2005). *Clear speech*. The Handbook of Speech Perception, pp. 207–235.

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Comm.* 49, 2–7. doi: 10.1016/j.specom.2006.10.003

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). *Wavenet: A generative model for raw audio*. arXiv 3499.

Waddell, K. (2021). *Hey Siri, is that you? Apple's new voices resonate with some Black iPhone users*. New York: Consumer Reports. Available at: https://www.consumerreports.org/digital-assistants/apples-new-sirivoices-resonate-with-some-black-iphone-users/.

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Comm.* 140, 50–70. doi: 10.1016/j.specom.2022.03.009

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* 5, 219–232. doi: 10.1177/1745691610369336

Wilt, H., Wu, Y., Trotter, A., and Adank, P. (2022). Automatic imitation of human and computer-generated vocal stimuli. *Psychon. Bull. Rev.* 30, 1093–1102. doi: 10.3758/s13423-022-02218-6

Wölfel, M., and McDonough, J. (2009). *Distant speech recognition*. New York: John Wiley and Sons.

Wood, T., and Merritt, T. (2018). *Varying speaking styles with neural text-to-speech*. Amazon Science.

Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P. R., et al. (2020). *See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers*. In: 22nd international conference on human-computer interaction with Mobile devices and services, pp. 1–9.

Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). *HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions*. In: Proceeding of Special Workshop in Maui (SWIM).

Zellou, G., Cohn, M., and Block, A. (2021). Partial compensation for coarticulatory vowel nasalization across concatenative and neural text-to-speech. *J. Acoust. Soc. Am.* 149, 3424–3436. doi: 10.1121/10.0004989

Zellou, G., Cohn, M., and Ferenc Segedin, B. (2021). Age-and gender-related differences in speech alignment toward humans and voice-AI. *Front. Commun.* 5:600361. doi: 10.3389/fcomm.2020.600361

Zellou, G., Cohn, M., and Pycha, A. (2023). Listener beliefs and perceptual learning: differences between device and human guises. *Language* 99, 692–725. doi: 10.1353/lan.2023.a914191

Zellou, G., and Lahrouchi, M. (2024). Linguistic disparities in cross-language automatic speech recognition transfer from Arabic to Tashlhiyt. *Sci. Rep.* 14:313. doi: 10.1038/s41598-023-50516-3

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Comm.* 51, 1039–1064. doi: 10.1016/j.specom.2009.04.004