One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices

JULIA CAMBRE, Human-Computer Interaction Institute, Carnegie Mellon University
CHINMAY KULKARNI, Human-Computer Interaction Institute, Carnegie Mellon University

When a smart device talks, what should its voice sound like? Voice-enabled devices are becoming a ubiquitous presence in our everyday lives. Simultaneously, speech synthesis technology is rapidly improving, making it possible to generate increasingly varied and realistic computerized voices. Despite the flexibility and richness of expression that technology now affords, today's most common voice assistants often have female-sounding, polite, and playful voices by default. In this paper, we examine the social consequences of voice design, and introduce a simple research framework for understanding how voice affects how we perceive and interact with smart devices. Based on the foundational paradigm of computers as social actors, and informed by research in human-robot interaction, this framework demonstrates how voice design depends on a complex interplay between characteristics of the user, device, and context. Through this framework, we propose a set of guiding questions to inform future research in the space of voice design for smart devices.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: voice interface; voice assistants; human-robot interaction; IoT; voice design; speech interfaces; intelligent personal assistant; voice user interface

ACM Reference Format:
Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 223 (November 2019), 19 pages. https://doi.org/10.1145/3359325

1 INTRODUCTION

The human voice is rich in social information. Independent of content, the sound of a voice conveys several signals that humans are naturally attuned to recognize, such as the gender, age, and personality of the speaker [35, 47, 59]. As Nass and Brave (2005) note in their book, *Wired For Speech*, these powerful responses to voice were evolved to facilitate human-human conversation, provoking a crucial research question: *"How will a voice-activated brain that associates voice with social relationships react when confronted with technologies that talk or listen?"* [59].

In the years since, voice technology has become ubiquitous: already, 46% of adults in the United States use a voice assistant on a daily basis [62], and estimates suggest that there will be upwards of 8 billion voice assistants worldwide by 2023 [69]. While smart speakers and smartphones may be largely driving this growth, there is also a growing trend towards embedding voice assistants in a diverse range of "smart" devices: these range from in-car navigation and entertainment systems, to microwaves, thermostats, and even toilets [15, 51, 81]. At the same time, speech synthesis technology

Authors' addresses: Julia Cambre, jcambre@cs.cmu.edu, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213; Chinmay Kulkarni, chinmayk@cs.cmu.edu, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213.

2573-0142/2019/11-ART223 $15.00
https://doi.org/10.1145/3359325

has also advanced considerably in recent years; new models based on deep neural networks such as WaveNet are now capable of generating increasingly varied and more human-sounding speech compared to prior approaches like concatenative or parametric synthesis [30, 63]. This explosion in the popularity and pervasiveness of voice interfaces—along with rapid improvements in speech technology—adds new urgency and complexity to the question Nass and Brave raised nearly 15 years ago.

Within the Human-Computer Interaction and Computer-Supported Cooperative Work community, this trend has not gone unnoticed. In recent years, researchers have studied voice assistants from a number of angles. Several papers have explored users' patterns of everyday use with common voice assistants like Alexa, Siri, and the Google Assistant [3, 44, 66, 70]. Others have considered usability challenges faced by natural language processing errors [58, 74], and future use scenarios such as leveraging speech to navigate videos [13] or promote workplace reflection [36]. There have also been efforts to establish a more theoretical or vision-setting perspective on voice technology: for example, Cohen et al. [18] and Shneiderman [72] have weighed in on the merits of voice as an interaction medium, while Murad et al. [56] proposed an initial set of design guidelines for voice interfaces. Within the CSCW community specifically, voice interactions have also received considerable attention in recent years, with papers and workshops on topics ranging from accessibility [12], to automated meeting support [49], Wizard of Oz prototyping techniques [45], privacy [37], multi-user interaction [67], and more. While these papers all offer useful perspectives on voice interface design, their focus has almost exclusively been on *what* voice assistants say in conversation, rather than on *how* they say it.

This paper poses a seemingly straightforward question: *What should the voices of our smart devices sound like?* Specifically, as we move towards a future in which users interact through speech with not just smartphones and smart speakers, but with an increasing array of everyday objects, selecting a voice identity for these smart devices remains an open design challenge with important social consequences.

This paper introduces a research framework for understanding the social implications of design decisions in voice design. To demonstrate the utility of this framework, we both summarize existing research using it, and discuss a sampling of new research questions it generates. To generate this framework, we consider the design space of smart device voices, and organize the literature around what we know about how the features of a synthesized voice shape our interactions with speech-enabled technology. In doing so, we rely heavily on research in human-robot interaction (HRI), while still incorporating research from other fields such as social psychology and design research.

We are not the first to propose a framework for voice design. For example, Clark et al [16] mapped out the existing space of research on voice in HCI through a recent review of 68 papers. Through this review, the authors suggest a set of open challenges for the field, including a need for further design work and studies of multi-user interaction contexts. Importantly, however, their review deliberately excluded papers focusing on *embodied* interfaces. Our framework complements Clark et al.'s review by focusing explicitly on this area of embodied voice design.

Our HRI-based perspective also distinguishes this paper from recent work that studies the design of speech interfaces with voice in isolation. For example, Sutton et al. propose a framework based on findings from socio-phonetics [71]. While studying voices in isolation prevents the confounding effects of voice with the effects of embodiment, in practice embodiment, form-factor, and contexts of use do indeed influence how people perceive voice interfaces and social robots [23, 28, 34, 50]. In our work, we hold that these attributes are not undesirable confounds, but necessary dimensions of analysis: smart devices necessarily will possess form, contexts of use, and perhaps even human-like embodiment. Thus, because embodiment and form and voice together affect perception is precisely

why we they should be studied in a holistic fashion. Therefore, an HRI-based perspective that combines embodiment and voice offers more holistic guidance that would be difficult if these factors were studied in isolation.

The lack of research frameworks that consider embodiment might also be responsible in part for current practice that seems to be moving towards a "one voice fits all" approach, with large companies embedding their respective assistant service across as many supporting devices as possible. Recent reports from Amazon indicate that there are over 28,000 Alexa-enabled smart home devices [10], meaning that a given user could own a microwave, car, smoke detector, and more, all of which speak with the same synthesized voice.

On the one hand, companies may favor this design choice as it helps solidify their brand identity and ensures a more consistent experience across products. Early work on speech interfaces has also suggested that using the same voice for multiple services can increase perceptions of intelligence in the voice persona [59, p.105-112]. On the other hand, the trend towards uniform assistant identities has drawn repeated criticism from popular press [76]. Feminist HCI researchers have also justly criticized these decisions [77], particularly because today's main voice assistants (e.g. Siri, Alexa, the Google Assistant) take on female, polite, and friendly voices by default in many locales. As journalist Chandra Steele writes, "companies have repeatedly launched these products with female voices and, in some cases, names. But when we can only see a woman, even an artificial one, in that position, we enforce a harmful culture" [76]. Indeed, a recent report by the UN cited artificial intelligence—and particularly the personification of many voice assistants as young women—as responsible for perpetuating harmful gender stereotypes [82]. To us, these design decisions and their corresponding critiques underscore the need for a framework that carefully considers embodiment, paralinguistic aspects of voice design, and their social implications together.

