"A Voice that Suits the Situation": Understanding the Needs and Challenges for Supporting End-User Voice Customization

Hyeon Jeong Byeon*
Chaerin Lee*
Jeemin Lee
Uran Oh†
cat@ewhain.net
chaerinlee@ewhain.net
hellofairy@ewhain.net
uran.oh@ewha.ac.kr
Ewha Womans University
Seodaemun-gu, Seoul, South Korea

ABSTRACT

Although there is a potential demand for customizing voices, most customization is limited to the visual appearance of a figure (*e.g.*, avatars). To better understand the users' need, we first conducted an online survey with 104 participants. Then we conducted a semi-structured interview with a prototype with 14 participants to identify design considerations for supporting voice customization. The results show that there is a desire for voice customization especially for non-face-to-face conversations with someone unfamiliar. In addition, the findings revealed that different voices are favored for different contexts from a better version of one's own voice for improving delivery to a completely different voice for securing identity. As future work, we plan to extend this study by investigating voice synthesis techniques for end-users who wish to design their own voices for various contexts.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

KEYWORDS

voice perception, voice customization, voice quality

ACM Reference Format:
Hyeon Jeong Byeon, Chaerin Lee, Jeemin Lee, and Uran Oh. 2022. "A Voice that Suits the Situation": Understanding the Needs and Challenges for Supporting End-User Voice Customization. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3491102.3501856

*Both authors contributed equally to this research.
†The corresponding author.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501856

1 INTRODUCTION

In recent years, the pandemic situation has required social distancing among people, making a shift in the way of interacting with others. We are witnessing many of these connections moving into the metaverse, a three-dimensional virtual space where people interact with each other using the metaphor of the real world [13]. Instead of physical co-presence, millions of avatars participate in social gathering events hosted on the metaverse ranging from online gaming to business meetings. As many studies have shown, these avatars are perceived as a representation of their owners as they are alike [33, 35]. Thus, avatar-based customization plays an important role by decreasing the self-discrepancy between the user and the character [33]. Noting the fact that higher avatar similarity leads to higher satisfaction and self-presence in a virtual world [14], many metaverse platforms provide users with detailed options to customize their avatars to look like themselves. Second Life (secondlife.com), for example, provides over 150 unique sliders to customize the appearance and the outfit of their avatar. A plethora of avatar-based customization options led to an active discussion among researchers about the user behaviors of avatar customization services. For instance, Drouin *et al.* [7] proposed an online community where users can beautify or even lie about themselves to look more attractive to others. Yet, metaverse platforms and studies on avatar-based customization fall short of details on the avatar's voices and only concentrate on the appearances although studies have shown that voice impacts the evaluation of other people's impressions [9]. In addition, considering that preferable voices that represent the user's profile vary on each social media platform [42], voice customization can be used to set different voices for different contexts. Also, as female users have a greater tendency to conceal their information for particular situations [28, 40], a customized voice can prevent them from revealing their identity. Lastly, people with speech or hearing disorders can benefit from voice customization. Indeed, companies (vocalid.ai, soundthinkers.co, and mov.acapela-group.com) have presented synthesized AI-voice personas which represent the unique personalities of individuals. VocalID (vocalid.ai), for instance, uses state-of-the-art machine learning and speech blending algorithms to create customized synthetic voices for people who use speech synthesizers. Discord (discord.com), a textual and vocal online communication

platform, allows users to experience voice customizing applications such as the Voicemod (voicemod.net) and the AV voice changer software[1]. While inspiring, these services are limited to serving as a voice customizing tool and they simply modify the original voice using voice filters rather than building a sophisticated voice that can be used to express oneself like customizing the appearance of an avatar. Moreover, these filtered voices are often exaggerated than the original voice, which hinders the utility of the system. Thus, in this paper, we explore the following research questions about end-user voice customization:

- *RQ1. How do individuals wish to change their voice in what circumstances?*
- *RQ2. What are the affecting factors for desiring certain voices?*
- *RQ3. How can we design a voice customization tool to support end-users?*

To understand the potential demands of voice customization especially for online communications, we first conducted a formative online survey with 104 participants. We investigated how individuals perceive their voice in terms of satisfaction and how they react to the idea of using customized voices under four different situations varying context (*e.g.*, social vs. school or work) and the person that they are talking to. The survey showed that the demands exist for voice customization regardless of the satisfaction level of one's own voice. Also, we found that types of situations affect the participants' willingness and concerns about voice customization. Next, to explore the design considerations for voice customization tools, we designed a web-based prototype which was served as a voice searching engine for choosing the desired voice and to collect participants' feedback with demonstrations. Then we conducted an in-depth interview with 14 participants recruited from the online survey to identify design considerations for supporting voice customization for end-users. Our findings uncover the participants' voice preferences under different situations. To be specific, we found that participants do not wish to or are reluctant to change their voice dramatically when talking to someone familiar such as friends and professors to maintain the intimacy they have built. On the other hand, they are more willing to use voice customization when talking to someone unfamiliar such as another online game player not to reveal their identities such as gender and age.

This paper makes the following contributions: (1) the assessment of needs for end-user voice customization under different situations, (2) the identification of voice customizing behavior of end-users, and (3) the design recommendations for a future end-user voice customization tool.

2 RELATED WORK

Our work is informed by studies on appearance customization, social roles of voices in communication, and voice preference.

2.1 Appearance Customization of Avatars

Users perceive avatars as themselves [33, 35], and it is found that the customization of avatars enables users to fulfill their urges for self-presentation [26]. In addition, it is shown that users are attracted to avatars that share identical visual features with themselves such as hairstyle, eye color, outfit, and facial characteristics [32]. Moreover, the visual similarity between the user and the avatar is found to be positively correlated with the level of satisfaction and self-presence in online space [14]. Meanwhile, it is also found that users customize their avatars to look more appropriate in virtual contexts [31] or to look more appealing to the others in virtual space [7]. Furthermore, a number of studies show that these appearance customizing behaviors establish cooperative behaviors among users and even impact their self-esteem [3, 6, 36].

However, regardless of the positive impacts for supporting avatar customization, most of these studies are limited to changes in appearance although there are potential needs for voice customization [9, 41]. As we expect the demands for voice-based online communication to grow (*e.g., Discord*), we focused on investigating the demands and user behaviors for voice customization more in depth under different scenarios and demonstrate a prototype for customizing voices for end-users.

