Voice Creator: Giving Customized Voice to the Voiceless for Online Communication

Hyeon Jeong Byeon
Ewha Womans University, Seoul, Republic of Korea
cat@ewhain.net

ABSTRACT

Voice plays an important role in online communications by increasing intimacy among people. Despite the advantages of using voice in computer-mediated communications (CMC), it is difficult for people with speech or hearing impairments to participate in such communication methods. In this study, we investigate different attributes of voices and how it affects users' preference. We also deployed a website called 'Voice Creator' for people who want to create an online voice by specifying the levels of different voice attributes: gender, age group, breathiness, smoothness, hoarseness, and variation. We plan to conduct a user study on the target users and study the behavior of voice customization in future work.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility**; **Accessibility systems and tools**; • **Human-centered computing** → **Accessibility**; **Accessibility technologies**.

KEYWORDS

speaking impairment, hearing impairment, voice customization, computer mediated communication

ACM Reference Format:
Hyeon Jeong Byeon. 2021. Voice Creator: Giving Customized Voice to the Voiceless for Online Communication. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3441852.3476476

1 INTRODUCTION AND RELATED WORKS

Several studies have proven that voice plays an important role in online communications [1, 2]. For instance, Dmitri et al. [1] have demonstrated that voice communications can lead to stronger relationships and increase bonding among online gaming community users than text-only communications. Despite the advantages of using voice in computer-mediated communication (CMC), it is difficult for people with speech- or hearing- impairments to participate in such communication methods. As Kuligowska et al. [3] has discussed, the present speech synthesizers have several limitations that obstruct synchronous communications: lack of emotions and prosody, low adaptation to the situation, and low authenticity compared to natural speech. These restraints make present speech synthesizers insufficient to meet the demands of the users who are seeking intimate communication. Furthermore, studies show that users tend to favor synthetic voices that sound like their personality traits due to the similarity-effect [4]. Also, the success of voice customization companies for people with speech disorders such as Vocal ID[1], Soundthinkers[2], and my-own-voice[3] shows that there is significant demand for customizing synthetic voices. Therefore, in this paper, we investigated the different attributes of voices and found out how each affects voice preference. Also, we proposed a website with an intuitive prototype design called 'Voice Creator' for people who want to create an online voice by customizing six voice features: gender, age group, breathiness, smoothness, hoarseness, and variation.

First, we defined the voice features by gathering the popular terms among preceding studies. Then, we conducted a survey asking the respondents to listen to a short recording and rate it based on the voice features. Based on the survey results, we gathered a labeled audio dataset. Finally, we deployed the Voice Creator, a website for customizing voices.

2 METHOD: DATA COLLECTION

To create a labeled audio dataset for voice customization, we conducted an online survey using the crowdsourcing platform Amazon Mechanical Turk (MTurk). Several studies have conducted a survey to ask respondents to evaluate the voice samples [5–10]. For instance, Colton et al. [9] defined the voice qualities in five dimensions: activity, purity, brightness, youthfulness, and texture. We gathered the general terms of vocal qualities and chose the seven most frequently used terms: breathiness, hoarseness, pitch, precision, speed, variation, and volume.

Next, we used the speech accent archive, proposed by Weinberger et al. [11]. It is an audio dataset recorded by 2140 different native and non-native speakers of English. All the speakers read the same text written in common English words. The dataset is labeled by the speaker's information: birthplace, native language, age, and gender. We used this information to design the Voice Creator in Section 3.

We asked the MTurk respondents to listen to a voice sample of the speech accent archive and mark each voice feature. We used the seven voice quality terms and added the term 'preference' to measure the pleasantness of the voice. Figure 1 shows a screenshot of the task we requested from the crowd workers. We minimalized texts on the survey by directly placing the two opposite values next to the slider. In consequence, the workers can mark the voice features by simply moving the slider to the appropriate value on a scale from 1 to 5.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ASSETS '21, October 18–22, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8306-6/21/10.
https://doi.org/10.1145/3441852.3476476

[1] https://vocalid.ai/
[2] https://soundthinkers.co/en/
[3] https://mov.acapela-group.com/

Figure 1: A screenshot of the task to collect voice labels from Amazon Mechanical Turk (MTurk)

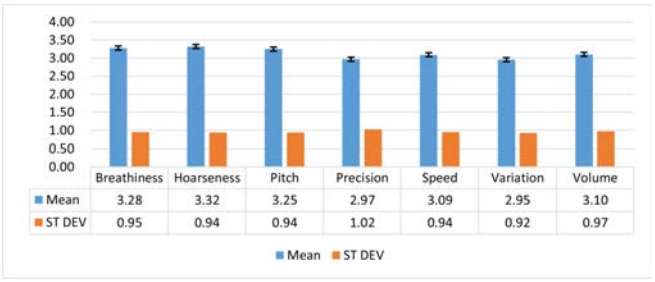

Figure 3: A chart of mean and standard deviation for each voice quality

3 SYSTEM: VOICE CREATOR

Based on the labeled audio data gathered from the crowd workers, we designed the Voice Creator, a voice customizing web application. As seen in Figure 2, the user can customize the voice through the voice element control bar. There are a total of six vocal option settings: gender, age group, breathiness, smoothness, hoarseness, and variation. The user first decides the gender and age group of the voice he/she will customize. Then, the user moves the slider to set the value for the remaining of four voice qualities by adjusting a value between 1 to 3. We left out the volume and speed factor from the controllable attributes because these features are quantitative and easily adjustable using simple techniques.