This paper contributes a novel research framework for understanding the design of smart device voices. Drawing upon literature from human-robot interaction and other fields, we synthesize three lenses that we believe are particularly useful in voice design: *user*, *device*, and *context*. Through this work, we hope to open the conversation and inform new research directions in the rapidly evolving space of voice-based interaction.

2 THEORETICAL FOUNDATIONS: FROM HUMAN-HUMAN TO HUMAN-COMPUTER

A rich history of research suggests that human-computer interactions largely parallel human-human interactions. In their 1994 paper, Nass et al. [61] asserted that humans engage with computers in ways that are consistently and fundamentally social: a user behaves towards computers as they might towards other human beings, despite knowing intuitively that computers are not animate [61]. This theory, known as the "Computers are Social Actors" paradigm, has become an influential blueprint for a long line of subsequent research in social computing. Through a series of five experimental studies, Nass et al. systematically replicated several key findings from literature in sociology and psychology that had been well-established patterns of interpersonal behavior. Particularly relevant to this discussion of voice interaction, their findings suggested that people naturally use voice (rather than a device's physical "box") to differentiate computer identities, and that people automatically ascribe gender stereotypes to computers as well. The conclusion that voice serves as the key feature for distinguishing between computers was elaborated over two experiments, where the authors first found that users considered different computer boxes that spoke with different voices as distinct intelligences, and built upon this to find that "subjects responded to different voices as if they were distinct social actors, and to the same voice as if it were the same social actor, regardless of whether the different voice was on the same or different computer" [61]. In the context of today's smart device ecosystem, this finding has important

implications, suggesting that users may consider all the devices that share a common voice (e.g. all Alexa-enabled objects) to have a common intelligence [50].

The CASA paradigm suggests that models of human-human interactions might inform how users might respond to social forms of technology. For example, it is possible that models of collaboration, trust, or even social support between people may apply to interactions between people and voice interfaces. This suggests a central role in this research for the CSCW community, which has long studied these models.

Within the scope of this paper, we focus specifically on aspects of impression formation and management (i.e. how users form initial impressions of others, and how they manage others' impressions of themselves.) Clearly, impression formation and management has immediate and profound effects on interaction. For example, consider the well-established phenomenon from the social psychological literature of *thin slicing*, which suggests that people make rapid, but often accurate judgments with only brief glimpses of behavior. For example, in a famous study by Ambady and Rosenthal [2], participants were able to judge the teaching ratings of college professors from short, silent video clips (from 10 seconds, down to even 2 seconds) with high accuracy compared to end-of-semester student ratings [2]. Others have shown similar effects for speech. McAleer et al. [47] investigated how robustly people could predict personality traits from an extremely brief sample of a speaker's voice. Participants listened to audio clips of various speakers saying the word "hello," resulting in sub-second exposure to each voice (recordings were 390ms on average). Listeners were highly consistent in how they rated perceived personality traits of the voices [47]. These results suggest that people form rapid judgments about a person's characteristics through their voice.

Similarly, does impression formation by thin slicing apply to voice-based agents as well? Indeed, results by Chang et al. suggest this may be the case [14]. They presented participants with 10 second clips of eight "candidate" voices for a caregiving robot, which varied along gender, age, and personality characteristics. Impression formation theory suggests that these short clips should lead participants to correlate them with personality traits, and indeed despite the range of potential voice options, participants overwhelmingly tended to prefer extroverted female voices, aligned with stereotypes of humans that take on caregiving roles.

Results such as these and others presented in the sections that follow suggest that because voice-based agents (as computers) are social agents, impression-formation and management processes might therefore have immediate and profound implications for how that device is perceived. This observation also suggests a framework for voice-design: designing voices can be seen as analogous to impression-management. Just as human impression management is mediated by physical characteristics, traits, and behaviors in context (or on-stage) [26], voice design can be seen as mediated by device characteristics, interactional traits with users, and contextual issues. Other scholars have also similarly hypothesized that voice design is analogous to designing for *performance* i.e. on-stage behaviors, but also that "we are a long way from realizing a sense of performance from speech systems" [6]. With our framework, we hope to fill in this gap.

3 THREE LENSES FOR VOICE DESIGN: USER, DEVICE, CONTEXT

While the theory of impression management and performance give an overall guiding principle, what concrete features must researchers and designers focus on while designing voice interfaces?

A large body of related work from nearby fields—particularly in human-robot interaction (HRI)—considers aspects of impression management and performance, albeit often indirectly as questions of embodiment and paralinguistics. We draw upon these results here to inform our design space, and to make concrete recommendations for future research. Specifically, we considered the broader

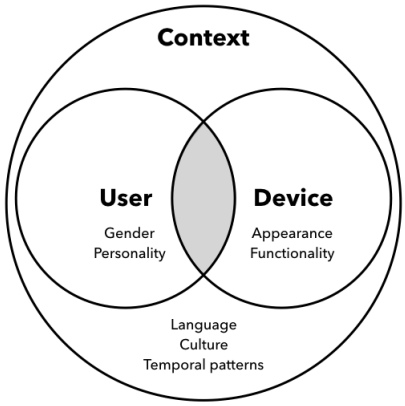

Fig. 1. Overview of the conceptual model for smart device voice design. Here, we argue that voice design should be considered through three lenses: user (representing aspects of the user's identity, such as gender and personality), device (concerning the smart device's appearance and functionality), and context (aspects of the situation in which the device is used, such as language and culture, or longitudinal changes). The amount of overlap depicted between user and device characteristics (light grey) may vary depending on the designer's goals for the interaction.

space of research on human-agent interactions (where an agent might be a robot or a disembodied voice), with an eye towards studies on impression formation and voice characteristics.

One point of departure from our impression management metaphor is that unlike humans, nearly every aspect of voice interfaces is malleable (humans, on the other hand, find it challenging to change physical attributes such as their height). Therefore, as an organizing framework we eschew the rich, nuanced models of impression formation and take inspiration from a simple model introduced by Mutlu et al. [57]. Mutlu et al. suggest that social interaction in human-robot interaction emerges from three components: *user attributes*, *robot attributes*, and *task structure*. In their model, user attributes constitute demographic information like the user's age or gender; robot attributes are aspects of robot's appearance or other features that suggest personality, such as voice; and task structure refers to whether the activity that the user and robot perform together involves cooperation, competition, or other shared behavior like planning.

Taking these three elements as a starting point, we propose a slightly modified version of the model for the particular use case of designing voices for smart devices, consisting of: *user*, *device*, and *context*. The following sections define and discuss related literature for each of these lenses in turn.

These lenses are one way to simplify the organization of this literature, but these lenses are not intended to be mutually exclusive; instead, we conceptualize them as modeled in Figure 1, where user-device relationship may share some amount of overlap, and are together situated within the broader contextual concerns we will describe. Note that this paper introduces our framework, but the task of filling it in is far from complete: the most obvious omission is around how linguistic content affects speech-based interaction, which is thoroughly studied in [16]. Specific linguistic aspects will enrich each of the lenses we describe.

