Social Media through Voice: Synthesized Voice Qualities and Self-presentation

LOTUS ZHANG, Human Centered Design and Engineering, University of Washington, USA
LUCY JIANG, Computer Science and Engineering, University of Washington, USA
NICOLE WASHINGTON, Human Centered Design and Engineering, University of Washington, USA
AUGUSTINA AO LIU, Human Centered Design and Engineering, University of Washington, USA
JINGYAO SHAO, Human Centered Design and Engineering, University of Washington, USA
ADAM FOURNEY, Microsoft Research AI, Microsoft Research, USA
MEREDITH RINGEL MORRIS, Microsoft Research, USA
LEAH FINDLATER, Human Centered Design and Engineering, University of Washington, USA

With advances in expressive speech synthesis and conversational understanding, an ever-increasing amount of digital content—including social and personal content—can be consumed through voice. Voice has long been known to convey personal characteristics and emotional states, both of which are prominent aspects of social media. Yet, no study has investigated voice design requirements for social media platforms. We interviewed 15 active social media users about their preferences on using synthesized voices to represent their profiles. Our findings show that participants want to have control over how a voice delivers their content, such as the personality and emotion with which the voice speaks, because these prosodic variations can impact users' online personas and interfere with impression management. We report motivations behind customizing or not customizing voice characteristics in different scenarios, and uncover key challenges around usability and the potential for stereotyping. We argue that synthesized speech for social media should be evaluated not only on listening experience and voice quality but also on its expressivity, degree of customizability, and ability to adapt to contexts (e.g., social media platforms, groups, individual posts). We discuss how our contribution confirms and extends knowledge of voice technology design and online self-presentation, and offer design considerations for voice personalization related to social interactions.

CCS Concepts: • **Human-centered computing** → **Sound-based input / output**; **Empirical studies in HCI**; **Social content sharing**; • **Social and professional topics** → *User characteristics.*

Additional Key Words and Phrases: voice interaction; voice synthesis; social media; online self-presentation

ACM Reference Format:
Lotus Zhang, Lucy Jiang, Nicole Washington, Augustina Ao Liu, Jingyao Shao, Adam Fourney, Meredith Ringel Morris, and Leah Findlater. 2021. Social Media through Voice: Synthesized Voice Qualities and Self-presentation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 161 (April 2021), 21 pages. https://doi.org/10.1145/3449235

Authors' addresses: Lotus Zhang, Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA; Lucy Jiang, Computer Science and Engineering, University of Washington, Seattle, Washington, USA; Nicole Washington, Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA; Augustina Ao Liu, Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA; Jingyao Shao, Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA; Adam Fourney, Microsoft Research AI, Microsoft Research, Redmond, Washington, USA; Meredith Ringel Morris, Microsoft Research, Redmond, Washington, USA; Leah Findlater, Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA.

2573-0142/2021/4-ART161 $15.00
https://doi.org/10.1145/3449235

1 INTRODUCTION

Voice interfaces have proliferated in recent years, enabling hands-free, intuitive, and accessible interactions [53]. Along with the widespread adoption of voice assistants such as Amazon's Alexa, Apple's Siri, Microsoft's Cortana, and Google's Assistant [55], an increasing number of web and mobile applications have started to support voice interaction [16, 50, 54].

In parallel, researchers have begun to move beyond the mechanics of speech recognition, speech synthesis, and conversational understanding to examine expressive characteristics of synthesized voices (e.g., [6, 8, 14, 29]). Increasing attention has shifted from *what* the voice should speak to *how* it should speak [7, 8]. This research has focused on users' preferences for synthesized voices when interacting with a voice agent or consuming content (e.g., listening to audio books or turn-by-turn navigation directions). For example, many studies suggest that listeners prefer voices that display characteristics similar or complementary to their own, such as personality and gender (e.g., [8, 35]).

In this paper, we instead explore a related but different question: how do individuals want their *own* content to sound? For social media sites in particular, a sense of individuality is critical [5, 33]. Unlike visual styling such as fonts, colors, or images, social media users currently have little to no control over how their content sounds when it is synthesized for voice interaction. Instead, the sound is dictated by settings on screen reader applications or voice assistants, where the default voices often lack expressiveness [14] and diversity [8]. With voice assistants beginning to provide audio-based access to general web content [40], we imagine that not only will website and voice application developers soon be able to choose from a range of synthesized voices, but end users themselves will be able to customize voices for their own content—perhaps in the form of a general "profile voice" or even more content-specific styling.

To investigate how social media users would want their own content to sound and how they would envision listeners responding to that sound, we conducted a semi-structured interview study with 15 participants. All participants were regular social media users with at least some voice interface experience. While most were younger adults, they represented a range of genders and ethnicities. To encourage participants to consider a wide range of expressive synthesized voice characteristics, we prompted each participant with audio clips of their own social media profile spoken by voices that varied in perceived gender, age, accent, and emotion. Participants were asked about how they felt about the presented voices in speaking their content and what considerations they would have in creating an ideal personalized voice if desired.

Our findings uncovered criteria for synthesized voices to properly present social media content, including expressivity, customizability, and context appropriateness. Participants generally wanted an authentic and consistent synthesized voice presentation of themselves, but also desired the ability to reflect the emotion of specific posts and to opt for more "fun" or more "formal" voices for different platforms (e.g., Twitter vs. LinkedIn). We also identified important challenges with voice personalization, perhaps most critically the possibility of perpetuating stereotypes through accented synthetic speech.

This paper makes the following contributions: (i) characterization of preferences for how to present one's own socially oriented content (as opposed to consuming voice-based content); (ii) empirical evidence for and interpretation of how synthesized voice preferences relate to online impression management theory; (iii) enumeration of technical, usability, and ethical challenges and design considerations for future work on socially situated, self-customized voice synthesis systems.

2 RELATED WORK

Our work is informed by literature on synthesized voice design, sociophonetics, and online presentation of self. We also provide a quick review of existing voice-based social media platforms.

2.1 Synthesized Voice Choice

What makes a good voice for a speech interface? Speech synthesis research has examined human perception of varying voice choices, discovering that listeners tend to be attracted to voices with similar "personal" characteristics as their own [35]. In an experiment where participants were asked to listen to book descriptions in synthesized voices, Lee et al. [35] found that a user's sense of social presence improved when the voice's personality matched the user's personality. Similarly, Braun et al. [6] reported that personalizing the voice assistant's personality for each user resulted in higher rates of likability and trust than a default personality. People also seemed to rate the speech more positively when the perceived voice gender [34] and accent [41, 67] matched their own or the content being spoken.

Besides the tendency to favor similar voices, early findings also indicated that natural human speech is preferable to and more understandable than computer-generated speech [22, 46, 63]. However, as the quality of speech synthesis improves, more recent studies have challenged those conclusions. In two experiments, Stern et al. [64] found that listeners do not prefer natural speech over synthesized speech when they see that the source of speech is a computer rather than a human. In fact, the increased level of humanness among smart speaker voices seems to introduce unrealistic expectations of these devices' intellectual and emotional capabilities, juxtaposing their limitations [12, 37].

Recently, speech synthesis researchers have begun to suggest that the lack of *qualitative understanding* in user needs for *specific application contexts* is one of the biggest open challenges of text-to-speech evaluation [76]. Within the Computer-Supported Cooperative Work community, Cambre et al. [8] proposed a research framework theorizing that voice design requirements for smart devices are shaped by and vary across users, devices, and contexts. To date, most voice evaluation work has focused on general use cases of smart speakers and voice assistants, and the most commonly used measures have been related to listening experience [76]. For example, voices for long-form texts (e.g., audio books) are evaluated primarily based on intelligibility, comprehensibility, and other subjective measures, such as likeability [7, 76].

While the above work focuses on how users perceive voices when *consuming* voice content, user preferences for voices that *present their content* and the social implications of those preferences have received much less attention.