2.2 Effects of Voice During Communications

While voice customization has not received much attention as appearance customization, voice does significantly impact communication in everyday activities. First, one's voice can be used to estimate the information about the person. For instance, a single utterance can reveal one's demographic information such as a speaker's age group, gender, ethnicity, and socioeconomic class of others [27, 30, 41]. In addition, speaking habits are often associated with one's personality [34, 41]. Second, voice impacts the delivery of the context. According to Mubarak *et al.* [22], the gender and accent of the speaker influence the credibility of information. For instance, their participants gave more credit to factual statements delivered by male voices than female voices, considered formal voices to be more accurate than informal ones regardless of the content. Third, vocal interactions increase the intimacy among people in virtual spaces. Studies have shown that compared to text-only communications, voice communication leads to intensely close experiences among users by increasing likeability and trust [37, 39].

Inspired by the studies that showed the significant role that voice has, we explored various voice attributes that may affect users' voice customizing behaviors.

2.3 Voice Preference

User behaviors about voice preferences have long been studied in the field of Human-Computer Interaction. For example, it is shown that people are attracted to a synthesized voice that shares some similarities with them such as the personality [4, 20], gender [19], or the accent [21, 29] of their own. Also, studies also found that people evaluate voice more positively when it is perceived to be appropriate given the context [19] such as textual content [4, 20, 21, 29, 41] or the type of online communication platforms. According to Zhang *et al.* [41], for instance, people prefer voices that match the user information and context of the social media platform. To be specific, people wished their Twitter profile to be delivered in a delightful voice, while their LinkedIn profile to be delivered in a professional voice. These behaviors also applied to individual posts. Meanwhile, several studies investigated the relationship between various voice

[1] https://www.audio4fun.com/voice-changer.htm

attributes and user preference such as the perception and stereotypes induced by voice attributes. [2, 24]. On the other hand, others focused on a specific voice attribute at a time. According to a study conducted by Niculescu *et al.* [23], the pitch of a robot's voice affects the overall user experience the most in terms of the interaction quality, the robot's appeal, and enjoyment. Moreover, Zhang *et al.* [42] focused on glottal flow-based attributes to improve the delivery of voice with high preference.

While there are numerous factors that can affect users' preferred voice, we focused on two contexts (social vs. business) and the level of intimacy towards the conversation partner(s). Also, we investigated types of preferred voices in terms of a set of voice qualities and why a certain voice is preferred over other voices under different circumstances.

3 FORMATIVE STUDY: AN ONLINE SURVEY

To explore the potential demands for customized voices during online communication, we first conducted an online survey as a formative study.

3.1 Methods

We used Google Forms to create the survey, which was distributed in Korean via word-of-mouth, the university board, and online communities for software developers, researchers, and online game users (*e.g.,* Kaggle Korea Facebook group, Discord servers related to online games). The survey lasted for 7 days and anyone aged between 18 and 65 were allowed to participate in the survey. It was designed to take approximately 10 minutes and it consists of 20 questions in total. The questions include demographic information, satisfaction of respondents' own voice, if and how they wish to change their voice under different situations (see the appendix for details).

3.2 Participants

We had 104 respondents (32 male, 71 female, 1 preferred not to disclose). Almost half of the respondents were aged between 25 and 34 (48.1%); 36.5% for between 18 and 24, 11.5% for between 35 and 44, 1.9% for both between 45 and 54, and between 55 and 64.

3.3 Findings

3.3.1 Satisfaction of One's Own Voice. To understand how individuals are satisfied or unsatisfied with their own voice, we asked participants to rate the level of satisfaction of their voice on a 5-point scale where 5 is the highest. As a result, we found that the majority of the participants (54.8%; 57 out of 104) rated 4 or more (13 respondents rated 5, 44 rated 4). Meanwhile, 34 participants (32.7%) were neutral, and the rest 12.5% of the respondents reported that they are not satisfied with their voice rating less than 3.

3.3.2 Situations when One Had Wished to Change Their Voice. To understand if participants had wished to change their voice and under which circumstances is so, we then asked participants to mark if they had wished to change their voice for each of the following four situations varying the context (*social* vs. *school or work-related*) and the familiarity with the conversation partners:

- $Social_F$: Socializing with someone familiar (*e.g.*, phone or video call with friends and family members)
- $Business_F$: Communicating with someone familiar at school or work (*e.g.*, online classes, online conference meetings)
- $Social_{UnF}$: Socializing with unfamiliar people (*e.g.*, in-game voice chats, YouTube live chats with audiences)
- $Business_{UnF}$: Communicating with unfamiliar people at school or work (*e.g.*, customer support telephone services, creating informative YouTube videos)

As a result, 80.8% of the participants (N = 84 out of 104) answered that they wish to have a different voice for at least one of the four situations, which includes the ones who are satisfied with their voice. Yet, most of the ones who reported none were the ones who were satisfied with their voice; of 20, 7 rated 5, 11 rated 4, 2 rated 3.

The number of responses for each situation is: 33 for $Social_F$, 27 for both $Business_F$ and $Social_{UnF}$, and 36 for $Business_{UnF}$. While we expected that the desire for changing their voice would depend on the situation. As participants were allowed to submit their own responses for any other situations they had wished to change their voice, one reported situations when having an oral presentation or singing in front of others.

3.3.3 Willingness for Using Voice Customization. We then further asked participants directly if they wish to use voice customization under each of the same four situations described in subsubsection 3.3.2 to investigate if there is a desire for voice customization and if it varies depending on the circumstances.

As shown in Figure 1, the number of participants who wish to use voice customization was the highest for $Business_{UnF}$ with 89 responses, followed by $Social_{UnF}$ (N = 82), $Business_F$ (N = 75) and $Social_F$ (N = 74). We also asked participants to report why they wish to use voice customization. Although the most dominant reason varied across the four situations, the top four reasons why they would use voice customization were the same as shown in Figure 2. The most dominant reason was the ability to communicate anonymously for $Social_F$, $Social_{UnF}$, and $Business_{UnF}$ (26 out of 74, 50 out of 81 and 55 out of 89, respectively). However, the dominant reason for $Business_F$ was that they wish others to concentrate only on the context (N = 41 out of 75).