Finally, even though the majority of our examples below are from the HRI and Communications community, there is a growing number of studies that directly addressed voice design in the context of smart devices. We hope our paper offers a guiding framework for such work in the future.

3.1 User

One of the most pervasive themes that emerged from the literature was a focus on the user's identity, and on how personal attributes affect responses to an agent (robot or voice). Through this lens, a person's characteristics (e.g. their gender, personality traits, etc.) serve as an anchor; studies that take this approach generally measured the user's attributes and looked for interactions based on whether the agent's attributes matched.

Within the domain of human-robot interaction, prior work has found that users can not only identify personality traits in a robot based on verbal and non-verbal behavior, but are attracted to robots that had a personality complementary to their own [39]. In a study by Lee et al., participants played with an AIBO, a social robot that resembles and plays like a dog. Equal numbers of introverted and extroverted participants were randomly assigned to interact with either an introverted or extroverted version of the AIBO dog; to simulate the AIBO's introversion / extroversion, the researchers adjusted features like the loudness and pitch of its synthesized voice, and manipulated the AIBO's physical movements to match personality traits (e.g. making larger and faster movements to signal extroversion). The study found strong evidence for a "complementarity attraction effect" with the AIBO: in other words, participants felt more positively towards an AIBO that complemented their own personality in introversion / extroversion, as measured by responses to ratings of intelligence and social attractiveness [39].

Interestingly, these findings are somewhat inconsistent with earlier work on disembodied computerized voices, which found that users preferred *voices* that exhibited a similar personality to their own [60]. As Lee et al. discuss, this discrepancy may be a consequence of how much sensory information people have when interacting with a voice versus a robot: "we believe that there is a fundamental difference between the interaction with disembodied agents and the interaction with embodied agent" [39].

Other work has investigated whether a user's age may also influence their perception of agent voices. Chang et al. [14] explored how different synthetic voices were perceived by baby boomers in Taiwan; their focus was on voices embedded within social robots given the potential future caregiving applications. In the study, participants first watched a prototype video of "ELLIQ" (a care-giving robot that reminds users to take medicine, call family, and so on), presented with Chinese language subtitles. Participants were then presented with 10 second clips of eight "candidate" voices, which were prerecorded human voices. The voices were chosen to vary on gender, age, and personality characteristics. Despite the range of potential voice options, participants overwhelmingly tended to prefer extroverted female voices; there was no significant difference in preference for younger versus older sounding voices [14].

3.2 Open questions about user-centric voice design

Individualization of voices To what extent should the voice of a smart device be tailored to its user? Voice assistants currently take a largely "one size fits all" approach in which each instance of a given device takes on the same voice by default; indeed, the same voice is often used *across* devices powered by the same company's assistant software (e.g. Alexa-enabled or Google Assistant-enabled devices). However, these studies on user characteristics suggest the alignment between user demographics and the demographics suggested by a device's voice likely play a crucial role in affecting interaction. This invites two open questions. First, if a device is used by more than one person, how might it adaptively individualize its voice to match multiple people? Previous work in HRI has found that robots which engage in vocal entrainment—changing the pitch, speaking rate, intensity, and other features of speech to mirror the user—have positive social outcomes by improving perceptions of rapport, trustworthiness, and learning [42, 43]. Such real-time voice

adaptations between smart devices and the user or users might yield similar benefits. Second, individualization may suggest either voices that are similar to users', or with attributes which are complementary. We examine this in more detail below.

Similarity vs. complementarity Following from the discrepant results between personality alignment preferences with robotic agents versus disembodied computer voices, one rich area for research is how the *degree of embodiment* affects similarity versus complementarity attraction effects in voice characteristics. Lee and Nass used a desktop computer with headphones as their source of their disembodied voice [60]. In the 18 years since, voice devices have vastly increased in diversity. Future research may thus investigate what characteristics of embodiment might work better with similar or complementary voices.

User preferences with multiple devices In the above studies, users were exposed to only a single robot or voice agent. One area that remains less explored within the user-level framing is understanding how robust these effects and preferences are across *multiple* devices. One possibility for future research is to investigate voice preferences for users who are surrounded by multiple robots or devices capable of interacting through speech. For instance, if an individual owns several smart home devices, should the devices all take on the same voice identity, or each speak with a subtly different voice?

3.3 Device

How might features of the device influence preferences and expectations for the device's voice? In what ways could a device's *appearance* or *stereotypes associated with its functionality* affect how people perceive it?

3.3.1 Appearance. The human-robot interaction community has long been interested in studying anthropomorphic tendencies towards robots. As one example, Kalegina et al. [32] systematically examined 157 robots with screen-based faces and coded for 76 nuanced features like eye color, mouth shape, and the presence of eyebrows. Through two surveys, they identified correlations between facial features and anthropomorphized traits; for example, robots that had cheeks were perceived as significantly more feminine and childlike than those without, whereas robots lacking a mouth were perceived as unfriendly and creepy [32]. Similarly, other studies have found that minimal visual cues can activate automatic stereotypes with robots. Replicating prior results finding that cues in the robot's appearance affected the perceived gender of the robot, [31] found that robots fashioned in gender-stereotypical ways (with pink earmuffs or with a black hat) were perceived as female and male, respectively.

These inclinations to anthropomorphize based on superficial characteristics of a robot's design can also reveal implicit biases that extend from human-human interactions into human-agent interactions. In a recent paper, Bartneck et al. found that people automatically attributed a race to a robot based on superficial physical characteristics, and revealed a bias towards both Black individuals and robot agents racialized as Black in the context of the study. The experiment adopted the "shooter bias" method from social psychology: participants were presented with a series of images in rapid succession, and were asked to simulate the role of a police officer deciding whether to "shoot" at the subject in the picture, who is either carrying a gun or some other harmless object, like a cell phone or wallet [7]. Previous studies depicting human subjects have found a tendency to shoot Black subjects more readily than White subjects. The authors also interpret their results as pointing to a troublesome trend in which most robots are stylized in such a way that they lack implied racial diversity. Whether and how people attribute race to robots is an active area of study, and it is likely that the relationships among designed color, interpreted race, and social consequences that result from either will become more clear over the next several years.

Similar research has specifically investigated how a robot's appearance shapes expectations and perceptions of the robot's voice. From both theoretical and empirical standpoints, much of the prior work in this space points to a fundamental mismatch between text-to-speech voices and the device or robot; according to Moore [55], voice-based systems face a "habitability gap" because endowing a system with a voice that sounds as human-like as possible misleads the user into overestimating the system's intelligence. To address this concern, Moore proposed a notion of "vocal appropriateness" in which the voice selected for a given artefact should set proper expectations for the user by "aligning an artefact's visual, vocal and behavioral affordances" [54] (e.g. giving a robot a voice that sounds robotic). Echoing the notion that device characteristics are a means of impression management, a recent provocation piece [6] suggests that instead of merely designing highly naturalistic voices, designers should also strive for abstraction and deliberate meaning (i.e. *performance*).