After setting the values, the user can click the 'Search' button to find the voice samples that satisfy the settings. The voice samples appear below the 'Search' button. The user can click on the 'play' button on the audio tag and listen to the generated output saying the sample phrase, "Please call Stella. Ask her to bring these things with her from the store." [11] The generated outputs are voice samples which were given the highest ratings by the crowd workers. The entire code[4] and website[5] are accessible via URL link.

Figure 2: A screenshot of the Voice Creator website

4 RESULTS

4.1 Preferred voices and their attributes

In total, 2176 speakers recorded the speech accent archive [11]. We requested three different crowd workers to rate each recording in eight attributes, as shown in Figure 1 in Section 2. Altogether, we gathered 6528 responses. Then we sorted each recording in a preference order by finding the average of the 'preference' value score. The top 27 recordings were rated over 4.5 points out of 5. The average preference of the recordings was 3.3262 points out of 5. Among the most preferred 27 voices, 8 were male voices, and 19 were female voices. 13 speakers were native speakers of English, and 14 speakers were non-native. Also, 12 speakers were in their 20's, and 7 speakers were in their 30's. In conclusion, the survey results indicate that female voices are preferable to male voices; the respondents preferred speakers in their 20's and 30's; being a native or a non-native speaker of English does not affect the preference.

4.2 The feasibility of voice attribute annotation by crowd workers

To summarize the results on the agreement of voice attributes, we calculated the standard deviation of the scores rated by three different crowd workers. Figure 3 shows the mean and standard deviation of each voice quality. Results show that 'precision' and 'volume' have larger values than the other voice quality attributes. This means that the respondent's ratings diverge most on the two attributes. However, the overall standard deviation is somewhat small enough to conclude that collecting voice attributes from crowd workers is feasible.

5 CONCLUSION AND FUTURE WORK

In this paper, we gathered a labeled audio dataset using the crowdsourcing platform Amazon Mechanical Turk (MTurk). We defined the seven voice attributes by determining the commonly used terms among papers and asked the respondents to rate the voice samples of the speech accent archive [11]. From the results, we found out the attributes that affect the preference of a voice the most, such as gender and age. Also, we discussed the feasibility of voice attribute annotation by calculating the standard deviation of the scores. The labeled audio dataset can be used for researchers who seek to create a customized voice considering people's preferences.

While the Voice Creator website is deployed and ready for usage, we were not able to interview the target users – individuals who have a speech or hearing disorder. Therefore, as a future task, we plan to conduct a user study on the target users and find out the preferred voices via interviews. Also, we plan to ask the crowd workers to rate the intelligibility factor of the speech accent archive [11] dataset, as this can be an important factor that affects the preference. Lastly, we plan to present the inter-rater reliability metric and precisely show the reliability of crowd sourced dataset.

[4] https://github.com/hyeonJeongByeon/VoiceCreator
[5] http://ec2-3-36-91-129.ap-northeast-2.compute.amazonaws.com:7777/

REFERENCES

[1] Williams Dmitri, Scott Caplan, and Li Xiong. 2004. "Can you hear me now? The impact of voice in an online gaming community." *Human communication research* 33.4, 427-449.

[2] 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, 336-365.

[3] Karolina Kuligowska, PawełKisielewicz, and Aleksandra Włodarz. 2018. "Speech synthesis systems: disadvantages and limitations." *Int J Res Eng Technol (UAE)* 7, 234-239.

[4] Matthew P. Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. "Speech synthesis for the generation of artificial personality." *IEEE transactions on affective computing* 11.2, 361-372.

[5] Klaus R. Scherer. 1978. "Personality inference from voice quality: The loud voice of extroversion." *European Journal of Social Psychology* 8.4, 467-487.

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

[7] 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.*

[8] Marylou Pausewang Gelfer. 1993. "A multidimensional scaling study of voice quality in females." *Phonetica* 50.1, 15-27.

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

[10] Marylou Pausewang Gelfer. 1988. "Perceptual attributes of voice: Development and use of rating scales." *Journal of Voice* 2, no. 4, 320-326.

[11] Weinberger, Steven. 2015. Speech Accent Archive. George Mason University. Retrieved from http://accent.gmu.edu