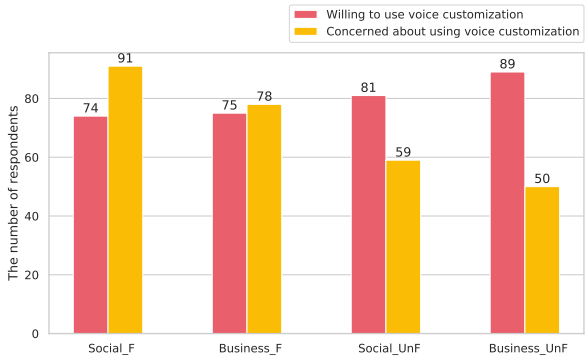

Figure 1: The number of respondents who are willing to use, or concerned about using voice customization for four situations. (N = 104)

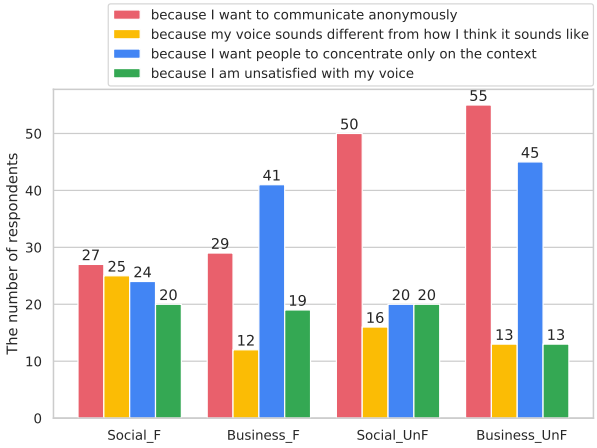

Figure 2: The number of respondents for different reasons for their willingness to use voice customization. (N = 104)

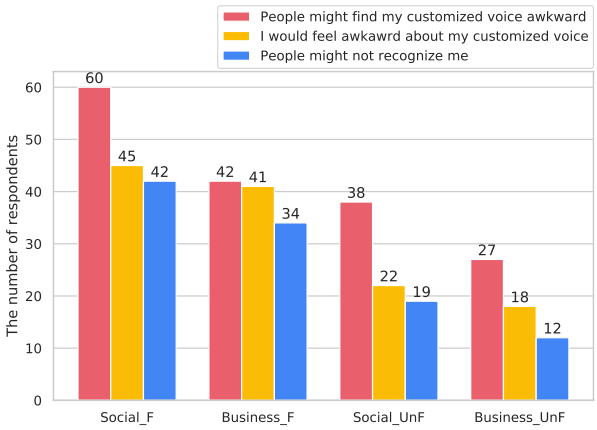

Figure 3: The number of respondents for each type of concern about voice customization for each situation. (N = 104)

3.3.4 Concerns for Using Voice Customization. We asked participants if they have any concerns using voice customization for each situation, and the results reflect the responses for the participants' willingness. The number of participants who showed concerns was the highest for $Social_F$ (N = 91), followed by $Business_F$ (N = 78), $Social_{UnF}$ (N = 59) and $Business_{UnF}$ (N = 50).

Again, we asked participants to specify their concerns for voice customization, which also varied across four circumstances. The results are shown in Figure 3. Regardless of the situation, the dominant reason was that others might find their customized voice awkward, particularly for $Social_F$. The concern about other people not recognizing him/her voice was the next. However, the number of responses was higher for $Business_F$ and $Social_F$ than the other two situations where they are communicating with someone who is unfamiliar.

3.3.5 Perceived Necessity for Supporting Voice Customization. Lastly, we asked participants about the perceived necessity for supporting voice customization for each situation in a 5-point Likert scale. Figure 4 suggests that most participants think that it is relatively unnecessary to support voice customization when talking to someone familiar; the average was 1.9 for $Social_F$ (SD = 1.0) and 2.5 for $Business_F$ (SD = 1.2). On the other hand, supporting voice customization was appreciated more for the situations at work or at school; 3.5 for $Social_{UnF}$ (SD = 1.2) and 3.7 for $Business_{UnF}$ (SD = 1.1).

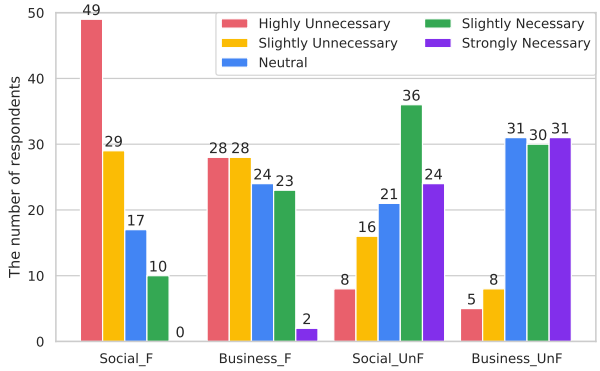

Figure 4: The perceived necessity for supporting voice customization for each situation in a 5-point Likert score.

4 MAIN STUDY: AN INTERVIEW WITH A DESIGN PROBE

We conducted a semi-structured interview with a design probe to have a deeper understanding of the needs for and reaction towards voice customization. We also had five specific scenarios derived from the ones we used in our survey to investigate if participants desire different voices for different situations.

4.1 Participants

We recruited 14 interviewees (11 female and 3 male) from our formative study respondents who showed their interest in participating in a follow-up interview study. They were 26.1 years old on average (SD = 5.6; range 19-38).

4.2 Apparatus

For our main study, we designed a prototype for voice customization, which is basically a search engine with voice attribute filters. As shown in Figure 5, it enables users to (1) specify the gender and the age group of the desired voice and different attributes that define voice quality, (2) listen to the voice samples that meet the users' filtering conditions after clicking the "search" button, and (3) choose their desired voice. As for the implementation, we used Django, a Python-based open-source web platform for our framework and S3 buckets in Amazon Web Services(AWS) to deploy the website. As for the voice dataset, we used Speech Accent Archive[38], which has 2176 utterance samples from 2176 different speakers diverse in age (between 6 to 97), gender, native language, the number of years of English practice. Each speaker read the same paragraph in English (*i.e.*, *"Please call Stella. Ask her to bring these things with*