Studies on robot voice design offer evidence of how the interplay between voice and appearance affect the user experience. For example, in a study to probe what mental images for a robot different voices evoked, McGinn and Torre [48] presented participants with a voice clip and asked the participants to match it to the corresponding robot from a selection of static images. The gender and naturalness of the voice had a substantial affect on which robot users believed the voice corresponded to, with participants often attributing the female voices to the robots described as having rounder, more friendly, and less mechanical appearances. Other work by Moore [54] tested the appropriateness of two candidate voices for the Nabaztag robot, a small robot resembling an animated character version of a rabbit. Participants were significantly more likely to perceive a childlike voice as "belonging to" or coming from the robot compared to an adult male voice, which provides evidence for the hypothesis that users expect a robot's voice to match its appearance. Follow-up work from Moore explored this notion of the alignment between robot form and voice through the design of MiRo, a biomimetic robot modeled after a "generic mammal." Following from the other characteristics of the robot's design, MiRo's voice was designed to directly mimic the "physical and behavioral characteristics of the robot" through real-time synthesis, and is modeled after mammalian vocal characteristics (e.g. "happier" vocalisations when in a state of high valence) [54].

3.3.2 Role. Other studies have shown that users' biases may manifest not just through appearance, but also in the *role* that the agent or device takes on.

Motivated by research on stereotypes and expectancy violation in human-human interaction, such as gender biases in hiring, Tay et al. [78] explored whether the same biases manifested in human-robot interactions; they were specifically interested in a growing category of robots that take on roles typically occupied by humans in social settings, such as providing at-home aid or acting as guides in public places like museums or train stations. Using similar methods as in the Lee et al. study mentioned previously [39], the researchers manipulated the robot's perceived personality (introversion / extroversion) and gender through voice (selecting male or female text-to-speech voices), and by giving the robots stereotypically gendered names (Joan vs. John). To manipulate gender roles, they constructed two scenarios related to healthcare and security. In these scenarios, the task itself was held constant, but participants were placed into different conditions based on the gender and personality that the robot presented. Gender did not significantly affect participants' perceived sense of trust in robots. However, participants responded more positively to robots that matched gender and personality stereotypes, particularly through perceived behavioral control, and more positive affective responses: "In addition to treating artificial agents as actual human beings, people transfer their traditional gender and personality stereotypes to social robots that undertake human occupational roles" [78].

3.4 Open questions about device-centric voice design

These findings demonstrate that features of the device itself—whether visual cues or associations between its functionality and other stereotyped social roles—can affect how users interact with the device. While the studies mentioned in this section all come from a human-robot interaction perspective, we expect that many of the same design considerations will transfer to smart devices as well.

Physical cues: appearance, movement, and sound While a voice-enabled smart device may not be embodied to the same degree as a robot, the voice nevertheless occupies a physical form. The fact that the voice is embedded within a physical object may therefore trigger many of the same anthropomorphic tendencies that HRI studies have reported in response to embodied robots. Much like in the studies cited above regarding robot appearance, certain qualities of the device itself may shape users' preferences and expectations in similar ways. For example, just as color can serve as a powerful cue in toys in gender stereotypical ways [65], the color of a smart device might predispose a user to expect a certain gendered voice from that device (e.g. expecting a blue blender to have a masculine-sounding voice, and a pink version of the blender model to have a feminine-sounding voice). The same might also be true of the relative size, stylization, and material form of a given smart device; for example, smaller devices, and those with colorful prints and durable materials might be perceived as a child's device (and therefore might take on a voice persona that sounds the appropriate age).

Considering voice design through the lens of the device also highlights physical properties of the device beyond its appearance and aesthetic. Smart devices may vary considerably in their range of movement, which may in turn impact the characteristics that a user expects from the device's voice. For example, certain devices will remain fixed in place (e.g. parking meters, kitchen stoves) by virtue of their size, whereas others may be portable (e.g. a smartwatch or baby stroller). Still other devices may *facilitate* movement, such as a car or electric scooter. Whether and how the voices for these devices should change if they are in motion vs. stationary, or adapt according to speed remains an open question.

Finally, with the exception of smart speakers, most voice-enabled devices perform some function *beyond* audio input and output; through their everyday operation, many of these devices emit sounds associated with their use. For instance, a vacuum cleaner and a refrigerator make distinct hums at different pitches, while a tea kettle may emit a high-pitched whistle. Within human-robot interaction contexts, these "consequential" sounds (e.g. the sound of servo motors actuating a robotic arm) negatively affect users' perceptions of a robot [53, 79]. How might the mechanical sounds that these devices produce shape a user's expectations around a smart device's voice?

Associations with functionality Device-level effects may also emerge in cultural associations with the device's function. Just as participants in the Tay et al. study felt more positively towards robots that conformed to social stereotypes about gendered occupational roles, the same may apply to how users' evaluate a device's voice [78]. Regardless of appearance, many everyday objects may be seen as gendered (e.g. through the lens of traditional gender roles in household work, a washing machine might be perceived as more feminine, whereas a power drill might be seen as a more masculine object); these gender associations might influence the voice that the user expects from the device.

Long-term use Exploring voice design through the lens of the device also invites interesting questions about the device's physicality and wear. Should the voice of a smart device *change over time*, just as a person's voice changes as they age? For many voice-enabled devices, the life-cycle of ownership may be too short to warrant such a change. However, other types of devices such as cars, built-in smart home systems, or other large investment-like purchases may remain in service

for decades, and potentially for generations. Despite their digitally-mediated form, these types of possessions may come to be what Ikemiya and Rosner et al. consider "heirloom artifacts" [29]. Should the voice that these heirloom objects take on evolve over several years of use, or could the "original" voice identity itself one day be considered a type of collectible? If the device itself becomes physically damaged, should that be reflected in the sound of the device's voice?

3.5 Context

Users' experiences with a technology rarely exist in isolation. Beyond the personal user attributes and characteristics of the device, the final consideration that may guide voice design is the surrounding context of use. By this, we are referring especially to broader cultural, temporal, or linguistic factors that may affect perceptions and expectations for voice devices. To our knowledge, these broader contextual factors have not received as much attention or study as user or device-level features thus far.

One recent case study that takes a contextual perspective describes the challenges involved in designing a voice for the Oakley Radar Pace, which are exercise sunglasses with an embedded voice persona that acts as a workout coach [20]. To develop personas that would be successful within the product's five language markets, the authors took both a cross-cultural and cross-linguistic perspective in considering how to design the voice assistant's persona along four dimensions (embodiment, register, gender, and personality traits such as authoritativeness vs. firmness) while also accounting for Oakley's brand identity. Across all locales, they found that users preferred a voice assistant that engaged using an informal tone (akin to a casual, friendly conversation). However, preferences for features like gender and personality traits differed by language and by region. For example, in order for the voice to sound credible as a coach, it was not sufficient to simply choose a voice with the correct language; they instead used a customized text-to-speech voice that matched in regional accent and "sounded like" it was from the locale [20].

