Voice Creator: 개인 맞춤형 목소리 생성 웹 어플리케이션 프로토타입

변현정*, 여수현*, 오유란**
*이화여자대학교 컴퓨터공학과
**교신저자, 이화여자대학교 컴퓨터공학과
cat@ewhain.net, yeo7764@ewhain.net, uran.oh@ewha.ac.kr

Voice Creator: A Vocal Customization Web Application Prototype

Hyeon Jeong Byeon*, Soohyun Yeo*, Uran Oh**
*Dept. of Computer Science and Engineering, Ewha Womans Univeristy
**Corresponding Author, Dept. of Computer Science and Engineering, Ewha Womans University

Abstract

Due to the important role of avatars in computer-mediated communication (CMC), a growing number of CMC-based services now support avatar customization options. However, in many cases, customization and personalization options are limited to visual features. In this paper, we propose and describe a prototype for a vocal customization web application. Titled Voice Creator, the app is designed for both able-bodied and speech- or hearing-impaired users who seek to communicate anonymously using digital voice identities.

1. Introduction

Avatars play an important role in constructing an identity in computer-mediated communication (CMC). They protect users' privacy when necessary and allow expressive freedom in anonymous online situations [1]. Therefore, many CMC-based services offer avatar customization options: to give a few examples, iPhone MEmoticons[1], computer games, social Virtual Reality, and many other online websites such as the Avatar maker[2]. However, we note that a majority of these services focus on only the visual features of the avatar. As a result, these avatars are unable to protect user privacy in synchronous communications contexts such as video chats and FaceTime audio calls - few offer the option of making alterations to the users' original voice information. Inspired by the fact that little has known about services that create customized voices, we designed an End-to-end model-based voice customizing web application prototype.

2. Related Works

2.1 The UX of Avatar Customization

Avatars are graphical representations of the user in virtual environments such as games. Many games let users customize their avatars using interfaces, referred to as Character Creation Interfaces or CCIs. Concerning the complexity and presentation of personalized options, CCIs differ greatly. McArthur [2] focused on User Experience (UX) in games that allow users to customize their avatar. According to this paper, interface widgets, additional options via sub-menus, or long lists of customization options to scroll can impact the user's self-representation in Massively Multiplayer Online games (MMOGs).

In our prototype, we targeted users seeking to communicate with other users in a virtual environment. We focused on providing an intuitive design for users to experience all of the options provided.

2.2 End-to-end Text-to-Speech Synthesis Model

The End-to-End speech synthesis model uses pairs of input text and raw spectrogram output to learn Deep Neural Network models. By minimizing the complexity of modern TTS designs, it uses a simple waveform synthesis technique. Wang et al. [3] proposed Tacotron, an End-to-End TTS model. Tacotron outperforms in terms of naturalness. "Deep Voice," proposed by Arik et al. [4], is another example of an End-to-End TTS synthesis model.

We focused on the advantage that Tacotron more easily allows various attributes such as speaker or language and even high-level features like the sentiment. We believe that voice identification features such as thickness, strength, pitch, accent, speed and pause could also become an attribute to

[1] https://support.apple.com/en-us/HT208986

[2] https://avatarmaker.com/

conditioning Tacotron. We referred to Voquent[3], an online voice-over service to decide the voice identification features.

3. Voice Creator

3.1 Web Application Prototype

We designed the prototype of the Voice Creator web application using Figma[4]. The web page consists of three parts: First, there is an introduction part on this homepage, and secondly, there is a part where users can create their voice and listen to the created voice. Finally, there is a part where users can see the characteristics of the voice that users saved and hear it again.

3.1.1 Introduction section

As seen in Figure 1, The introduction part contains the introduction and usage of this page.

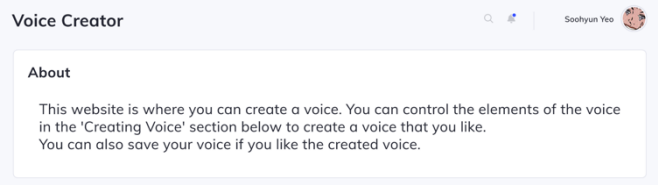

Figure 1. Introduction part of the Voice Creator

3.1.2 "Create Voice" section

In the "Create Voice" section, the user can create a voice directly through the voice element control bar. As seen in Figure 2, there are six vocal elements: thickness, strength, pitch, accent, speed and pause. Users can set the value of the "thickness" element by entering a number between 1 -100, 1 standing for most thin and 100 standing most thick. The "strength" element can be adjusted by a number between 1-100, 1 standing for the weakest, and 100 standing for the strongest. Pitch refers to the audio frequency of the voice. Users can set the value of the "pitch" element by entering a number between 1-100, 1 standing for the lowest, and 100 standing for the highest. Users can set the value of the "accent" element by setting a number between 1-100, 1 standing for most monotonic, and 100 standing for a strongest accent. The "speed" element of the voice can be adjusted by a number between 1-100, 1 standing for the slowest, and 100 standing for the fastest. Lastly, the users can set the value of the "pause" element by setting a number between 1-100, 1 standing for the least pauses, and 100 standing for the maximum pauses.