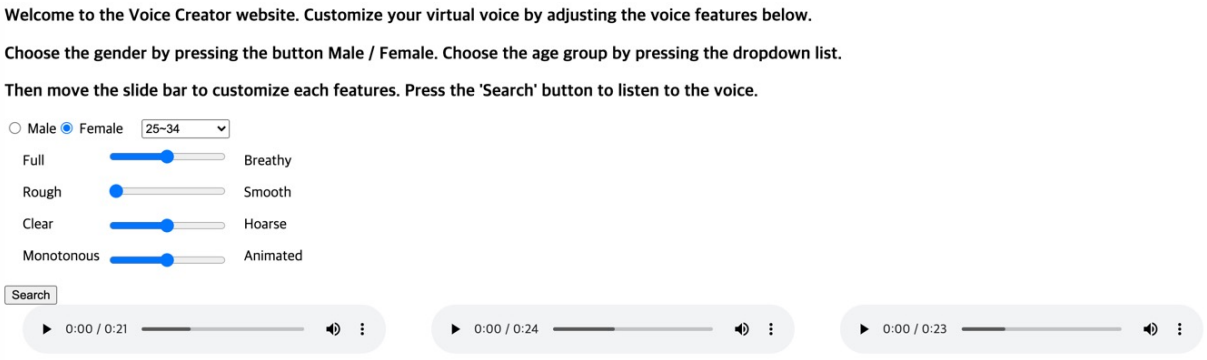

Figure 5: A screenshot example of our voice customization prototype used in the main study. Participants could choose *gender*, *age group* from drop-down menus and specify *breathiness*, *smoothness*, *hoarseness*, and *variation* in a 3-point scale with sliders. Clicking the "search" button would show all audio recordings, three in this case, that meet the criteria above for participants to listen to before choosing a specific voice.

her from the store(...)". Then we annotated each voice sample using Amazon Mechanical Turk (mTurk) with the following seven voice attributes: *breathiness (full ↔ breathy)*, *hoarseness (clear ↔ hoarse)*, *pitch (low ↔ high)*, *smoothness (rough ↔ smooth)*, *speed (slow ↔ rapid)*, *variation (monotonous ↔ animated)* and *volume (soft ↔ loud)*. These voice attributes were selected based on the papers in the field of phonetics that studied voice qualities (*i.e.*, [5, 8, 10–12, 25]). For mTurk crowd workers, their task was to listen to one of the utterance samples described above then rate each attribute in a 5-point scale. During the task, the definition of each attribute was provided as we did not expect them to have background knowledge in phonetics.

While our dataset has utterances with seven different voice quality attributes, we removed *pitch*, *speed*, and *volume* in our prototype[2] as these are easily manipulable with existing voice synthesize techniques [16, 17] and focused on the rest.

For the main study, we used a modified version of the prototype to support a 3-point scale that corresponds to 2 to 4 from the 5-point scale collected from crowd workers since voice samples with the extreme values (*i.e.*, 1, and 5) were rare, and to log participant IDs and their desired voices for different scenarios as well as to collect participants' opinions about the prototype[3]. The interview was conducted using an online video conferencing platform (*i.e.*, Zoom) with a screen-sharing feature turned on. Each session was recorded and transcribed.

4.3 Procedure

The interview was conducted using Zoom, and it began by submiing a consent form via Google Form. We first asked demographic information of the participants and voice-related experiences, then we showed and explained our voice customization prototype and asked participants to search for the voice that is most similar to theirs. Then, we presented the following five different scenarios, derived and modified from the situations from our formative study, and asked how participants would create a customized voice using our prototype for each:

- ***Holiday Greetings***: *"You are going to call your family and friends who are far away to send holiday greetings and catch up with the latest news for the holidays."*
- ***Online Classes***: *"You plan to participate as a student in a real-time online lecture for the upcoming semester. Assume that the class has an active Q&A session, presentation, and discussion, and you already know other classmates and the instructor face-to-face."*
- ***Team-based Games***: *"You want to communicate with other members of the same team in a game where it needs teamwork (e.g., Overwatch, Battleground). At this time, You've decided to participate by using the voice talk feature instead of chatting where you had to type on the keyboard and read the text on the screen."*
- ***Customer Services***: *"A delivered item was damaged so you've decided to call the customer service or the product manufacturer to ask what can be done."*
- ***YouTube Channel***: *"You want to create informative YouTube content (e.g., documentaries) on a subject that you have expertise in. So you are trying to edit a video by adding video clips, subtitles, and voice. This video can be watched by a large number of unspecified users on YouTube."*

For each scenario, we asked them to describe the voice they wish to have, then to choose a specific voice using our prototype and asked reasons for their choice. After going through all five scenarios, participants were asked to provide feedback about the prototype. Participants' screens were shared throughout the study.

4.4 Data Analysis

As for analyzing interview responses, two researchers developed initial codebooks together for each question, then independently coded a randomly selected 10% of the responses, following an iterative coding process [15]. We had 1 to 3 iterations for each question to refine codes. The Cohen's kappa across all codes after the final iteration was 0.83 on average (SD = 0.06, range from 0.76 to 0.97).

[2]http://ec2-3-36-91-129.ap-northeast-2.compute.amazonaws.com:7777/
[3]The project repository can be found at https://github.com/hyeonJeongByeon/VoiceCreator.

4.5 Findings

4.5.1 The Desire for Changing One's Own Voice. Next, we asked participants in what situations they wished to change their voice if any. All but three participants (P1, P3-P4) specified that they had wished to have a different voice in the following situations: personal conversations in daily life (N = 3) , online platforms like games (P2, P9, P11), delivery orders (P5, P12), and work-related situations including business calls (P6, P10).

> *"If I create a video for an anonymous report or for the public interests sometimes, I would have to use my voice for the narration. But I want to use an anonymous voice in this case." (P6)*

4.5.2 The Voice Attributes of Interests. We further asked them how they wish to change the attributes of the voice, five participants answered pitch (N = 5 out of 11); three of them wished the pitch of their voice to be lower (P2, P7, P8) while the remaining two (P12, P13) hoped to have a high-pitched voice. Other attributes were hoarseness and volume; P8 wished to have a hoarse voice and P14 wished her voice to be louder. Although not related to voice attributes four participants (P6, P8-P9, P11), who were all female, wished to have a voice that would not reveal their identity such as their gender. In addition, P10 and P13 wished to have a voice that would sound older.