Indeed, when companies fail to perform proper market research and tailor voice design to a cultural demographic, the resulting product can backfire in sales and customer satisfaction. According to Nass and Brave [59, p. 55], this was the case for BMW when they released a car navigation system in the German market with a female voice. Shortly after the car's release, the automaker received complaints from customers demanding that they change the system's voice; as Nass and Brave report, drivers in Germany felt uncomfortable and untrusting of receiving driving directions from women. In response, BMW issued a product recall to update the car's navigation system, yielding to pressure to find a voice that matched their branding and cultural concerns [59].

Both the Oakley Radar Pace and the BMW navigation systems point to the complexity of designing voices for smart devices across contexts; for both products, there was no globally-optimal voice that would have resonated with users across domains or devices. Instead, the reasons that a voice succeeded or failed was influenced by the surrounding cultural context.

3.6 Open questions about context-centric voice design

Looking beyond culture, how else might context influence voice design for smart objects?

Linguistic cues Language is integral to a context of use. Indeed, according to the Sapir-Whorf hypothesis, the language that an individual speaks shapes the way they think [21]. One such way in which this manifests is through grammatical constructs: many languages are grammatically gendered, meaning that the words for inanimate objects are masculine or feminine. For example, the word for "dishwasher" is masculine in Spanish, yet feminine in German. Would these linguistic cues make it more likely for a Spanish speaker to expect a male voice from a smart dishwasher, and more likely for a German speaker to expect a female voice from the same smart dishwasher? Prior research by Boroditsky et al. [11] suggests that German and Spanish speakers were more

Lens	Open Question	Theme
User	To what extent should the voice of a smart device be tailored to its user?	Individualization
User	If a device is used by more than one person, how might it adaptively individualize its voice to match multiple people?	Individualization
User	Under what circumstances should voice characteristics be similar to a user's, and when should voice characteristics be complementary?	Similarity vs. complementarity
User	How does the degree of embodiment of a voice agent affect whether synthesized voices should be similar or complementary?	Similarity vs. complementarity
User	If an individual owns or interacts with several smart devices, should all the devices share the same voice identity, or each speak with a subtly different voice?	Multi-device ecosystems
Device	Could the color, size, stylization, and material form of a smart device trigger stereotypes about the voice's gender or other characteristics?	Physical cues: appearance
Device	Should the voice of a smart device change or adapt depending on its range of motion?	Physical cues: movement
Device	How might the mechanical sounds that smart devices produce (associated with their regular function) shape a user's expectations around a smart device's voice?	Physical cues: sound
Device	How do associations with a device's function (e.g. with traditional gender roles) affect expectations of a device's voice?	Associations with functionality
Device	Should the voice of a smart device change over time?	Long-term use
Device	If a device becomes physically damaged or worn, should that be reflected in the sound of the device's voice?	Long-term use
Context	For users who speak a language with grammatically gender, could the grammatical gender of the noun for a smart device influence the gender that users expect from its voice?	Linguistic cues
Context	As voice assistants like Siri and Alexa grow in popularity, will people come to expect that all voice-enabled devices take on the same, often female-sounding voice?	Longitudinal trends

Table 1. Summary of open research questions suggested by the User-Device-Context framework for smart device voices

likely to describe the same inanimate objects with adjectives consistent with their grammatical gender. Future work should investigate whether the same holds true for voice-enabled devices, and to what extent prior exposure to other voice assistants mediates this effect.

Longitudinal trends Another rich opportunity for future research is to explore *temporal patterns* in the voices that users expect from smart devices. A growing concern (e.g. see [76] and [14]) for why today's common female voice assistants are problematic is the potential for habituation: as voice assistants like Siri and Alexa grow in popularity, will people come to expect that *all* voice-enabled devices sound the same, and take on a female voice? Monitoring whether users' expectations trend towards uniformity may be an interesting research agenda over the long term.

4 DISCUSSION

So far, we have outlined the implications and specific directions for future work created by our *user-device-context* framework. The open questions associated with each frame are summarized in Table 1. Such a framework also suggests research and design opportunities which cut across these frames, some of which we outline below.

4.1 Can voices be deliberate persuasive design?

According to Fogg, Cuellar, and Danielson, persuasion can be defined as a "noncoercive attempt to change attitudes or behaviors" [24]. As the studies discussed in this paper have demonstrated, people naturally form social judgments on the basis of voices, both human and synthesized. In these studies, the end goal of understanding voice preferences was largely seen as an attempt to *match* user preferences and expectations. For example, Mutlu et al. suggested that "designers of interactive experiences should make sure that the interaction style of the robot fits the task structure and the individual attributes of users" [57]. Rather than seeking voice *fit*, we see the space of deliberate *misalignment* in expectations as an interesting area for future exploration. In other words, could voice instead be used as a means of persuasion?

One such recent exploration of this space comes from the *Intimate Futures* project. Taking a feminist HCI perspective, Søndergaard and Hansen [77] use design fiction as a means of exploring possible futures for voice assistants, particularly to challenge the status quo of feminized digital personal assistants. In motivating their contribution, the authors argue that the publicly articulated advertisements and narratives around voice assistants are problematic: "Intentionally or not, these objects are political entities that bring with them particular ethical and philosophical questions that we need to investigate also through design" [77]. With this framing in mind, they created two fictionalized voice assistants which they expressed through video prototypes. The first, "Aya," explores the interplay between gender and sexual harassment towards voice assistants; in a series of scenarios, AYA pushes back against problematic comments, using strategies like humor or overt threats to confront the user's sexual harassment. The second design fiction introduces "U," an assistant situated in the user's bathroom to support women's health. In the narrative with U, the fictionalized user discusses intimate topics such as menstruation and sexual activity, ultimately trusting the voice assistant to provide birth control advice. Reflecting on their design process, the authors share that they intentionally used a lower-pitched, potentially male-sounding voice for the voice assistant for women's health ("U") as a way of "troubli[ng] the dichotomy that connects male voices to professional work tasks and female voices to domestic, social, and personal tasks" [77]. Though Aya and U are still speculative in nature, their designs (if realized) could in many ways be seen as a form of persuasion; they both push the boundaries of how users typically interact with voice-enabled devices. The designers' explanation of how they selected a voice for U demonstrates how the voice identity in particular can play a role in persuasion by challenging expectations of what is conventional or comfortable.

To revisit the example of BMW's car interface mishap, as another speculative example, suppose that the company had decided *not* to issue a recall, and instead insisted that using a female voice was a deliberate design choice. What would have been the downstream consequences on users' perceptions of their car, and on the bias against accepting driving directions from women? On what timeframe would these attitude changes occur, if at all?

More generally, one open question for future work is to explore is whether giving devices features that *violate* stereotype expectations would affect users' beliefs and behaviors. For example, if an oven (which may hold stereotypical associations with female gender roles of housework) took on a male voice, would users come to view the device and associated tasks (baking) in a more gender-neutral way?

4.2 Going beyond the "once voice fits all" approach

Today's voice assistants are largely modeled after human speech: for instance, the personas of Siri, Alexa, and the Google Assistant are intended to sound as natural as possible. From both the technical and design perspective, however, synthesized speech need not mimic human speech.

Here, we share several ideas on how voice design might go beyond interaction metaphors drawn from human-human interaction.