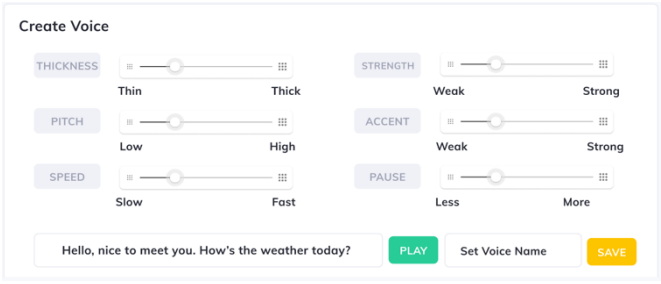

Figure 2. Create Voice part of the Voice Creator

After specifying the elements of the voice, the user can click the 'PLAY' button at the bottom to hear a sample clip of the generated output saying the sample phrase, "Hello, nice to meet you. How's the weather today?" If the user is satisfied with the generated voice output and wants to save it, the user records the name of the voice and presses the 'SAVE' button.

3.1.3 Collection section

The collection part gathers the voices saved in the "Create Voice" part. Users can listen to the customized voices and check the value of the voice components. As seen in Figure 3, you can visually check the vocal waves.

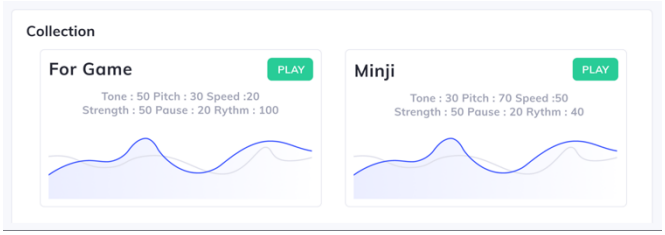

Figure 3. Collection part of the Voice Creator

3.2 Speech synthesizer model

We believe that an End-to-End speech synthesizer model is appropriate for the Voice Creator. The Emotional End-to-End Neural Speech synthesizer proposed by Lee et al. [5] outperforms the naturalness of speech in various situations. In this work, the authors used emotion and intensity as a variable. We suggest an End-to-End Neural Speech synthesizer that customizes voice features by six voice identification features which are thickness, strength, pitch, accent, speed and pause.

4. Conclusion

We designed the prototype of Voice Creator, a web application that allows users to create a customized digital voice. Unlike other voice synthesizer services, Voice Creator enables users to create customized voices into six features – thickness, strength, pitch, accent, speed and pause, and let users customize the voice precisely. Voice Creator can meet

[3] https://www.voquent.com/the-vocal-characteristics-that-speak-to-your-character/

[4] https://www.figma.com/

the demands of those users looking to create and use a computer-generated voice for computer-mediated communication. A potential user base would be people who want to communicate anonymously on the voice-based social network service Clubhouse[5]. Also, it will be useful for people with hearing or speech impairments who want to make a new digital voice to express themselves in a virtual environment.

Our future tasks are to implement the voice customizing model and to distribute the Voice Creator web application using the End-to-End Neural speech synthesizer model and Amazon Web services.

Reference

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

[2] McArthur, V., The UX of avatar customization. In Proceedings of the 2017 CHI conference on human factors in computing systems, 5029-5033, 2017.

[3] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S. and Le, Q., Tacotron: Towards end-to-end speech synthesis. arXiv preprint: 1703.10135, 2017.

[4] Arik, S.Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J. and Sengupta, S., Deep voice: Real-time neural text-to-speech. International Conference on Machine Learning, PMLR, 2017, 195-204.

[5] Lee, Y., Rabiee, A., and Lee, S.Y., Emotional end-to-end neural speech synthesizer. arXiv preprint: 1711.05447, 2017.

[5] https://apps.apple.com/us/app/clubhouse-drop-in-audio-chat/id1503133294