4.5.3 Prior Experience with Voice-Changing Services. We also asked participants if they had ever changed voices using voice-changing services. As a result, all but three participants (P9, P11, P14) reported *'yes'*, which was higher than our expectation. Most of the participants (N = 10 out of 11) answered that they used a voice filtering service when talking to someone they are familiar with, while P10 used a modulated voice during a random chat. Moreover, while all but one tried changing their own voice, P6 altered the voice of someone else. We further asked why they tried changing their voice. Nine out of 11 participants (P2-P5, P7-P8, P10, P12-P13) said they tried it for fun, and some tried the voice modulator as they were curious (P1, P8). However, none of them perceived the voice-changing services for long-term use. As pointed out by P13, it could be due to a lack of available services that can be used for typical communications. Indeed, most of the participants tried a voice filter that would change their voice in a funny way like *My Talking Tom*[4].

4.5.4 The Desired Voices for Each Scenario. To investigate if participants wished to have different voices under different circumstances, we asked participants to describe the voice that they would like to have and choose a voice by specifying the values of the four voice attributes for each of the five scenarios. The attribute values are shown in Figure 6.

Holiday Greetings. As for *Holiday Greetings*, participants wished to use their own voice as is (P2, P7, P11, P13). P13 said, *"(...) I don't want to become less familiar when approaching my family or friends, and I actually don't know if that's needed."* The rest wished for a better version of their voice. For example, P5, P10 and P14 wished their voices would sound brighter, reflecting the results shown in Figure 6.

[4]https://apps.apple.com/us/app/my-talking-tom/id657500465

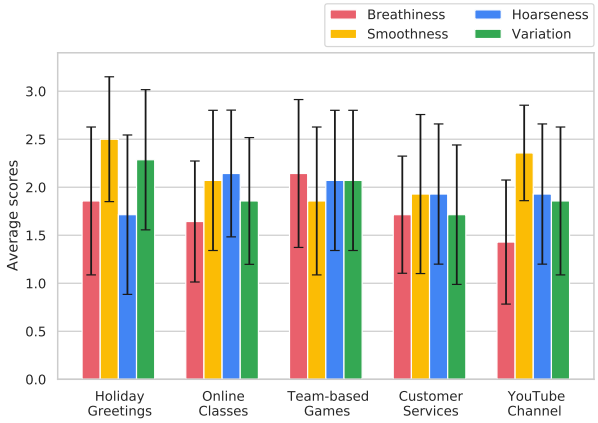

Figure 6: Average scores of each voice attribute for each scenario. The error bars indicate standard deviations. Note that each has three levels (1-3).

> *"I wish it [my voice] to be a little brighter. I think my family would want me to have fun in my life, so I'd like to have a voice that has a lot of changes in pitch so that I can let them know I'm really enjoying my life." (P10)*

In addition, the desire for *smoothness* was and *clearness* (the opposite of *hoarseness* were higher compared to other scenarios as well. Indeed, regarding the delivery (pronunciation, volume), P4 wished for a clear voice. P5 wished her voice to be louder so that her grandma could hear her better.

Online Classes. Similar to *Holiday Greetings* but slightly greater number of participants (N = 7; P1, P3-P4, P9, P11-P13) wished to use their own voices for online classes. One specified that she is reluctant to use a modulated voice in class because she thinks it is not polite when asking questions to professors. Yet, 7 participants wished for a voice that has a good delivery. They wished their voices to sound smart and calm (N = 2) and clear (N = 2).

In terms of voice attributes, two participants (P8, P14) wished less variation in pitch and one (P2) wished a voice with a low breathing sound. In addition, P1 wanted a trembling tone correction.

Team-based Games. Unlike other scenarios, gender-related responses (N = 7) was the most common with a distinct tendency for preferring male voices (P3, P5, P8, P10-P11, P13), except for P2 who would like to communicate in a neutral voice without revealing her gender unless she knows the other person well. Note that two participants (P12, P14), who answered that they would use their own voices, had no experience using voice to communicate when playing online games. P10 commented that,

> *"In the previous situations, I actually chose female voices for all other scenarios. But, in this case, I wish to use a male voice. I felt it many times while playing the game that if the speaker is a woman, there is a tendency to look down on gaming abilities a bit."*

Similar to *Online Class* scenario, participants were also interested in the delivery of their voice (N = 6). Participants also wished to change the perceived age of the voice (N = 6). Four of them (P4-P5, P11, P13) wished for an older voice while the other two (P1, P10)

preferred a younger voice. In fact, in the voice selection process, 8 out of 14 participants chose a voice older than their age. Some participants (P4-P5, P8) wished for a powerful voice that could help to avoid conflicts. For instance, P4 wished to have a deep, low-pitched voice with strong leadership.

Customer Services. Similar to the *Online Class* scenario, most of the responses wished to have a voice with good delivery such as low in pitch with little variation (N = 10) where the scores for all four attributes are below 2 (neutral).In addition, half of the participants chose voices that do not show their emotions (N = 7). P9 described,

> *"Since this is a situation where I need to go through the process of returning [the purchased item], I'd be speaking in an angry tone no matter how kind the operator behaves. So I want to communicate in a calm voice that does not show emotions."*

Similarly, 6 participants wished for a strong and determined voice. As an exception, P6 chose a breathy voice of an old man with high variation in pitch because she thinks that the operator would deal with the complaints faster if she sounds angry.

Interestingly, unlike all the other scenarios, none of the participants answered that they would like to use their own voice as is. P12 specified,

> *"It would be their first time to hear my voice for the consultant at a customer service center, and they will not hear my voice again. So, I thought it would be okay if the voice is not similar to my own voice."*

YouTube Channel. Again, the most common content was about delivery. Particularly, variations in pitch were mentioned the most (N = 6; P1-P2, P4, P7, P9, P11), followed by having a good voice to listen to (N = 6; P1,P3-P4,P6,P10-P11). P9 said, *"Since the purpose is to deliver information, I think it's good to have a voice that can convey information and information only monotonously without changes in pitch nor emotions."* Some participants (P8, P10, P12) specified voices of professions that talk for large audiences regularly such as voice actors, announcers, and professors. P8 described, *" I chose it [the voice] because I felt that it was the most similar to the voices of the BBC documentary voice actors. It was a moderate tone that gives credibility and the speaking speed was also good so that I think it would be great for a documentary video."* Indeed, according to the attribute scores for voices that participants selected, their desire for having a clear voice was the strongest in this scenario compared to others (see Figure 6). The *smoothness* was also high, which is another attribute related to the delivery of the content.