In the field of accessibility, research on augmentative and alternative communication (AAC) devices noted the challenges of using synthesized voices for self-expression due to their lack of expressivity. The challenges center on supporting users' conversational pacing, personality expression, and identity presentation, which limit their ability to authentically express themselves [31, 57]. In recent years, advances in speech synthesis have begun to address technical issues related to expressivity (e.g., [45]), with many commercial text-to-speech engines producing more humanistic and emotion-rich voices (e.g., IBM Watson [78], MaryTTS [39]). Focusing on AAC output, Fiannaca et al. [19] developed two interfaces to adjust expressive speech parameters such as choosing and refining voice emotions. HCI researchers have also looked into different ways to render emoticons with voice to augment the social experience of voice interaction [27]. Finally, the emergence of voice font technology [23, 43, 74], which leverages machine learning to create a synthesized voice through a recording of the user's own voice, makes more diverse voice options possible. Combined, these advances provide support for voice to enter mainstream social-interaction-heavy applications, such as social media. Yet, research on synthesized voice choices for social media is sparse. Recognizing this gap, we seek to investigate voice design requirements for presenting personal content on social media.

2.2 Sociophonetic Considerations of Voice

While voice design for social media content has not garnered much attention, voice is known to influence social experiences in everyday life. Combining sociolinguistics and phonetics, the field of sociophonetics examines this phenomenon, including how social information is perceived through phonetic details [15]. Along with the spoken content, social categories and personality traits can be extracted from auditory input in a fairly consistent way [15]. People are capable of estimating a speaker's age [61], ethnicity [58, 69], gender and sexuality [49, 65], and socioeconomic class [61] from speech alone. A series of studies also suggests that speaking rate, amplitude, and accent of speech impact perception of a speaker's personality traits [1, 2, 68, 70]. Some argue that people's perception of this social information, especially gender, is along a spectrum rather than in rigid categories [48]. At the same time, listeners' expectations regarding the speaker also influence how they perceive social information from speech [51].

Recently, Human-Computer Interaction (HCI) literature has begun incorporating sociophonetic findings into voice technology research. For example, Sutton et al. [66] proposed three design strategies rooted in sociophonetics for more inclusive and natural voice user interfaces: individualism, context awareness, and diversification. The focal point of these strategies has been on how voice should be designed for consuming information. Our study instead explores these considerations with respect to how users want their *own* content to sound. With synthesized voices being highly malleable, social media users have the opportunity to hide or reveal aspects of their identity that are usually tied to their real voices—a question to investigate is how people consider employing this option to present themselves online.

2.3 Online Presentation of Self

How people want others to perceive themselves on social media has long been studied in HCI, CSCW, and the sociology of technology and science. Most social media activities involve self-presentation—making social connections, sharing identities, and updating statuses [30]. Past work has looked into social media users' decision-making around identity presentation [5, 18, 33], content management [82], privacy control [59], and audience management [38], often drawing on Goffman's theatrical metaphor [20] and Hogan's exhibition approach [26].

Goffman's dramaturgical approach uses stage and performance as metaphors to describe how a person's self-presentation tends to be selective across different contexts [20]. The notion of the "front stage", as opposed to the private "back stage", describes scenarios where a performance is given in the presence of an audience. On social media, similar to offline situations, people have a set of audiences[38]. However, unlike how physical interactions usually have specific audiences, social media interactions "collapse multiple audiences into single contexts" [38] and display users' content in a form rather similar to an exhibition [26]. Users typically engage in audience management techniques to cope with presentation challenges within these "exhibitions" [4, 38]. For example, many people use multiple accounts [38], only post things that are non-offensive to the broadest group of audiences (known as the lowest common denominator effect) [26], and strategically conceal information from different audience groups [38]. Engaging in these techniques often heightens the difficulty of balancing personal authenticity and audience expectations [4, 38]. Mainstream social media often endorses positive self-presentation, which may fuel negative social comparison among users at times and further restrict an individual from freely expressing themselves [9, 17, 28, 79]. Xiao et al. [80] found that by using fake accounts, or "finstas", with only close friends, users can present themselves more authentically and disregard social pressures from the social comparison that is rampant on social media. Zhao et al. [82] also found that people create social media content not only for others, but for themselves, to create an "archive" of the meaningful parts of their life.

Dramaturgical analysis for offline scenarios often involves nonverbal signals [13, 20], but online presentation of self has primarily focused on how users select their textual or graphical content. There is little known about whether people apply impression management strategies when choosing synthesized voices to represent themselves. Does the emphasized difficulty in navigating impression management through collapsed contexts introduce special challenges for voice design and create a space for exploring voice-customization? If so, how do users propose changing their synthesized voice presentations to cope with these challenges? We explore these questions in our study.

2.4 Voice-based Social Media

Many social media sites are starting to support voice-based content delivery. Some recording-based voice forums also exist where users can record and listen to voice messages. A set of research-based voice forums including Sangeet Swara [72, 73], Baang [73], and Gurgan Idol [32] are targeted to users with low literacy and socioeconomic barriers to accessing online information. These services are accessed via toll-free phone calls in local languages without requiring internet connectivity, and research has focused on usability, financial sustainability, integrity, and equality. Recording-based auditory social platforms that target a broader set of users also exist, such as Clubhouse [11], Shootwords [62], Audlist [52], and HearMeOut [25]. These sites advertise that integrating audio interaction can make online connections more authentic, engaging, and convenient. However, the fact that these services only support recording-based interactions can be problematic, especially for users who cannot record their own voice, prefer to compose content using text-based methods, or want to create content that can be consumed in multiple modalities. Privacy issues related to voice recordings are another concern [24, 81]. Therefore, our study mainly focuses on exploring how current social media users would choose to present themselves online using a synthesized voice.

3 METHOD

To explore how social media users react to the idea of using customized synthesized voices to present their content, we conducted a semi-structured interview study with 15 participants. During the study, we prompted participants to envision a range of synthesized voice characteristics by playing audio examples, some of which included the participant's own social media content. Questions covered participants' reactions to the overall idea of using a profile voice, the voice characteristics presented, and preferences for customized or default voices in a variety of contexts. This study was approved by the Institutional Review Board at our university.

Throughout Sections 3 and 4, we use *bio information* to refer to the introductory descriptions that social media users provide about themselves on a profile page, and *profile* as an overarching term that includes both this bio information and *posts*.

3.1 Participants

Fifteen active social media users with some exposure to voice interaction technology participated in our study. For diversity, we broadly recruited on four social media platforms (Facebook, Instagram, Reddit, Twitter) and through word of mouth. Participants were screened on the following criteria: they must i) have a Facebook, Instagram, LinkedIn, and/or Twitter account with at least two posts and one written paragraph of bio information; ii) use social media platform(s) at least once a week; iii) have posted at least once within the last month; vi) have some experience with smart speakers or voice-based interaction such as Amazon Echo, Apple Siri, or Google Home. Participants ranged in age from 19 to 47. These details, along with self-reported gender, ethnicity, and primary social media platform, are shown in Table 1. All participants were volunteers and were compensated for their time with a $20 gift card.

Table 1. Participant demographics and initial voice preferences among a set of five voices that varied only in terms of perceived gender and age. We list gender, age, and ethnicity in participants' own words (e.g., "female", "woman").

ID	Gender	Age	Ethnicity	Primary Social Media Platform	Preferred Voice
1	Female	20	African American, Caucasian	Instagram	Older masculine
2	Female	20-something	Black, white	Twitter	Gender neutral
3	Female	30	White	Facebook	Younger feminine
4	Genderqueer	20	White	Instagram	Younger masculine
5	Female	19	Black	Instagram	Younger feminine
6	Male	25	Mixed race, but mostly white	Instagram	Gender neutral
7	Female	47	African American	Facebook	Older feminine
8	Male	20	African American	Twitter	Younger masculine
9	Female	20	Asian	Twitter	Younger feminine
10	Woman	20	Caucasian	Facebook	Younger feminine
11	Male	23	Half black, half Filipino	Instagram	Younger masculine
12	Female	21	Asian	Instagram	Younger feminine
13	Female	20	Asian, Caucasian	Instagram	Younger feminine
14	Female	21	White	Instagram	Older masculine
15	Male	32	African American	Instagram	Older feminine

3.2 Procedure

The interviews were designed to take up to 60 minutes and were conducted remotely via video conferencing software. Beforehand, participants shared their bio information and two example posts from their primary social media account. For consistency, the same research team member conducted all interviews, each of which consisted of four parts:

First, participants were asked about their demographic information, language proficiency, technology and social media usage, and experience with voice technologies.