Multiple voices Rather than having a one-to-one correspondence between smart devices and voices, designers may consider devices using multiple voices to suggest multiple personalities co-inhabiting the same device. With certain "multi-purpose" voice-enabled devices like the Amazon Echo or Google Home smart speaker system, this poly-voice model is already in place to some degree; if a user activates a third-party app on such devices, the app itself may speak with a different voice identity than that of the device's built-in assistant. Still other models may be possible: for instance, several voices could speak in unison (akin to the sound of a chorus) to suggest a consensus among different agents, or a device could use a different voice when the default agent fails, similar to how a manager might speak to a customer if a line-employee fails to help them.

Gendered or not? Gender has been a primary concern of both robot and voice assistant designers; often, a robot or voice assistant's voice is chosen such that its perceived gender matches with the intended persona of the device or task. For example, in designing Jibo, a social robot intended for a home environment, the designers explained their choice of a male voice as follows: "when you do robot design, you can try to make robots gender ambiguous, but it's fascinating that people pick up on cues and assign gender [...] Male voices when you talk about information tend to be held to be more credible" [41]. Similar justifications have been offered by corporations on the design of their respective voice assistant personas; according to the team behind Microsoft's Cortana assistant, user research revealed that "respondents worldwide preferred a female to a male assistant, ideally in her 20's, 30's tops — not one surveyed population expressed enthusiasm for a middle-aged assistant. She should be professional, but not a stiff; solicitous, but not a pest; cheeky, not biting" [19, 20].

As the research highlighted in this paper illustrates, decisions about whether a smart device reflects a gender through its voice and other characteristics has profound impacts on user interaction, and particular gender choices lead to stereotypes and other social implications [35]. Indeed, the voices of both social robots and voice assistants are often gendered as male or female by default [46], which reinforces a problematic notion that gender is a binary [33].

How might users respond to a smart device with a voice that is ambiguously gendered, or designed to be gender neutral? Researchers at *Project Q*, an initiative to bring a gender-neutral option to voice interfaces [1] suggest one approach to such design. Starting with "donated" voice recordings from more than two dozen individuals who identify as male, female, transgender, or non-binary, the team conducted a large-scale survey to select one voice which respondents found to be gender neutral [46]. Further research could study the effects of such voices on users, and how to synthesize them appropriately.

Human-like or robotic? While Nass and Brave consider speech "the fundamental means of *human* communication" [59] (our emphasis) the voices that smart devices take on need not emulate the characteristics we commonly associate with real human speech.

For example, voice interfaces may use non-speech sounds. Audio interfaces such as screen readers have a long history of using "earcons" – brief audio clips that act as audio icons and efficiently signal activity or state [9, 22, 25]. Future smart device interfaces might consider expanding the range of non-human-like sounds, to convey information or enhance the expressiveness or playfulness of communication.

Designers might also consider creating speech voices that are distinctly and deliberately *non-human*. In what cases might it be appropriate to use a voice that is clearly non-human or robot-like in quality? In 2018, a demo of Google's artificial intelligence technology, Duplex, was met with some controversy as the system deliberately mimicked human voices, including hesitations and other disfluencies like "umm" and "ahh", to make the listener believe they were talking to a human [38].

A distinctly non-human voice might mitigate this criticism by clarifying the identity of the speaker. Recent research from Moore [54, 55] and Aylett et al. [6] similarly argue for designing voices for robots and other smart devices sound deliberately non-human.

4.3 Findings from other fields

This paper considers the design of voices with literature and perspectives drawn primarily from the field of human-robot interaction. This perspective leads to a particular framework of analyzing the design of voices and future research questions. However, other literatures and perspectives are valuable as well.

4.3.1 Sociophonetic theory. A different perspective is based in sociophonetics, specifically how social factors affect the production of (human) speech, and how speech is used to draw social inferences. As voice assistants currently exhibit only a limited phonetic ability (e.g. accents and vocal quality), they lack the rich social and cultural identity inherent in human speech. Research in this vein suggests that social cues from voice add meaningful and systematic information over and above the linguistic content in speech, and that socio-psychological processing of speech leads to computers being seen as social actors. Similar to our own work, Sutton et al. [71] suggest that individualization and context of use affect how users are affected by voice. Our work complements and amplifies this research by suggesting how other factors such as embodiment or lack thereof affects voice design, and suggests ways to deliberately design voices for social persuasion. In addition, a human-robot interaction perspective also allows us to suggest research questions based on non-human-sounding voices.

4.3.2 Linguistic content and speech. While the aim of this paper was to underscore the importance of how a voice sounds on how a smart device is perceived, a user's experience of interacting with a voice-enabled device is ultimately a function of not just how the device sounds, but also what the agent says. In practice, the linguistic content and paralinguistic features of voice design are deeply intertwined [59]. According to prior work, users expect consistency between voice characteristics and linguistic content; people are more likely to trust and like a speech interface when the voice and content match in personality (e.g. both sounding like an extrovert and using phrases that an extrovert would use in conversation) [60]. With this in mind, it is important for voice designers to also consider what a voice-based agent says, and how well aligned it is with the voice identity.

Several studies have considered how linguistic content affects user perceptions and experience with conversational agents, and how linguistic and paralinguistic features of speech interact. Research in this area explored a wide range of communicative strategies such as politeness [17, 75, 80], vagueness [17], and apology or compensation strategies in response to breakdowns [40], among others. These studies have found linguistic content can affect user perceptions and experience in ways that largely mirror the user, device, and context frames discussed in this paper. For example, prior work has considered whether the dialect a robot spoke with (Modern Standard dialect of Arabic compared to a local dialect) affected perceptions of credibility, and how credibility was mediated by the agent's perceived knowledge of the domain [5]. Taken together, these studies offer helpful and complementary perspectives to the framework presented in this paper, and provide guidelines for designing dialogue scripts that are authentic and consistent with how the agent's voice sounds.

4.3.3 Technical considerations in speech synthesis. A major consideration in proposing new forms of voice design for smart devices is whether these novel types of speech are technically feasible. While a full review of the speech synthesis community is beyond the scope of this paper, many of the ideas and open questions posed already intersect with active areas of research in technical

communities such as INTERSPEECH. These communities have explored several themes relevant the ideas discussed in this paper, such as methods for synthesizing emotional voice content (e.g. adjusting pitch and intonation contours to suggest happy, sad, or neutral-sounding speech) [52] and identifying more robust measures for subjective evaluations of synthetic voices [83]. Other research efforts have focused on speech recognition, for example in modeling the prosody of a user's speech [73], and improving gender [27] and personality [4] recognition. These improvements may prove relevant particularly in modeling and designing with the "user-centric" frame in mind, like building voice assistants that are tailored to each user's personality, or even to in-the-moment fluctuations in mood. At present, the development frameworks for building voice assistants or skills (e.g. with the Alexa Skills Kit or Actions on Google) offer only a small set of pre-defined voices to choose from, with limited control over expressiveness through Speech Synthesis Markup Language (SSML) [64]. Through the framework introduced in this paper, we hope to inspire speech synthesis developers to make richer and more varied control over voices available to designers.