Meanwhile, five participants (P1, P5, P11-P13) wished to use their own voice. P1 wished not to make any changes in his voice to better credit the work that he created. P13 also explained that,

> *"With this level [of commitment], it would mean that others might look up to find out who made the content [with good intention]. So I don't think it's necessary to keep my gender a secret. And I won't be embarrassed because this would be something that I know very well."*

4.5.5 The Scenario where Voice Customization is Most and Least Wanted. Among the five scenarios, participants were asked about the situation in which they would use voice customization the most and the least. As a result, 7 out of 14 participants chose *Team-based Games* as the most likely situation for voice customization, followed by *YouTube Channel* (N = 4), and *Online Classes* and *Customer Service* (N = 2 for both), while no one chose *Holiday Greetings*. Note that one participant chose both *Team-based Games* and *YouTube Channel*.

On the other hand, *Holiday Greetings* was chosen as the least likely scenario for voice customization by 10 participants, followed by *Online Classes* (N = 5). Note that one participant selected both scenarios. Meanwhile, none of the participants chose the other three scenarios. P10 said that voice customization would not be required for talking with acquaintances.

> *" I will use the voice customization function less in Holiday Greetings than in the other scenarios. Because they know my voice well, so I want to reduce uncertainty even if I talk on the phone after a long time. Besides, I think it is still vague whether the technology of this web service has been 100% proven."*

Additionally, participants were asked why they wished for a similar or different voice for different scenarios. As a result, nine participants answered that they chose a voice that suited the situation. P8 specified that there is a voice more or less favorable depending on the situation.

> *"People are expected to wear different clothes depending on the situation. Likewise, I think it is the same with the voice. It needs to be tailored for the given situation. If I speak in a bright tone at a funeral, for example, it might be a little rude. So I think that it is more important to adjust the tone or volume of the voice according to the situation and thus the situational context."*

Five participants answered that it was because of anonymity or privacy. For example, P10 shared an uncomfortable experience of noticing that she encountered her game partner at work through the voice. And P13 said, "It doesn't seem to make much sense to use a different voice once the identity is revealed."

4.5.6 Voice Attributes Considered the Most and Least Useful. We first asked the participants about the most useful voice attributes and unnecessary ones while using our prototype. As a result, the most useful attributes when searching for a specific voice were *age group*, *gender*, *smoothness*, and *variation* (N = 4 for each); *breathiness* and *hoarseness* were considered less useful (N = 2 for each). In general, participants tended to choose attributes that they believe as a distinctive feature for distinguishing one voice from another.

We then asked about the voice attributes they thought were unnecessary, and most of the participants (N = 8) answered *'none'*. Meanwhile, *breathiness* was considered as the least useful attribute by four participants, followed by *hoarseness* and *variation* (N = 1 for each). None of them considered *age group* and *gender* not useful.

In addition to the six voice attributes, we asked what other attributes participants wish to customize if any. Six participants responded that they wish to be able to change *pitch* (including vocal range). Next, *speed* was mentioned by four participants. Some participants (N = 3) wished to be able to add *emotion* to their customized voice. *Preciseness* of the pronunciation and *volume* were chosen by two participants. *Vibration*, *character*, and *tone(timbre)* were also mentioned by one participant each.

4.5.7 Feedback on our Voice Customization Prototype. At the end of the study, we asked participants to provide feedback on our prototype, which served as a voice searching engine for participants to find a voice they wish to use for each scenario.

In terms of positive feedback, participants like that they have a variety of options to choose from (N = 8); P2 and P11 particularly like that they could specify the age group. Participants also appreciated that they were provided with voice samples that they can listen to before choosing a voice (N = 6); P2 and P9 thought this enabled them to choose a specific voice that they desire without having to guess what the voice would sound like. Some liked the prototype as it helps them to learn what types of voice they prefer (P5) and which attribute to work on to improve the delivery of a voice (P6). P14 valued that the prototype did not require her voice which can reveal her identity.

As for rooms for improvements, five participants requested to have 4 to 5 levels to tune each attribute while we had 3. Five participants wished the prototype to present at least one and a possibly similar number of search results for a specified set of voice attributes. Moreover, some participants did not like having voice samples that had different accents (P4, P5). P4, in particular, suggested using a synthesized voice with less variation in accents. Related, P14 recommended displaying a text label that describes the accent of the voice sample. In addition, two participants thought that voice labels were sometimes inaccurate. Participants also wished for improved audio quality and support for various languages other than English (N = 1 for each). P12 had an interesting suggestion about a feature where a user can specify famous figures (*e.g.*, news anchors, movie stars) by their names and use that voice as a start. She also wished for the prototype to show values for each attribute given a voice so that she can work on training her voice to sound like the voice she desires to have as her own voice.

Meanwhile, serious concern was raised by P6. While she appreciated the voice customization can be used to secure one's identity, she urged that the technique should be carefully considered as it can be used for crimes.

5 DISCUSSION

Here we summarize findings from the study, discuss design recommendations for supporting end-user voice customization, and how it can be implemented in practice.

5.1 The Needs for Voice Customization

Our findings showed the need for voice customization. Most participants had wished for having a different voice or wished to use voice customization regardless of one's satisfaction with their own voice. In addition, we have confirmed that people are willing to customize their voices differently under different circumstances, and their desire for voice customization was the strongest for anonymity for situations when talking to someone unfamiliar. Our in-depth interview revealed that concealing gender and age is the reason for desiring voice customization. Confirming prior findings [28, 40], these behaviors were frequently observed among female participants who enjoy playing online games, where they easily get into troubles with others when their gender is revealed as female. Moreover, all interview participants wished to use voice customization when they need to hide their anger and try to be polite to others (*e.g.* using a calm voice when making complaints). While both survey and interview findings indicate the need for voice customization, the existing support is limited to simple voice filters to play around with friends. We hope that our exploratory work to show the proof-of-concept can urge other engineers and researchers in the related field to get involved in realizing the idea.