Second, we elicited reactions to the general idea of interacting with social media content via voice, introducing it through the following scenario: *"...imagine that you and others can listen to and interact with social media through voice-based technologies like smart speakers. Think about what you would want your bio information and posts to sound like, and what other people might think when they listen to that information."* To provide concrete examples and encourage interviewees to envision a range of possible voice choices, we played a portion of their own social media profile information using five example voices. The samples were selected to represent a range of perceived ages and genders: younger and older male and female voices, and a vocally androgynous (gender

neutral) voice. These five voices were generated with IBM Waston [77] and Natural Reader [36]. We controlled the length of all audio clips to be roughly 10 seconds, which is adequate time to read typical update-type posts (e.g., 240 characters on Twitter). All emoticons were converted to spoken text through the site Unicode Common Locale Data Repository [71].

Participants listened to their profiles read out by the five voice samples (randomly ordered per participant) without mentioning the intended differences in gender and age. After each clip, participants were asked to describe the voice, rate their level of agreement on a 5-point Likert scale with the statement, "I would choose to use this voice for my social media profile," and provide a rationale for their rating (we focus on the qualitative rationale in Section 4, but the rating data is included in supplementary materials). We also asked participants what factors they considered when evaluating the voice samples and which of their identities (e.g., gender, ethnicity) they would be comfortable with representing through a synthesized voice.

The third part of the study focused on more advanced synthesized voice qualities. Here, we probed reactions to possible advanced qualities by playing two sets of example voices that demonstrated different accents (Indian English and British English) and different emotions (sad, angry, happy, and scared), all of which were generated with Voicery [75]. The accented samples said *"the quick brown fox jumped over the lazy dog,"* whereas the emotional samples said the following with the words changing to match the emotion: *"That was not just a [bad/good] dream. It made me [sad/angry/scared/happy]."* We explained that these examples were meant to encourage participants to think about advanced voice qualities, and that synthesis of these qualities would likely improve in the coming years. After discussing the two sets of clips, we asked for participants' initial reactions and whether they felt that being able to vary accents and/or emotions would be useful or not for their own social media voice profiles. Interviewees were then asked more generally what characteristics of synthesized speech, if any, might be useful for social media users to customize how their profiles sound, and what voice they would ideally want for their own profile and why.

Lastly, the interview closed with questions to contrast default voices (e.g., the voices of Siri or Alexa) versus customized voices for social media profiles, and whether voice preferences would change in different contexts (e.g., for different platforms or types of content).

3.3 Data Analysis

We performed thematic analysis on our interview transcripts, focusing on understanding voice characteristics important to participants' self-presentation and how the role of context influences profile voice choices. To start the analysis, the first author read through all of the transcripts to obtain a global view of the data, and derived a list of initial codes that included both deductive and inductive codes. The first two authors then each independently coded all transcripts, meeting to share coded transcripts and memos, to resolve conflicts, and to iterate on the structure of the codebook every two to three participants. After the first round of coding, we identified five main themes: general responses, feedback on accented voices, feedback on emotional voices, factors considered when evaluating a voice, and when and how to personalize profile voices (codebook shared as supplementary material). The codebook was again reviewed, discussed, and regrouped by the first two authors, which led to the final themes we present in this paper.

4 FINDINGS

Here, we present voice considerations unique to self-presentation on social media.

4.1 Voice Choice by Content Producers

Past discussions around synthesized voice choice predominately focused on listeners' experiences. Yet, a significant part of online content, and almost all of social media, is generated by users themselves. What voices do users think should be used to speak their own content?

When asked to evaluate synthesized voice options for their own social media content, participants mentioned a variety of factors that roughly fall into two categories: how the voice *represents* characteristics of themselves, and the overall *quality* of the voice. Table 2 lists the full set of factors, including those we explicitly mentioned in interview questions (*gender*, *age*, *accent* and *emotion*) as well as others that spontaneously emerged during the interviews.

Table 2. Factors impacting voice personalization preference.

Category	Factor
Representation	Personality
	Gender
	Pitch
	Age
	Accent
	Emotion (of content)
Overall Quality	Naturalness
	Accuracy and Clarity
	Timbre
Other	Basic voice characteristics (speed, cadence, volume)
	Non-text sounds (e.g., music, animal sounds)

While many factors arose, participants weighed them differently. Perhaps most critical was the capability to naturally represent who they are and accurately reflect the meaning of their messages. Below, we present why and how respondents thought certain aspects of synthesized voices were important to the presentation of their content.

4.1.1 Wanting a representative voice. All interviewees expressed the desire to have a representative synthesized voice. For example, P4 said, *"I'm reading these posts [their own social media posts] in my head essentially like I'm hearing my own voice. It would be strange if someone else is there."* In describing how voices would or would not represent them, participants mentioned primarily persistent personal characteristics (e.g., *gender*, *personality*) but also the *emotion* of their content (e.g., a happy or sad post). From Table 1, we can see that the gender and age of the preferred voice tended to match the participant's gender and age, with some exceptions.

Personality: While reviewing the voice samples, almost all interviewees commented on whether or not the perceived personality of a voice matched their social media persona ($N = 14$). Examples of a disconnect in personality included that the synthesized voice was not *"positive"* (P1), *"quirky"* (P3), *"confident"* (P5), or *"friendly"* (P15) enough. Authentically representing their personality was important to some participants ($N = 5$), as one participant noted: *"I don't want it to sound like something that I'm not, because I'm not trying to fool anybody, or I'm not trying to have a different persona"* (P11). Others were more concerned about matching the personality of the voice to their online persona but not necessarily as an authentic representation of themselves ($N = 6$). For example, P12 said: *"Because, as I mentioned, [my post is] very positive. But my personality is... I have*

up[s] and I have down[s], but I only show the good side, which... it's not my personality, it's just part of it.". Participants overall tended to prefer presenting a positive self-image on social media. Many found it strange when their posts were spoken in ways that did not correspond with this level of excitement and positivity: *"All of my language or posts are very, I would say, energetic or positive. 'Oh, I finished this skydive!' or 'Oh, I finished the internship, 'so exciting!' but he sounds so tired, that makes me feel like, is he laughing at my post?"* (P12).

Gender and Pitch: As shown in Table 1, nearly all interviewees wanted their profile voice to sound like the gender they identify with, but to differing extents. Many explicitly did not want to use a voice that they felt clearly differed from their own gender ($N = 8$):

> "Even though this voice is the best [quality] voice out of all the ones that I have heard so far, it's still a female, and the delivery of the words and the way the female is speaking is better, but I still, for my profile, I would want it to represent me. I'm a male, and I don't talk in that manner." (P15)

However, people have differing opinions on how important gender is as part of their identity. Two respondents were willing to prioritize voice quality over having a gender-match with the voice, as P1 commented:

> "I think that [gender] did a lot with how I felt it identified with me. But then once I listened to the last voice then I didn't, like, care as much about [gender]. So then I was no longer taking gender into account cause I just, like, prefer that voice." (P1)

Pitch is one aspect of voice that is relevant to gender. About half of our participants mentioned wanting to have a pitch that represents the gender they identify with. Nevertheless, pitch can vary within a gender category. Our participants appreciated when the voice not only matched their gender identity, but also represented who they are with respect to the average voice within that category. For example, P2 said, *"It [gender neutral voice] did sound deeper than number four [younger female voice], which I as a female have a deeper voice than average. So that's something I appreciated because I identify with it".* In this vein, a few people were unhappy with how masculine or feminine the gendered voices sounded. For example, when asked to describe his ideal voice, P6 hoped for more neutral options to better represent himself on the gender spectrum: *"...something that is male, but... not extremely male I guess".* This suggests that a spectrum of pitches would be more useful than a small, pre-defined set.