4.3.4 Ethnomethodologies of voice interaction. Finally, other research has considered the design of voice interfaces through an ethnomethodological perspective. Such studies have considered the everyday, naturalistic usage of common voice assistants like Alexa and Google, focusing on common use cases, elements of anthropomorphism, and use in family or other multi-user contexts, especially through conversational analysis [3, 8, 66, 68, 70]. Such a perspective suggests that "conversations" with voice devices are qualitatively distinct from conversations among humans. Even so, such a perspective still suggests that people consider voice-based devices to be social actors, with genders, identities, and agendas, and is largely consistent with the HRI perspective we explore in this paper.

5 CONCLUSION

The research reviewed in this paper demonstrates that the voice a smart device speaks with has profound social consequences for interaction. We propose a concrete framework that focuses on users, devices and contexts to better harness voice as an interaction technique. Moving forward, we hope this research moves voice interfaces away from the current one-voice-fits-all approach.

The research framework we articulate also naturally suggests areas for future research, some of which we have outlined in this paper. We hope that this framework creates powerful and currently under-explored opportunities to use voice in a way that deliberately shapes users' experiences with smart devices.

6 ACKNOWLEDGEMENTS

Our deepest thanks to Geoff Kaufman, Queenie Kravitz, and Samantha Reig for their encouragement and insights; this paper would not have been possible without their support. Thanks also to our reviewers for their time and helpful feedback. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

REFERENCES

[1] 2019. Meet Q. The First Genderless Voice. https://www.genderlessvoice.com

[2] Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology* 64, 3 (1993), 431.

[3] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (April 2019), 17:1–17:28. https://doi.org/10.1145/3311956

[4] Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. 2018. Deep Personality Recognition for Deception Detection. In *Proc. Interspeech 2018*. 421–425. https://doi.org/10.21437/Interspeech.2018-2269

[5] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. ACM Press, Portland, Oregon, USA, 157–164. https://doi.org/10.1145/2696454.2696464

[6] Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You Have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, alt08:1–alt08:10. https://doi.org/10.1145/3290607.3310422 event-place: Glasgow, Scotland Uk.

[7] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots And Racism. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 196–204. https://doi.org/10.1145/3171221.3171260 event-place: Chicago, IL, USA.

[8] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 243:1–243:13. https://doi.org/10.1145/3290605.3300473 event-place: Glasgow, Scotland Uk.

[9] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and Icons: Their Structure and Common Design Principles (Abstract Only). *SIGCHI Bull.* 21, 1 (Aug. 1989), 123–124. https://doi.org/10.1145/67880.1046599

[10] Dieter Bohn. 2019. Amazon says 100 million Alexa devices have been sold. https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp

[11] Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought* (2003), 61–79.

[12] Robin N. Brewer, Leah Findlater, Joseph 'Jofish' Kaye, Walter Lasecki, Cosmin Munteanu, and Astrid Weber. 2018. Accessible Voice Interfaces. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18)*. ACM, New York, NY, USA, 441–446. https://doi.org/10.1145/3272973.3273006 event-place: Jersey City, NJ, USA.

[13] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (2019), 11.

[14] Rebecca Cherng-Shiow Chang, Hsi-Peng Lu, and Peishan Yang. 2018. Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Computers in Human Behavior* 84 (July 2018), 194–210. https://doi.org/10.1016/j.chb.2018.02.025

[15] Brian X. Chen. 2019. *Devices That Will Invade Your Life in 2019 (and What's Overhyped)*.

[16] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. *arXiv preprint arXiv:1810.06828* (2018).

[17] Leigh Clark, Abdulmalik Ofemile, Svenja Adolphs, and Tom Rodden. 2016. A Multimodal Approach to Assessing User Experiences with Agent Helpers. *ACM Trans. Interact. Intell. Syst.* 6, 4 (Nov. 2016), 29:1–29:31. https://doi.org/10.1145/2983926

[18] Phil Cohen, Adam Cheyer, Eric Horvitz, Rana El Kaliouby, and Steve Whittaker. 2016. On the Future of Personal Assistants. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16* (2016), 1032–1037. https://doi.org/10.1145/2851581.2886425

[19] Dan Kedmey. 2015. Microsoft's Cortana Gets a Crash Course in Cultural Sensitivity | Time. *Time Magazine* (July 2015). http://time.com/3960670/windows-10-cortana/

[20] Andreea Danielescu and Gwen Christian. 2018. A Bot is Not a Polyglot. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–9. https://doi.org/10.1145/3170427.3174366

[21] Guy Deutscher. 2010. *Through the language glass: Why the world looks different in other languages*. Metropolitan Books.

[22] W. Keith Edwards and Elizabeth D. Mynatt. 1994. An Architecture for Transforming Graphical Interfaces. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology (UIST '94)*. ACM, New York, NY, USA, 39–47. https://doi.org/10.1145/192426.192443 event-place: Marina del Rey, California, USA.

[23] Kerstin Fischer, Katrin S Lohan, and Kilian Foth. 2012. Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 463–470.

[24] BJ Fogg, Gregory Cuellar, and David Danielson. 2019. Motivating, influencing, and persuading users: An introduction to captology. (2019).

[25] William W Gaver. 1989. The SonicFinder: An interface that uses auditory icons. *Human–Computer Interaction* 4, 1 (1989), 67–94.

[26] Erving Goffman. 1978. *The presentation of self in everyday life*. Harmondsworth London.

[27] Rajat Hebbar, Krishna Somandepalli, and Shrikanth Narayanan. 2018. Improving Gender Identification in Movie Audio Using Cross-Domain Data. In *Proc. Interspeech 2018*. 282–286. https://doi.org/10.21437/Interspeech.2018-1462

[28] Laura Hoffmann, Nikolai Bock, and Astrid M. Rosenthal v.d. Pütten. 2018. The Peculiarities of Robot Embodiment (EmCorp-Scale): Development, Validation and Initial Test of the Embodiment and Corporeality of Artificial Agents Scale. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 370–378. https://doi.org/10.1145/3171221.3171242 event-place: Chicago, IL, USA.

[29] Miwa Ikemiya and Daniela K. Rosner. 2014. Broken Probes: Toward the Design of Worn Media. *Personal Ubiquitous Comput.* 18, 3 (March 2014), 671–683. https://doi.org/10.1007/s00779-013-0690-y

[30] James Vincent. 2018. Google launches more realistic text-to-speech service powered by DeepMind's AI - The Verge. https://www.theverge.com/2018/3/27/17167200/google-ai-speech-tts-cloud-deepmind-wavenet

[31] Eun Hwa Jung, T. Franklin Waddell, and S. Shyam Sundar. 2016. Feminizing Robots: User Responses to Gender Cues on Robot Body and Screen. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. ACM Press, Santa Clara, California, USA, 3107–3113. https://doi.org/10.1145/2851581.2892428

[32] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the Design Space of Rendered Robot Faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*. ACM Press, Chicago, IL, USA, 96–104. https://doi.org/10.1145/3171221.3171286

[33] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357

[34] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[35] Sei Jin Ko, Charles M. Judd, and Irene V. Blair. 2006. What the Voice Reveals: Within- and Between-Category Stereotyping on the Basis of Voice. *Personality and Social Psychology Bulletin* 32, 6 (2006), 806–819. https://doi.org/10.1177/0146167206286627

[36] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. *Proceedings of the 2018 Designing Interactive Systems Conference* (2018), 881–894. https://doi.org/10.1145/3196709.3196784

[37] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 102:1–102:31. https://doi.org/10.1145/3274371

[38] Lauren Goode. 2018. How Google's Eerie Robot Phone Calls Hint at AI's Future. *Wired* (May 2018). https://www.wired.com/story/google-duplex-phone-calls-ai-future/

[39] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *Journal of Communication* 56, 4 (2006), 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x

[40] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 203–210. https://doi.org/10.1109/HRI.2010.5453195

[41] Lily Hay Newman. 2014. This Social Robot Is Adorable. But Will Families Actually Want One? *Slate* (July 2014). https://slate.com/technology/2014/07/social-robotics-expert-cynthia-breazeal-debuts-jibo-a-family-robot.html

[42] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 255–262.