5.2 The Challenges and Feasibility for Supporting Voice Customization

We designed and implemented a prototype to better understand the feasibility of supporting end-user voice customization and to elicit participants' in-depth feedback. As a result, we could confirm that our prototype is easy to use and it can enable participants to choose the voice they wish to use instead of their own voice for different circumstances. However, to be able to serve as an end-to-end system, two more modules should be implemented: a voice attribute recognizer and a voice synthesizer. While our current web-based prototype lacks in the number of voice samples, it is impossible to guarantee the diversity of voices with every possible combination of different voice attributes with more levels. Thus, we plan to investigate voice synthesis techniques to augment voice samples. In addition, we also plan to work on modules that can automatically compute each attribute given an input voice, users' own voice in particular, and produce a synthesized voice in real-time as one speaks. Yet, the development of voice synthesizers is still in progress. Several companies offer software programs based on deep learning models for commercial usage. However, they do not support end-user customization where users can choose their own voice by tuning various attributes. Thus, we plan to open our dataset to the public for the future progress of voice synthesizer and customization services.

In addition, there are concerns about abusing the synthesized voice, such as identity theft. In fact, there has been a case of cybercrimes using an AI-based software that imitated the chief executive's voice and demanded remittance. Similar to Deepfake, careful consideration is needed since one can easily deceive others with more realistic synthesized voices.

5.3 Design Recommendations for Voice Customization

One of the objectives of this study is to provide a foundation of for designing an end-user voice customization tool to developers and user experience designers/researchers for platforms or interfaces that supports voice-based communication including synthesized voices. Thus, we present recommendations based on our findings.

5.3.1 Voice Attributes to be Supported. Being able to choose a particular voice with various voice attribute options is desired. To be specific, reflecting the findings from subsubsection 4.5.6, we suggest supporting the following voice attributes as a default since these are considered to be useful for making distinctive changes in voices: *smoothness*, *variation*. In addition, *pitch*, *speed*, and *volume* should be considered as well, not only because these were requested by the participants but also because these are the defining characteristics of a voice [5, 8, 10–12, 25]. In addition, when these attributes are

supported, we recommend supporting multiple levels; more than three if possible.

5.3.2 Use of Hashtags for Different Voice Categories. We recommend providing a list of hashtags that users can choose from as in a prior study that categorizes singing voices using hashtags [18]. Similar to how *gender* and *age group* in our prototype, some of the characteristics more applicable to be considered as nominal data. The use of hashtags would be also useful for describing *emotional states* (*e.g.*, happy, sad) and *personality* (*e.g.*, friendly, stubborn) of a voice; the ones mentioned by the participants. For instance, users can type *#joyful* or *#formal* on the search bar, and then adjust the voice qualities in detail and find the exact voice that they are looking for.

5.3.3 Suggesting Similar Voices but Better. We found that participants have a strong concern about using customized voice when talking to acquaintances such as family members or classmates as they think it is awkward or even rude. Yet, they wish to have a voice that is pleasant to listen to regardless of how satisfied they are with their own voice. This suggests that there is a need for beautifying one's own voice with small changes so that the customized voice is still recognized as their own voice by others instead of an arbitrary voice that sounds much better. Thus, we highly recommend a future end-user voice customization tool to incorporate users' own voice when generating synthesized voices or listing voice candidates that sound like them.

5.3.4 Profiling Context-specific Voices. Our findings showed that the desired voice changes depending on the context. For instance, participants preferred to have high variation in pitch when talking to someone familiar while they wish to have a monotonous voice when talking to a stranger (*i.e.*, customer service operator). Also, they wished to have a clear voice when the content delivery is important such as when taking a class or creating an informative YouTube video. In addition, they valued *smoothness* of a voice for situations where they can be more casual like Holiday Greetings and YouTube Channel scenarios. Meanwhile, participants chose the most neutral voice for playing games. Thus, we suggest supporting users to create and save multiple voice profiles so that they can easily switch to different voices that are appropriate for the given context without having to customize a new voice every time.

5.4 Use of Voice Customization in Practice

Considering the pandemic where most of the physical co-presence is almost impossible, a growing number of people are beginning to socialize with others in a metaverse; the monthly active users of Roblox (roblox.com), one of the metaverse platforms, has increased from 35 to 150 million in three years[1], and since vocal communications are known to bring users closer compared to text-based communications [37, 39], we believe that a voice customization feature can be applied to reduce the disparity between users' own voice and the virtual avatars with different appearances and enhance the level of immersion as a result.

Furthermore, we expect voice customizing services to enable people to have difficulties in vocalizing with their own voice to have a voice of their choice given the contexts instead of relying on a commercial text-to-speech (TTS) synthesizers that sound the same. We plan to further explore the idea of using voice customization for improving the accessibility in voice-based communication.

5.5 Limitations and Future Work

As in other survey and interview studies, ours have several limitations. First of all, the survey results may not be as accurate as the responses were self-reported. In addition, since there were more female participants than male participants for both formative and main studies[5], larger sample size is needed to confirm if ones' unwillingness to reveal their gender is specific to female users. In addition, as the studies were conducted in Korean, our results may not be generalized to users who have different cultural backgrounds. Moreover, the voice data we had for our prototype lacks samples for extreme values and diversity in terms of supported languages. Also, while the annotated voice attributes can be perceived as wrong as they can be subjective. Thus, human verification or auto-labeling would be needed for the attribute values to be objective. Finally, although we presented our prototype as a voice customization tool, its feature was limited to an interactive voice search engine where audio recordings can be filtered in terms of the specified attributes. Future work should investigate an end-to-end system including the real-time conversion of one's own voice into the customized voice.

6 CONCLUSION

To better understand the need and challenges for supporting end-user voice customization, we first conducted an online survey with 104 participants. The results confirmed that there is a demand for changing one's voices under different situations particularly for non-face-to-face conversations at work or at school. We then conducted a semi-structured interview with 14 participants with a prototype that enable participants to search for a specific voice they wish to use. The findings revealed that different voices are favored for different contexts with improved delivery or anonymity. We also identified useful voice attributes for supporting voice customization. While we used voice recordings to demonstrate the idea to participants, we plan to extend this study by investigating voice synthesis techniques for end-users where they can create a new voice or tune their own voice as needed by various voice attributes.

ACKNOWLEDGMENTS

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2020-0-01460) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). Also, this work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF2021R1F1A105278611).

REFERENCES

[1] [n.d.]. Business of Apps Roblox Statistics. https://www.businessofapps.com/data/roblox-statistics/. Accessed: 2022-01-06.
[2] David W Addington. 1968. The relationship of selected vocal characteristics to personality perception. (1968).
[3] Max V Birk and Regan L Mandryk. 2018. Combating attrition in digital self-improvement programs using avatar customization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–15.