Age: While not considered as important as personality and gender, age was also deemed to be a part of our participants' identities. When a voice sounded too different from their age, participants found it to interfere with their online persona, such as: *"I imagine that I don't sound like an old woman. Uh yeah, like I said before, it's just not consistent with how I imagine my voice"* (P4). However, more minor differences in age were not a concern to most participants: *"I didn't take in my age per se. I think it comes through when you try to fit it to your personality or who you show yourself to be, but I didn't focus on my age as a reason to pick one over the other"* (P9). Note that our participants were primarily younger adults, so these findings may differ for older users.

Accent: Participants who consider accent to be a part of their identity felt that incorporating accents into their voice presentations would help support their self-expression, using P10's words:

> "Well, I would feel more comfortable and confident in having the way that I speak represented, especially in media and being able to have a voice that actually matches the way that I speak, rather than just one that matches the way everyone else speaks." (P10)

Emotion: Besides the above identity-related factors, we probed for interviewees' responses to synthesized voice emotions by playing four voice samples designed to sound sad, angry, happy,

and scared. All respondents felt that being able to change voice emotion to match specific social media posts could help in delivering meaning. As P13 said, *"I think emotions are what's lacking in text and social media, and to be able to convey that in a more auditory format could add an extra layer to what people are missing in their social media"*. P11 also expressed enthusiasm: *"I want people to hear how I see things. I think this changes the game a lot."*

Participants also often called out when the voice's emotion or tone did not match with the post content. For example, one participant specifically commented on how distracting the mismatch between voice and content could be after listening to a voice that he perceived as emotionless: *"They can see that in the post, I had a good time, but when you hear it that way, I feel like it takes away from the picture, and it makes you focus on that voice"* (P15). As another example, when asked about voice choices for Instagram, P15 said, *"If it was an exciting moment, I would want my voice to be more exciting and maybe have some emphasis on certain words. If it's sad, I'd rather it be neutral, just depending on the picture."*

Despite the importance of emotion customization, there were also concerns around simulating and presenting emotion with algorithms:

> "It's very subjective, emotional. Sometimes when people are posting really sad stuff, maybe people find it funny or... I don't want the voice assistant to judge whether this is a sad post or a happy post. That's one gray area that I'm not sure about." (P12)

Beyond the emotions we presented, some participants suggested additional emotions that would be useful to represent. For example, four participants mentioned sarcasm: *"I think it would be super helpful, especially if you could get across the sarcastic tone. That would save a lot of people a lot of trouble because... sarcasm does not... come across well through texts"* (P2).

4.1.2 Wanting a good-quality voice. Voice quality is critical to the listening experience and understanding of content, and thus has always been an important measure in evaluating speech synthesis [7]. When interviewees evaluated voices for their own social media content, we observed other voice-quality related considerations besides comprehension and listening pleasure—the key point is that an unnatural, unclear voice may not sound representative regardless of what other characteristics can be manipulated. Among voice quality concerns, participants most commonly emphasized naturalness, accuracy, and clarity.

Naturalness: The extent to which a synthesized voice should be human-like varies based on context of use [12]. Almost all participants ($N = 13$) preferred a natural-sounding voice—this was a characteristic that we did not explicitly ask about, emphasizing the importance of naturalness for voice-based social media delivery. Indeed, the majority of participants said that naturalness was the *most* important characteristic of a good voice to them, sometimes even beyond representativeness, with P8 saying, *"I think out of all of the voices it [younger masculine voice] checked the most boxes that mattered to me, and the ones that I didn't check, they were less important, because I think it sounded the most life-like. I think that's probably the most important."* Participants did not like voices that sounded too flat, out of concern that the voice would not accurately convey the content's meaning and would impede the user's ability to represent themselves, such as: *"Doesn't sound right. The way the voice sounds reading my bio and my post, it just sounds bland, and I feel like it takes away from what I want people to focus on. The voice is still a little robotic, and it doesn't represent who I am, as a person"* (P15). Combined, participants described a natural-sounding voice as sounding like a human, with appropriate inflections, a smooth flow of words, and a natural tone (e.g., P11: *"I liked the way it had different inflections and different tones."*)

Accuracy and clarity: More than half of our participants chose voices based on whether the voices spoke accurately and clearly ($N = 8$). For example, one participant was worried about the voice being too fast for listeners to understand, which relates to clarity: *"That was definitely really*

fast. If I was trying to understand something from having it spoken to me, I would want that to be a little slower" (P2). Besides the clarity of the speech, interviewees also shared concerns about whether the voice would pronounce words properly. For example, four people were specifically worried about names or terms on their profile being mispronounced: *"It said, 'My dog's name is Milan', and it said 'Mulan' or something. It just was saying certain words wrong, I noticed that. I guess... I don't know if it was the punctuation, but the flow of the sentence wasn't right"* (P15).

4.1.3 Other voice considerations. Additionally, other less common suggestions included being able to control voice speed, cadence, and volume, and to support non-text elements, such as animal sounds, unnatural voice distortions, and signature background music, as recorded in Table 2. Together, they reflect social media users' interests in expressing themselves in creative ways and their desire to have more control over their online presentations.

4.2 Voice Choice Across Social Contexts

In the previous section, we explored preferences for synthesized social profile voices in general, but participants also reflected on how those preferences may change based on the social context. We observed four common considerations: (1) the need to sound authentic and consistent regardless of social context, the desire to adjust voice tones and phonetic style across (2) content type and (3) audience, and (4) differences for private versus public social media posts.

First, participants preferred to keep the key characteristics of their synthesized voice consistent ($N = 13$).

For example, P5 emphasized the importance of consistency by reflecting on how their friends and contacts would react to voices that did not sound like her: *"...if I would have chose like number one [younger feminine voice] or three [younger masculine voice], for example, they would have been, like, What is she doing? That doesn't sound like her at all"* (P5). Additionally, P3 commented on the confusion that could arise when switching between voices: *"And I think it would get pretty confusing for the listeners, you know, that one day Lori [pseudonym] sounds like an old man, and the next day, Lori sounds like [inaudible], you know, I feel like it's probably easiest for everybody if it's consistent."*

Second, while maintaining a voice's overall consistency was important, many participants also felt that it would be useful to manipulate some characteristics depending on the specific content ($N = 13$). For example, P12 said, *"If we switch off [to a] different sound [voice]? That's weird. But if it's one person but with a different tone, that's fine."* Such nuanced changes were seen as particularly useful for sharing emotion ($N = 11$), such as: *"If it's consistent with the possible exception of adding that sort of emotional intonation, I think you have the same voice"* (P3). For some participants, being able to feel others' emotion meant deeper social connections and better mutual understanding. P2, for example, felt that emotion brings people "closer together" and stated:

> "I go through phases with my tweets where if I'm feeling particularly sad in a week, all of the tweets that I put out will be very sad. But then the second I'd bounced back, I would love to have something happy. Just so that it's not in constant sadness on my timeline for everyone." (P2)

Some participants only wanted emotional changes on voice synthesis for social media posts but not for everyday, functional applications. Most participants suggested that neutral, objective, or supplementary content, such as their bio information, and visually heavy content such as Instagram posts, did not need to be customized: *"Basic profile information would probably just be a standard like if I was giving a speech, like not really emotional, [but it should be] very formal."* (P2)

Third, people present themselves and interact differently on different social media platforms, and, accordingly, participants expressed considerations for *audience expectations* and *community culture*. About half of our participants reported that they would use a more professional voice for

career-oriented platforms such as LinkedIn ($N = 7$). For example, P9 said, *"If there was more options that came out, maybe one that sounded a bit more formal, maybe I would choose that for a LinkedIn or a Facebook kind of environment. Something more professional for those kinds of settings or more even-toned"* (P9). A few participants made similar comments about choosing a more professional voice on Twitter, while other participants felt they had more freedom in presenting themselves on Twitter.