[43] Nichola Lubold, Erin Walker, Heather Pon-Barry, and Amy Ogan. 2018. Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics. In *International Conference on Artificial Intelligence in Education*. Springer, 282–296.

[44] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 5286–5297. https://doi.org/10.1145/2858036.2858288

[45] Nikolas Martelaro and Wendy Ju. 2017. WoZ Way: Enabling Real-time Remote Interaction Prototyping & Observation in On-road Vehicles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, Portland, Oregon, USA, 169–182. https://doi.org/10.1145/2998181.2998293

[46] Matt Simon. 2019. The Genderless Digital Voice the World Needs Right Now. https://www.wired.com/story/the-genderless-digital-voice-the-world-needs-right-now/

[47] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices. *PLoS ONE* 9, 3 (March 2014), e90779. https://doi.org/10.1371/journal.pone.0090779

[48] C. McGinn and I. Torre. 2019. Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 211–221.

https://doi.org/10.1109/HRI.2019.8673305

[49] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335 event-place: Portland, Oregon, USA.

[50] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. [n. d.]. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2019 - DIS '19*.

[51] Rani Molla. 2018. Voice tech like Alexa and Siri hasn't found its true calling yet: Inside the voice assistant 'revolution'. *Recode* (2018). https://www.recode.net/2018/11/12/17765390/voice-alexa-siri-assistant-amazon-echo-google-assistant

[52] Juan Manuel Montero, Juana M Gutierrez-Arriola, Sira Palazuelos, Emilia Enriquez, Santiago Aguilera, and José Manuel Pardo. 1998. Emotional speech synthesis: From speech database to TTS. In *Fifth International Conference on Spoken Language Processing*.

[53] Dylan Moore, Hamish Tennent, Nikolas Martelaro, and Wendy Ju. 2017. Making noise intentional: A study of servo sound perception. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 12–21.

[54] Roger K Moore. 2017. Appropriate Voices for Artefacts: Some Key Insights. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*.

[55] Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Kristiina Jokinen and Graham Wilcock (Eds.). Springer Singapore, Singapore, 281–291. https://doi.org/10.1007/978-981-10-2585-3_22

[56] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design Guidelines for Hands-free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. ACM, New York, NY, USA, 269–276. https://doi.org/10.1145/3236112.3236149

[57] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica Hodgins, and Sara Kiesler. 2006. Task Structure and User Attributes as Elements of Human-Robot Interaction Design. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Univ. of Hertfordshire, Hatfield, UK, 74–79. https://doi.org/10.1109/ROMAN.2006.314397

[58] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 6:1–6:7. https://doi.org/10.1145/3173574.3173580 event-place: Montreal QC, Canada.

[59] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press.

[60] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[61] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703 event-place: Boston, Massachusetts, USA.

[62] Kenneth Olmstead. 2017. *Nearly half of Americans use digital voice assistants, mostly on their smartphones*. Technical Report. Pew Research Center. https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

[63] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[64] Sarah Perez. 2018. Alexa developers get 8 free voices to use in skills, courtesy of Amazon Polly. *TechCrunch* (May 2018). https://techcrunch.com/2018/05/16/alexa-developers-get-8-free-voices-to-use-in-skills-courtesy-of-amazon-polly/

[65] Martha L Picariello, Danna N Greenberg, and David B Pillemer. 1990. Children's sex-related stereotyping of colors. *Child Development* 61, 5 (1990), 1453–1460.

[66] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–12. https://doi.org/doi.org/10.1145/3173574.3174214

[67] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

event-place: Portland, Oregon, USA.

[68] A Purington, J G Taft, S Sannon, N N Bazarova, and S H Taylor. 2017. "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. *Conference on Human Factors in Computing Systems - Proceedings* Part F1276 (2017), 2853–2859. https://doi.org/10.1145/3027063.3053246

[69] Sara Perez. 2019. Report: Voice assistants in use to triple to 8 billion by 2023 | TechCrunch. https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/

[70] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18* (2018), 857–868. https://doi.org/10.1145/3196709.3196772

[71] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 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 - CHI '19*. 1–14. https://doi.org/10.1145/3290605.3300833

[72] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65. https://doi.org/10.1145/348941.348990

[73] Berrak Sisman and Haizhou Li. 2018. Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion. In *Proc. Interspeech 2018*. 52–56. https://doi.org/10.21437/Interspeech.2018-1499

[74] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and Correcting Less Accessible Content in Voice Interfaces. (2018), 1–13. https://doi.org/10.1145/3173574.3173870

[75] Vasant Srinivasan and Leila Takayama. 2016. Help Me Please: Robot Politeness Strategies for Soliciting Help From Humans. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4945–4955. https://doi.org/10.1145/2858036.2858217 event-place: San Jose, California, USA.

[76] Chandra Steele. 2018. The Real Reason Voice Assistants Are Female (and Why it Matters). *PCMag* (2018).

[77] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18* (2018), 869–880. https://doi.org/10.1145/3196709.3196766

[78] Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* 38 (Sept. 2014), 75–84. https://doi.org/10.1016/j.chb.2014.05.014

[79] Hamish Tennent, Dylan Moore, Malte Jung, and Wendy Ju. 2017. Good vibrations: How consequential sounds affect perception of robotic arms. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 928–935.

[80] Cristen Torrey, Susan Fussell, and Sara Kiesler. 2013. How a Robot Should Give Advice. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13)*. IEEE Press, Piscataway, NJ, USA, 275–282. http://dl.acm.org/citation.cfm?id=2447556.2447666 event-place: Tokyo, Japan.

[81] James Vincent. 2019. Kohler's smart toilet promises a 'fully-immersive experience'. *The Verge* (2019). https://www.theverge.com/2019/1/6/18170575/kohler-konnect-bathroom-smart-gadgets-numi-intelligent-toilet-ces-2019

[82] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education*. Technical Report. UNESCO, EQUALS Skills Coalition. https://unesdoc.unesco.org/ark:/48223/pf0000367416.locale=en

[83] Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are We Using Enough Listeners? No!—An Empirically-Supported Critique of Interspeech 2014 TTS Evaluations. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Received April 2019; revised June 2019; accepted August 2019