[5]Note that unbalanced male-to-female ratio could be due to the high participation from the researchers' institution, which is a women's college.

[4] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[5] Raymond H Colton and Jo A Estill. 1981. Elements of voice quality: perceptual, acoustic, and physiologic aspects. In *Speech and Language*. Vol. 5. Elsevier, 311–403.

[6] Igor Dolgov, William J Graves, Matthew R Nearents, Jeremy D Schwark, and C Brooks Volkman. 2014. Effects of cooperative gaming and avatar customization on subsequent spontaneous helping behavior. *Computers in human behavior* 33 (2014), 49–55.

[7] Michelle Drouin, Daniel Miller, Shaun MJ Wehle, and Elisa Hernandez. 2016. Why do people lie online?"Because everyone lies on the internet". *Computers in Human Behavior* 64 (2016), 134–142.

[8] T Ehrette, N Chateau, Christophe d'Alessandro, and V Maffiolo. 2002. Prosodic parameters of perceived emotions in vocal server voices. In *Speech Prosody 2002, International Conference*.

[9] Jerry Bryan Fuller, Tim Barnett, Kim Hester, Clint Relyea, and Len Frey. 2007. An exploratory examination of voice behavior from an impression management perspective. *Journal of Managerial Issues* (2007), 134–151.

[10] Marylou Pausewang Gelfer. 1988. Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice* 2, 4 (1988), 320–326.

[11] Marylou Pausewang Gelfer. 1993. A multidimensional scaling study of voice quality in females. *Phonetica* 50, 1 (1993), 15–27.

[12] Christer Gobl and Ailbhe Ní Chasaide. 2000. Testing affective correlates of voice quality through analysis and resynthesis. In *ISCA tutorial and research workshop (ITRW) on Speech and Emotion*.

[13] Mark Grimshaw. 2014. *The Oxford handbook of virtuality*. Oxford University Press.

[14] Rosalie Hooi and Hichang Cho. 2014. Avatar-driven self-disclosure: The virtual me is the actual me. *Computers in Human Behavior* 39 (2014), 20–28.

[15] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.

[16] Yukara Ikemiya, Katsutoshi Itoyama, and Hiroshi G Okuno. 2014. Transferring vocal expression of f0 contour using singing voice synthesizer. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 250–259.

[17] Mohammad Muzammil Khan and Anam Saiyeda. 2021. Reader: Speech Synthesizer and Speech Recognizer. In *International Conference on Innovative Computing and Communications*. Springer, 877–886.

[18] Keunhyoung Luke Kim, Jongpil Lee, Sangeun Kum, Chae Lin Park, and Juhan Nam. 2020. Semantic Tagging of Singing Voices in Popular Music Recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 1656–1668.

[19] 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 (2007), 310–329.

[20] Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 289–296.

[21] 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 (2003), 419.

[22] Eman Mubarak, Tooba Shahid, Maryam Mustafa, and Mustafa Naseem. 2020. Does Gender and Accent of Voice Matter? An Interactive Voice Response (IVR) experiment. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–5.

[23] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics* 5, 2 (2013), 171–191.

[24] Marilena Phillips. 2017. Talking the talk: The effect of vocalics in an interview. (2017).

[25] Klaus R Scherer. 1978. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology* 8, 4 (1978), 467–487.

[26] Eliane Schlemmer, Daiana Trein, and Cristoffer Oliveira. 2009. The metaverse: Telepresence in 3D avatar-driven digital-virtual worlds. *@ tic. revista d'innovació educativa* 2 (2009), 26–32.

[27] Richard J Sebastian and Ellen Bouchard Ryan. 2018. Speech cues and social evaluation: Markers of ethnicity, social class, and age. In *Recent advances in language, communication, and social psychology*. Routledge, 112–143.

[28] Stefano Taddei and Bastianina Contena. 2013. Privacy, trust and control: Which relationships with online self-disclosure? *Computers in human behavior* 29, 3 (2013), 821–826.

[29] 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 (2011), 253–262.

[30] Sonja A Trent. 1995. Voice quality: Listener identification of African-American versus Caucasian speakers. *The Journal of the Acoustical Society of America* 98, 5 (1995), 2936–2936.

[31] Stefano Triberti, Ilaria Durosini, Filippo Aschieri, Daniela Villani, and Giuseppe Riva. 2017. A frame effect in avatar customization: How users' attitudes towards their avatars may change depending on virtual context. (2017).

[32] Selen Turkay. 2012. User experiences with avatar customization in Second Life and Lord of the Rings Online. In *Proceedings on Teachers College Educational Technology Conference*. Citeseer, 78–79.

[33] Selen Turkay and Charles K Kinzer. 2015. The effects of avatar-based customization on player identification. In *Gamification: Concepts, methodologies, tools, and applications*. IGI Global, 247–272.

[34] Kyle James Tusing and James Price Dillard. 2000. The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research* 26, 1 (2000), 148–171.

[35] Asimina Vasalou and Adam N Joinson. 2009. Me, myself and I: The role of interactional context on self-presentation through avatars. *Computers in human behavior* 25, 2 (2009), 510–520.

[36] Daniela Villani, Elena Gatti, Stefano Triberti, Emanuela Confalonieri, and Giuseppe Riva. 2016. Exploration of virtual body-representation in adolescence: the role of age and sex in avatar customization. *SpringerPlus* 5, 1 (2016), 1–13.

[37] Greg Wadley, Marcus Carter, and Martin Gibbs. 2015. Voice in virtual worlds: The design, use, and influence of voice chat in online play. *Human–Computer Interaction* 30, 3-4 (2015), 336–365.

[38] Steven Weinberger. 2015. Speech accent archive. george mason university. *Online:< http://accent. gmu. edu* (2015).

[39] Dmitri Williams, Scott Caplan, and Li Xiong. 2007. Can you hear me now? The impact of voice in an online gaming community. *Human communication research* 33, 4 (2007), 427–449.

[40] Kuang-Wen Wu, Shaio Yan Huang, David C Yen, and Irina Popova. 2012. The effect of online privacy policy on consumer privacy concern and trust. *Computers in human behavior* 28, 3 (2012), 889–897.

[41] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[42] Zhaoyan Zhang. 2021. Voice Feature Selection to Improve Performance of Machine Learning Models for Voice Production Inversion. *Journal of Voice* (2021).