Some participants specifically explained how these audience differences across platforms would influence their voice preferences ($N = 6$), such as P2 wanting to use a more positive voice on Facebook to connect with family members with whom she was not close enough to talk about serious topics.

Finally, for some participants, which personal characteristics to disclose through a synthesized voice also depended on how private or public the post content would be. For example, P13 had a private account where they could envision using a customized voice, but *"...maybe I wouldn't be as comfortable letting someone hear my voice if they weren't a friend, a public account. I guess, in that regard I would have the [default] Alexa voice."* P1 also mentioned that she would feel more comfortable using the "sad or angry" voices with a more private Instagram account but *"I feel like my main [account] that I use more only shows, like, the happy highlight..."* (P1). P6 also specifically pointed out that certain platforms are inherently more anonymous, such as Reddit, and that it would make more sense to use a default voice on it because *"...I wouldn't necessarily want everyone to know more about me."*

4.3 Challenges of Voice Customization for Self-presentations

While almost all interviewees shared excitement about navigating social media with this emerging mode of interaction, doubts about identity-based voice personalization arose. Below, we describe three main concerns: the possibility of stereotyping, misrepresentation, and impacts on usability.

Customization could result in stereotyping: The possibility of stereotyping was mentioned by more than half of our participants ($N = 9$), most notably related to genders and accents. An accented voice, in particular, may be inappropriately associated with certain groups of people, as mentioned by P4, *"I could see people using it maliciously as well, like make fun of accents"*, and P9, *"Especially with English, like Asian accents in English or something. I feel like some people will say it's too stereotyped, that [the] synthesizer makes it sound weird."* Some participants were specifically concerned that people would use this opportunity for cyber-bullying. P5, for example, said that *"It would provide an opportunity for the mean-hearted people in this world to mock and put others down, like (those who) with speech impairment[s] or disorder[s]."*

Also related to accents, many participants pointed out that there could be challenges with representing ethnicity through voice. For example, in response to a question about which aspects of their identity they would feel comfortable representing through a custom voice, two white participants (P4, P14), questioned what a "white-sounding" voice might sound like, while one Asian participant expressed a similar sentiment. Two participants (P7 and P8) who are people of color were concerned about perpetuating stereotypes with voices that would match their ethnicity: *"I would feel weird if a voice sounded overly African American in terms of [being too] stereotypical, because I would feel like they're just doing too much"* (P8). At the same time, P15, who identified as African American, felt that having a nuanced range of voices could provide an opportunity to counter stereotypes, saying: *"maybe if we had voices that sounded like mine or sounded like the different variations of people, then people wouldn't have that set stereotype of what an African-American man would sound like."*

A few people mentioned similar challenges around gender stereotypes, such as P13: *"I feel bad assuming that these voices sound like a specific gender."* P4, who identified as gender-queer, told us

that they did not feel fully comfortable representing gender through a synthesized voice. These findings suggest that speech perception is highly influenced by social stereotypes, calling for greater awareness of this phenomena in voice user interface design.

Customization may lead to misrepresentation: A few respondents were concerned about how customized voices may incorrectly intensify their posts and inadvertently offend listeners ($N = 3$): *"I don't really care about being inflammatory, but generally speaking, I try to post things that I know won't piss everyone off because the thing is like I do have family on there, I won't start fights online for no reason"* (P2). These participants would try to use a neutral or just a default voice to avoid this situation: *"So then it's not like... almost like someone can't judge what voice I chose, because it's just so neutral that it doesn't matter, like I almost didn't really choose anything"* (P14). Two participants also shared their unwillingness towards using a voice that attempted to be realistic but fell short, resulting in an "uncanny valley" [47] quality that could influence perceptions of the content. In this situation, a default or "normal" (P12) voice was preferable.

Potential impacts on usability: Important usability concerns arose related to the potential impacts of customized voices, which will need to be explored in future work. One of these was the burden of customization, commented on by four participants. For example, P14 felt that, *"If I didn't have the time to set it up, I'd probably just want to choose a default one"* unless she was *"really motivated to make my profile more fun or like more personalized to me."* P9 had a similar reaction, acknowledging that: *"I think I probably would only do it for maybe a few posts for me personally. I'm kind of lazier when it comes to social media, but I feel like it would be useful in the long run."* Another concern from one (sighted) participant was how customized voices would impact usability for blind users who rely on screen reader audio—here, P3 said: *"too much customization would impede [a screen reader user's] ability to actually get the content, which is a lot more important than the voice it's presented in."*

5 DISCUSSION

Through interviews with 15 participants, we explored how social media users may want to present their content for voice-based platforms. While past work has investigated synthesized voice preferences when interacting with and *consuming* voice-based content (e.g., different forms of texts [3]), our study is the first to examine preferences for how to present one's *own* socially oriented content. Besides common voice criteria such as intelligibility and likeability [7, 76], our findings suggest that these customized voices should match, or at least not interfere with, social media users' online personas as well as the specific content being conveyed. When a customized voice fails to satisfy these requirements, or when voice customization disrupts other needs such as usability, users would rather have a default, generic voice to represent them. These findings provide empirical knowledge to understanding what makes a voice suitable to represent a social profile, contributing to emerging discussions on voice interaction design for social computing [8, 42, 56].

In the following sections, we discuss how findings from this study add to the theoretical knowledge of *voice interaction design* and *online presentation of self*, and discuss when and how they may apply to other types of voice applications. We wrap up with design considerations for better supporting synthesized voice presentation of one's self online.

5.1 Online Self-Presentation and Synthesized Voice

Following patterns of online impression management [38], our participants tended to consider their synthesized voice presentations with a set of imagined audiences in mind. While past research has studied self-presentation on social media with visual and textual content (e.g., [5, 18, 33, 38, 82]), our study expands on this literature to identify its similarities with voice-based impression management and to reveal unique aspects that arise with the use of voice.

Some aspects of identity, such as gender, age, and accent, can be conveyed especially quickly and saliently through voice [15, 61]. Because these aspects may not be as easily extracted through textual information, synthesized voices allow users to be more mindful about their identity presentation choices, providing them with different and explicit opportunities for impression management on social media. For example, to maintain a consistent impression and to avoid being perceived as inauthentic, our participants judged sample voices based on both the extent to which the voice sounded like them and how other people would perceive that voice. To them, the key to having an authentic voice presentation was to convey the right personality and the important aspects of their identity. Furthermore, we observed how the lowest common denominator effect [26] might play out for synthesized voice selection: when in doubt about the appropriateness of a voice, many participants opted for what they saw as neutral defaults (e.g., Siri's or Alexa's default voice).

There was also evidence of other common impression management behaviors in participants' voice choices, including self-enhancement (i.e., the attempt to present one's self to others in a favorable light) [20] and conformity to the norm in a given social situation (e.g., a specific social media platform) [20, 38]. For example, social media sites are known to endorse positive self-presentation [9, 17], and this cultural expectation was reflected in our respondents' voice choices. Participants commonly favored positive and upbeat voices, even if they did not accurately reflect their emotional state.

Similar to how social media users adopt different verbal and visual styles across scenarios [38, 82], participants considered how they might adapt synthesized voices to appropriately convey their persona on a given platform and to modify the tone to be appropriate for a specific post.

Yet, synthesized voice options also present unique challenges due to limitations even in state-of-the-art speech synthesis. Participants emphasized the importance of high quality synthesized speech and raised concerns about how limitations such as mistaken pronunciation, not fully accurate emotional expression, and the "uncanny valley" effect could impact listeners' perceptions of a user's content and who they are. Inaccurate pronunciation could make one sound less professional, and, in the case of misspeaking the name of a close friend or family member, could also cause conflict in established close relationships. Low-quality manipulations of emotional expression and accents also have the potential to change the meaning of the content or even offend others. These concerns will need to be revisited as speech synthesis capabilities expand.

Researchers have recently called for more work in voice design theory, particularly in specific contexts [10], and the above discussion can inform the extension of existing theory to users as *producers* of a voice. Cambre et al. [8] propose a framework for designing voices for smart devices by considering the lenses of user, context, and device. While our focus on social media content tends to be device independent, the lenses of user and context are relevant. For the user lens, Cambre et al.'s framework considers computers to be social actors and suggests that voice designers take into account users' personal preferences for that social interaction (e.g., matching or complementing user characteristics). The focus is currently on how to make personality and other social characteristics of the computer or agent attractive to the user. Instead, when the design problem flips to focus on how to support individual users in representing themselves and their own social content, a voice designer needs to consider how (e.g., through personality, gender) and to what extent the user wants to reveal their own identities and persona.

Cambre et al. also look at voice design through the lens of the larger cultural and historical context of a smart device's use. Analogously, our findings show that voice preferences vary across social contexts, including different platforms, groups, or specific posts.

We speculate that, for platforms that support user-generated content that is not closely linked to one's identity (e.g., wiki pages or platforms that support anonymous participation such as Reddit), users will likely have a different set of voice requirements.

5.2 Design Considerations

Our study is forward-looking with the goal of identifying promising directions for future research and design exploration. Our findings show that with an expansion in voice-based interaction, providing configurable voices for spoken personal content could give social media users increased and welcome control over their self-presentation. Yet, there are also important technical, usability, and ethical challenges that will need to be addressed. Here, we offer design considerations for future work on socially situated, self-customized voice synthesis systems.

5.2.1 Voice customization options. Among all possible synthesized voice characteristics, what options should designers provide, and how? Our findings suggest that the overall quality and naturalness of the voice is a critical foundation for any customization. Indeed, current limitations with voice synthesis mean that perceived quality may be a barrier to adoption of fine-grained control over emotional tones and accents in the near term. New design features could address some of these current technical limitations with voice synthesis, such as allowing users to add phonetic gloss to names and other words that are likely to be pronounced incorrectly. Once a baseline level of quality is established, the question then becomes how to allow users to manipulate the voice. Some voice characteristics should be supported along a spectrum rather than enforcing rigid categories, such as gender and age (though age was not deemed as important as other characteristics by some participants). If supporting different accents, as many text-to-speech engines already do in a limited form, providing a nuanced set of options will likely be necessary to counter and prevent the risks of stereotyping, as discussed below. Finally, and perhaps most importantly, any voice customization should be optional, with designers providing a default voice and giving the user veto power if the system ever automatically adapts the voice to the content or user.

5.2.2 Reducing user effort. While we have argued that users should have fine-grained and nuanced control over how a voice sounds, that customization takes effort. While participants commonly wanted a representative voice, they did not necessarily need voices that were *identical* to their own. In fact, a voice that tries to sound exactly like the real person but fails may create an "uncanny valley" effect. An important question for future work then is to compare the value-effort trade-off of providing users with fine-grained control versus having a relatively small set of predefined voices. Another means of reducing setup effort might be to prompt users to specify a small set of adjectives (e.g., 3-5) that describe what they want to present about themselves and having the system automatically configure an initial profile voice based on that input.

Beyond an overall profile voice, adjusting phonetic characteristics for individual posts or other content could be particularly time consuming. Taking inspiration from AAC designs such as the Expressive Keyboard and Voicesetting Editor [19], we suggest having one-click voice tone options available when users compose their content, along with an always-available default or neutral voice option.

Automatic adaptation may also be attractive in helping to identify the appropriate voice tone to use, but any prediction mistakes could lead to confusion and misunderstandings, a concern raised by our participants.

5.2.3 Potential harms with personalizing synthesized voices. Important ethical considerations related to voice personalization arose in our study. Participants, especially non-Caucasian participants, expressed concerns that voices improperly associated with their ethnic and cultural background would propagate harmful stereotypes about their identity. One of the participants was also particularly worried about voice options for people with speech disorders. Designers must be attuned to any and all unintended stereotypes that can possibly result from voice technology development and must be transparent in communicating any algorithmic decisions made on users' identity

labels when generating voices. Whenever necessary, users should be able to change the vocal characteristics determined by the algorithm. As one of our participants said, we should not solely rely on algorithms to judge what a post, or a person, should sound like.

Further, highly personalized, natural-sounding voices can become identifiable information. If users' personal voices are usable by anyone on social media, there is a risk of individuals potentially using these voices for malicious purposes, such as deception, cyberbullying, or crime. While the voice synthesis community has been aware of potential ethical issues related to voice cloning [44, 60], it is unclear how the potential misuse of personal synthesized voices will be regulated. We suggest for designers to make the code of conduct and legal consequences of using voice synthesis very clear. Policy makers should also consider effective solutions to prevent harms from happening. A likely controversial policy possibility is to restrict the use of voice characteristics that do not apply to users' identities, such as by requesting users to register their personalized voices before usage. However, the implementation of this rule would be at the cost of limiting users' freedom of expression.

5.2.4 Toward broader audio content. Our study focused on presenting textual information via voice, yet social media content often contains non-text content as well: emojis, images, and videos. Existing explorations to voice out emojis include reading them aloud and playing sound clips of laughter or sighing [19]. As for pictures, current solutions rely on alternative text, which is often missing. If voice-based access to social media is adopted as a complement to more traditional access, users may become more motivated to write out image captions, thus improving the accessibility of these sites. Additionally, the audio modality opens new possibilities for self-expression, such as including signature background music as envisioned by some participants. How and whether to support this content will be important for future work to explore, and more effort is needed to identify and iterate on ways non-text elements can be delivered through voice.

5.3 Limitations

Our work has several limitations. First, this is an exploratory study in a new research area. While we provided concrete audio examples using participants' own social media content, participants still needed to speculate on their use of a future technology. Participant responses may have been impacted by a novelty effect and demand characteristics, which could have led to artificially positive responses to the overall idea of customizing voice; real patterns of adoption will no doubt differ. A critical next step in this work is to revisit the findings with a fully functional system.

Second, we explicitly sought to recruit a diverse set of participants and had some success in terms of gender and ethnicity; however, almost all participants were younger adults aged 20-30 and a majority were women. Including a wider range of participant ages could have generated additional themes and changed the importance of some themes (e.g., perhaps responses to representing age through voice would have been different). While our participants had experience with many social media platforms, they primarily used Instagram, Facebook and Twitter, which limits the degree to which findings may generalize to other platforms. Future studies should also consult users who rely heavily on audio interfaces for access, including screen reader and AAC users, about their perspectives on voice customization for self-presentation. Third, generating expressive synthesized voices is an open area of research [21, 57] and the examples we played for participants, particularly for emotions and accents, were not perceived to be fully natural. The lower quality of voice samples may have impacted participant responses, although there was overall a positive response to the idea of manipulating emotion at least to some extent. Finally, similar studies for other types of user-generated content, such as blogs and crowd-sourced review platforms, may also be valuable

for understanding how user requirements and preferences for audio-renderings of content vary (or stay consistent) across media types.

6 CONCLUSION

In summary, we presented an interview study that explored voice requirements specific to social media settings. Our findings uncover criteria for synthesized voices from content producers' perspectives and show how impression management strategies extend to preferences for synthesized voice presentation. As with visual and textual content [38, 82], voice presentation involves balancing authenticity and audience expectations. A favourable voice option should sound natural and also represent users' personal characteristics, emotion, and tone. These preferences on a representative voice may be overtaken by the concern of upsetting audiences with inappropriate or incomprehensible voice renderings. We collected a set of strategies that social media users proposed in light of this concern, including using a neutral or default voice and having multiple voice settings for different accounts. Our investigation also identified ethical concerns around propagating stereotypes with voice generated based on personal identities. Highlighting the diverse needs and concerns around voice renderings of social media content can inform designers of the need to give users more control over choosing voice options, attune to being transparent in communicating voice generation algorithms, and focus on creating intuitive methods to adjust voice parameters. Our work confirms and extends previous work on sociophonetics and online presentation of self, providing new insights into voice design for social and personal content.

ACKNOWLEDGMENTS

This work was funded in part by Mozilla and Microsoft Research.

REFERENCES

[1] David W. Addington. 1968. The relationship of selected vocal characteristics to personality perception. *Speech Monographs* 35, 4 (Nov. 1968), 492–503. https://doi.org/10.1080/03637756809375599

[2] Charles D. Aronovitch. 1976. The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology* 99, 2 (1976), 207–220. https://doi.org/10.1080/00224545.1976.9924774

[3] Alan Black and Keiichi Tokuda. 2005. The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech 2005*. 77–80.

[4] Danah Boyd. 2006. Friends, friendsters, and top 8: Writing community into being on social network sites. *First Monday* 11, 12 (Dec. 2006). https://doi.org/10.5210/fm.v11i12.1418

[5] Danah Boyd. 2007. Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media Volume* (2007).

[6] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces a real world driving study. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, New York, USA, 1–11. https://doi.org/10.1145/3290605.3300270

[7] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (apr 2020), 1–13. https://doi.org/10.1145/3313831.3376789

[8] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019). https://doi.org/10.1145/3359325

[9] Hui-Tzu Grace Chou and Nicholas Edge. 2012. "They Are Happier and Having Better Lives than I Am": The Impact of Using Facebook on Perceptions of Others' Lives. *Cyberpsychology, Behavior, and Social Networking* 15, 2 (Feb. 2012). https://doi.org/10.1089/cyber.2011.0324

[10] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016 arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/33525046/iwz016.pdf

[11] Josh Constine. 2020. Clubhouse voice chat leads a wave of spontaneous social apps. https://techcrunch.com/2020/04/18/clubhouse-app-chat-rooms/

[12] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017*. Association for Computing Machinery, Inc, New York, NY, USA, 1–12. https://doi.org/10.1145/3098279.3098539

[13] Bella M. DePaulo. 1992. Nonverbal behavior and self-presentation. *Psychological Bulletin* 2, 111 (1992), 203–243.

[14] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2019*. Association for Computing Machinery, Inc, New York, NY, USA, 1–12. https://doi.org/10.1145/3338286.3340116

[15] Katie Drager. 2010. Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass* 4, 7 (Jul. 2010), 473–480. https://doi.org/10.1111/j.1749-818X.2010.00210.x

[16] Edison Research and Triton Digital. 2019. *The Infinite Dial 2019 - Edison Research*. Technical Report. Edison Research and Triton Digital. https://www.edisonresearch.com/infinite-dial-2019/

[17] Ana Elias, Rosalind Gill, and Christina Scharff. 2017. Aesthetic Labour: Beauty Politics in Neoliberalism. In *Aesthetic Labour*. Palgrave Macmillan UK, 3–49. https://doi.org/10.1057/978-1-137-47765-1_1

[18] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication* 11, 2 (Jan. 2006), 415–441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

[19] Alexander J. Fiannaca, Ann Paradiso, Jon Campbell, and Meredith Ringel Morris. 2018. Voicesetting: Voice Authoring UIs for Improved Expressivity in Augmentative Communication. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173857

[20] Erving Goffman. 1990. *The presentation of self in everyday life*. Penguin, London.

[21] D Govind and · S R Mahadeva Prasanna. 2013. Expressive speech synthesis: a review. *Int J Speech Technol* 16 (2013), 237–260. https://doi.org/10.1007/s10772-012-9180-2

[22] Beth G. Greene, John S. Logan, and David B. Pisoni. 1986. Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers* 18, 2 (Mar. 1986), 100–107. https://doi.org/10.3758/BF03201008

[23] Acapela Group. 2020. Custom text to voice online. https://www.acapela-group.com/voices/custom-voices/

[24] J. Alex Halderman, Brent Waters, and Edward W. Felten. 2004. Privacy management for portable recording devices. In *WPES'04: Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society*. ACM Press, New York, New York, USA, 16–24. https://doi.org/10.1145/1029179.1029183

[25] HearMeOut. 2017. Voice Social Media Startup HearMeOut Formally Launches in U.S. https://www.prnewswire.com/news-releases/voice-social-media-upstart-hearmeout-formally-launches-in-us-300421352.html

[26] Bernie Hogan. 2010. The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society* 30, 6 (dec 2010), 377–386. https://doi.org/10.1177/0270467610385893

[27] Jiaxiong Hu, Qianyao Xu, Yingqing Xu, and Limin Paul Fu. 2019. Emojilization: An automated method for speech to emoji-labeled text. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313071

[28] Melissa G. Hunt, Rachel Marx, Courtney Lipson, and Jordyn Young. 2018. No More FOMO: Limiting Social Media Decreases Loneliness and Depression. *Journal of Social and Clinical Psychology* 37, 10 (2018), 751–768.

[29] Jesin James, Catherine Inez Watson, and Bruce MacDonald. 2018. Artificial empathy in social robots: An analysis of emotions in speech. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 632–637.

[30] Adam N. Joinson. 2008. 'Looking at', 'looking up' or 'keeping up with' people? Motives and uses of Facebook. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM Press, New York, New York, USA, 1027–1036. https://doi.org/10.1145/1357054.1357213

[31] Shaun K. Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. "at times avuncular and cantankerous, with the reflexes ol a mongoose": Understanding self-expression through augmentative and alternative communication devices. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Association for Computing Machinery, New York, NY, USA, 1166–1179. https://doi.org/10.1145/2998181.2998284

[32] Zahir Koradia, Piyush Aggarwal, Gaurav Luthra, Aaditeshwar Seth, and Gram Vaani. 2013. Gurgaon idol: A singing competition over community radio and IVRS. In *Proceedings of the 3rd ACM Symposium on Computing for Development, DEV 2013*. https://doi.org/10.1145/2442882.2442890

[33] Nicole C. Krämer and Stephan Winter. 2008. Impression Management 2.0: The Relationship of Self-Esteem, Extraversion, Self-Efficacy, and Self-Presentation Within Social Networking Sites. *Journal of Media Psychology* 20, 3 (2008), 106–116. https://doi.org/10.1027/1864-1105.20.3.106
[34] Kwan Min Lee, Katharine Liao, and Seoungho Ryu. 2007. Children's responses to computer-synthesized speech in educational media: Gender consistency and gender similarity effects. *Human Communication Research* 33, 3 (Jul. 2007), 310–329. https://doi.org/10.1111/j.1468-2958.2007.00301.x
[35] Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the conference on Human factors in computing systems - CHI '03*. ACM Press, New York, New York, USA, 289. https://doi.org/10.1145/642611.642662
[36] NaturalSoft Ltd. 2020. Free Text to Speech Online with Natural Voices. https://www.naturalreaders.com/online/
[37] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288
[38] Alice E. Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13, 1 (2011), 114–133. https://doi.org/10.1177/1461444810365313 arXiv:https://doi.org/10.1177/1461444810365313
[39] MaryTTS. 2020. The MARY Text-to-Speech System (MaryTTS). http://mary.dfki.de/
[40] Yossi Matias. 2020. Easier access to web pages: Ask Google Assistant to read it aloud. https://www.blog.google/products/assistant/easier-access-web-pages-let-assistant-read-it-aloud/
[41] Richard E. Mayer, Kristina Sobko, and Patricia D. Mautone. 2003. Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology* 95, 2 (Jun. 2003), 419–425. https://doi.org/10.1037/0022-0663.95.2.419
[42] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335
[43] Microsoft. 2020. Custom Voice. https://speech.microsoft.com/customvoice
[44] Microsoft. 2020. Microsoft's Responsible AI guidelines. https://speech.microsoft.com/customvoice
[45] Timothy Mills, H. Timothy Bunnell, and Rupal Patel. 2014. Towards Personalized Speech Synthesis for Augmentative and Alternative Communication. *Augmentative and Alternative Communication* 30, 3 (Sep. 2014), 226–236. https://doi.org/10.3109/07434618.2014.924026
[46] Pat Mirenda, Douglas Eicher, and David R. Beukelman. 1989. Synthetic and Natural Speech Preferences of Male and Female Listeners in Four Age Groups. *Journal of Speech, Language, and Hearing Research* 32, 1 (Mar. 1989), 175–183. https://doi.org/10.1044/jshr.3201.175
[47] M. Mori, K. F. MacDorman, and N. Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (2012), 98–100.
[48] John W. Mullennix, Keith A. Johnson, Meral Topcu-Durgun, and Lynn M. Farnsworth. 1995. The perceptual representation of voice gender. *Journal of the Acoustical Society of America* 98, 6 (1995), 3080–3095. https://doi.org/10.1121/1.413832
[49] Benjamin Munson and Molly Babel. 2007. Loose Lips and Silver Tongues, or, Projecting Sexual Orientation Through Speech. *Language and Linguistics Compass* 1, 5 (Sep. 2007), 416–449. https://doi.org/10.1111/j.1749-818x.2007.00028.x
[50] Casey Newton. 2018. Pocket redesigns its mobile apps to emphasize listening - The Verge. *The Verge* (2018). https://www.theverge.com/2018/10/11/17961564/pocket-redesign-listening-amazon-polly
[51] Nancy Niedzielski. 1999. The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology* 18, 1 (Mar. 1999), 62–85. https://doi.org/10.1177/0261927X99018001005
[52] Will Nigri and Rafael Perez. 2020. Audlist - Audio Social Network. Speak anywhere, be heard everywhere. https://audlist.com/en/
[53] Cathy Pearl. 2016. *Designing voice user interfaces : principles of conversational experience.* O'Reilly Media. 278 pages.
[54] Sarah Perez. 2017. Audm turns long-form print journalism into professionally narrated digital audio. https://techcrunch.com/2017/07/14/audm-turns-long-form-print-journalism-into-professionally-narrated-digital-audio/
[55] Victoria Petrock. 2019. The Who, What, When, Where and Why of US Voice Assistants - eMarketer Trends, Forecasts & Statistics. (2019). https://www.emarketer.com/content/voice-assistant-use-reaches-critical-mass
[56] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298
[57] Graham Pullin and Shannon Hennig. 2015. 17 Ways to Say Yes: Toward Nuanced Tone of Voice in AAC and Speech Technology. *Augmentative and Alternative Communication* 31, 2 (2015), 170–180. https://doi.org/10.3109/07434618.2015.1037930 PMID: 25965913.

[58] Thomas Purnell, William Idsardi, and John Baugh. 1999. Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology* 18, 1 (Mar. 1999), 10–30. https://doi.org/10.1177/0261927X99018001002

[59] Laura Robinson. 2007. The cyberself: the self-ing project goes online, symbolic interaction in the digital age. *New Media & Society* 9, 1 (Feb. 2007), 93–110. https://doi.org/10.1177/1461444807072216

[60] Kristen M. Scott, Simone Ashby, David A. Braude, and Matthew P. Aylett. 2019. Who Owns Your Voice? Ethically Sourced Voices for Non-Commercial Tts Applications. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 17, 3 pages. https://doi.org/10.1145/3342775.3342793

[61] Richard J. Sebastian and Ellen Bouchard Ryan. 2019. Speech Cues and Social Evaluation: Markers of Ethnicity, Social Class, and Age. In *Recent Advances in Language, Communication, and Social Psychology*. Routledge, 112–143. https://doi.org/10.4324/9780429436178-5

[62] Shootwords. 2020. Shootwords - Future Message, Voice Comments. https://shootwords.com/login

[63] Steven E. Stern, John W. Mullennix, Corrie Lynn Dyson, and Stephen J. Wilson. 1999. The persuasiveness of synthetic speech versus human speech. *Human Factors* 41, 4 (1999), 588–595. https://doi.org/10.1518/001872099779656680

[64] Steven E. Stern, John W. Mullennix, and Ilya Yaroslavsky. 2006. Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human Computer Studies* 64, 1 (Jan. 2006), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

[65] Elizabeth A. Strand. 1999. Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology* 18, 1 (Mar. 1999), 86–100. https://doi.org/10.1177/0261927X99018001006

[66] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300833

[67] Rie Tamagawa, Catherine I. Watson, I. Han Kuo, Bruce A. Macdonald, and Elizabeth Broadbent. 2011. The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics* 3, 3 (Aug. 2011), 253–262. https://doi.org/10.1007/s12369-011-0100-4

[68] Erik R. Thomas. 2002. Sociophonetic Applications of Speech Perception Experiments. *American Speech* 77 (2002), 115 – 147.

[69] Sonja A Trent. 1995. Voice quality: Listener identification of African-American versus Caucasian speakers. *Citation: The Journal of the Acoustical Society of America* 98 (1995), 2936. https://doi.org/10.1121/1.414099

[70] KJ Tusing and James Dillard. 2000. The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research* 26 (Jan. 2000), 148–171. https://doi.org/10.1093/hcr/26.1.148

[71] Unicode. 2020. Full Emoji List, v13.0. https://unicode.org/emoji/charts/full-emoji-list.html

[72] Aditya Vashistha, Richard Anderson, Abhinav Garg, and Agha Ali Raza. 2019. Threats, abuses, flirting, and blackmail: Gender inequity in social media voice forums. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300302

[73] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. Recall: Crowdsourcing on basic phones to financially sustain voice forums. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300399

[74] Sean Vasquez and Mike Lewis. 2019. MelNet: A Generative Model for Audio in the Frequency Domain. (2019). arXiv:1906.01083 [eess.AS]

[75] Voicery. 2020. Voicery Text-to-Speech. https://www.voicery.com/

[76] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Henter, Sébastien Le Maguer, Zofia Malisz, Eva Szekely, Christina Tånnander, and Jana Voße. 2019. Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Proc. 10th ISCA Speech Synthesis Workshop*. 105–110. https://doi.org/10.21437/SSW.2019-19

[77] IBM Waston. 2020. Text to Speech Demo. https://text-to-speech-demo.ng.bluemix.net/

[78] IBM Waston. 2020. Watson Text to Speech - Overview | IBM. https://www.ibm.com/cloud/watson-text-to-speech

[79] Sophie F Waterloo, Susanne E Baumgartner, Jochen Peter, and Patti M Valkenburg. 2018. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New Media & Society* 20, 5 (May. 2018), 1813–1831. https://doi.org/10.1177/1461444817707349

[80] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. 2020. Random, Messy, Funny, Raw: Finstas as Intimate Reconfigurations of Social Media. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376424

[81] Eric Zeng, Shrirang Mare, Franziska Roesner, and Paul G Allen. 2017. End User Security and Privacy Concerns with Smart Homes End User Security & Privacy Concerns with Smart Homes. *Proceedings of the Thirteenth Symposium on Usable Privacy and Security* Soups (2017), 255–272. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/zeng

[82] Xuan Zhao, Niloufar Salehi, Sasha Naranjit, Sara Alwaalan, Stephen Voida, and Dan Cosley. 2013. *The Many Faces of Facebook: Experiencing Social Media as Performance, Exhibition, and Personal Archive.* Association for Computing Machinery, New York, NY, USA. 1–10 pages. https://doi.org/10.1145/2470654.2470656

Received June 2020; revised October 2020; accepted December 2020