PROJECT SPECIFICATION FOR CSC 475 2025

Isaiah Doyle
University of Victoria
isaiahdoyle@uvic.ca

Aileen Klassen
University of Victoria
aileenklassen@uvic.ca

Elijah Larmer
University of Victoria
elijahlarmer@uvic.ca

ABSTRACT

The state of programs designed to support voice training are lacking with respect to timbral nuance. For trans people undergoing voice training, available applications tend to favour pitch as the primary – or in some cases sole – measure of progress. To accommodate the many parameters that factor into voice perception, we propose a tool to allow users to mimic a resynthesized version of their voice using applied timbral descriptors.

1. INTRODUCTION

This project is a component of a larger application that supports people undergoing voice training in developing their preferred vocal timbre by mimicking a synthesized (or otherwise modified) version of their voice with any desired timbral modifications. Current applications (DevExtras, 2018; Seek & Nitz, 2020; Antoni & Speechtools Ltd., 2013) rely largely on pitch to distinguish vocal characteristics. The human voice is far more nuanced than this, and as such timbral development of those undergoing voice training is paramount to the users' success (Hawley & Hancock, 2024). Semantic timbral descriptors (e.g., breathier, huskier, higher) will be given by the user to apply to a recording of their voice, and the resulting output can be tweaked further using additional descriptors. By using the user's voice as the primary input source, our goal is to promote a healthy and informed way to explore the timbral possibilities of one's voice. Analysis of the user's voice will be detailed with respect to a number of timbral descriptors and made available through an accessible user interface, such that the program can be used as a tool for speech pathologists and the public (e.g., trans people seeking a more feminine/masculine/neutral voice, people with speech impairments (Barkmeier-Kraemer et. al), voice actors) alike.

2. METHODOLOGY

We plan on using Python as our coding language for this project. We anticipate using libraries like Coqui TTS, Parselmouth (Jadoul., Boer, & Ravignani, 2024), and Librosa to support vocal analysis and synthesis steps. For testing, we plan on using our own voices and possibly the voices of other volunteers. Any training will utilize publicly available datasets like Mozilla's (2024) Common Voice among others (Schwoebel, 2021). Linear predictive coding (Kim; Kunigami) and/or PRAAT (Magdin et. al., 2019) synthesis will likely be a significant component of the project as a way to reproduce the user's voice. From there, we plan to explore ways the user will be able to interface with the program in order to adjust their voice quality in real time.

3. TIMELINE

There will be a significant learning curve in implementing this project, so the precise details are subject to change throughout the term. The general timeline resembles the following:

1. Before commencing the project, it's important that all members are on the same page and agree to and understand the project details and distribution of work. Amendments should be made to this timeline as research progresses and implementation details are decided on.

2. The first major component of the project is to effectively clone a vocal sample using either linear predictive coding or a text-to-speech algorithm. Using existing resources as reference material, we expect to have a working copy by early March.

    - Feb. 25: consider best speech synthesis algorithm for our use case (Aileen)
    - Mar. 7: implement the algorithm to reproduce a target voice (Elijah)

3. The next step involves training a model to interpret timbral descriptors with respect to voice. This will include deciding on whether to support a discrete number of descriptors, and labelling dataset samples

Attribution: Doyle, I., Klassen, A., & Larmer, E., "Project Specification for CSC 475 2025", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

with those descriptors. Alternatively, some descriptors can be trained by applying DSP-based effects (e.g., a breathiness effect) to the dataset. This should be ready by mid-March.

- Feb. 28: determine which descriptors to use, if a restricted set (Isaiah)
- Mar. 11: generate dataset with descriptor labels (Aileen, Elijah, Isaiah)

4. With a replica of the user's voice and a dataset populated with timbral descriptors, use input descriptors to favour particular voices in the dataset to resynthesize the user's voice.

   - Mar. 25: implement vocal resynthesis favouring dataset voices corresponding to given timbral descriptors (Isaiah)
   - Apr. 4: create a minimal UI (Aileen)
   - Apr. 4: final testing (Aileen, Elijah, Isaiah)
   - Apr. 11: final adjustments (Elijah)

4. REFERENCES

[1] Antoni, C., Speechtools Ltd. (2013). ChristellaVoiceUp. Accessed: Feb. 18, 2025 [Online.] Available: https://www.christellaantoni.co.uk/transgender-voice/voiceupapp/

[2] Alter, I. L., Chadwick, K. A., Andreadis, K., Coleman, R., Pitti, M., Ezell, J.M., Rameau, A. "Developing a mobile application for gender-affirming voice training: A community-engaged approach" Laryngoscope Investig Otolaryngol. 2024. DOI: 10.1002/lio2.70043 Accessed: Feb. 13, 2025. [Online.] Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11645500

[3] Barkmeier-Kraemer, J. M., Craig, J. N., Harmon, A. B., Hillman, R. R., Jacobson, J., Patel, R. R., Ruddy, B. H., Stemple, J. C., Sumida, Y. A., Tanner, K., Theis, S. M., van Mersbergen, M. R., & Verdun, L. P. "Voice Disorders." asha.org. Accessed: Feb. 12, 2025. [Online.] Available: https://www.asha.org/practice-portal/clinical-topics/voice-disorders

[4] DevExtras. (2018). Voice Tools. Accessed: Feb. 18, 2025. [Online.] Available: https://devextras.com/voicetools/

[5] Hawley, J. L. & Hancock, A. B. "Incorporating Mobile App Technology in Voice Modification Protocol for Transgender Women." Journal of Voice. 2024. DOI: 10.1016/j.jvoice.2021.09.001. Accessed: Feb. 11, 2025. [Online.] Available: https://www.sciencedirect.com/science/article/abs/pii/S089219972100299X

[6] Jadoul, Y., Boer, B. & Ravignani, A. "Parselmouth for bioacoustics: automated acoustic analysis in Python." The International Journal of Animal Sound and its Recording. 2024. DOI: 10.1080/09524622.2023.2259327. Accessed: Feb. 11, 2025. [Online.] Available: https://iris.uniroma1.it/retrieve/ad22f915-c173-46fb-8c71-5f635b0f16f3/Jadoul_etal2024_Parselmouth%20for%20bioacoustics%20%20automated%20acoustic%20analysis%20in%20Python-3.pdf

[7] Kim, H. "Linear Predictive Coding is All-Pole Resonance Modeling." ccrma.stanford.edu. Accessed Feb. 12, 2025. Available: https://ccrma.stanford.edu/~hskim08/lpc

[8] Kunigami, G. "Linear Predictive Coding in Python." kuniga.ma. Accessed: Feb, 2025. [Online.] Available: https://www.kuniga.me/blog/2021/05/13/lpc-in-python.html

[9] Magdin, M., Sulka, T., Tomanová, J., & Vozar, M. "Voice Analysis Using PRAAT Software and Classification of User Emotional State." International Journal of Interactive Multimedia and Artificial Intelligence. 2019. DOI:10.9781/ijimai.2019.03.004. Accessed: Feb. 10, 2025. [Online.] Available: https://www.researchgate.net/publication/331881418_Voice_Analysis_Using_PRAAT_Software_and_Classification_of_User_Emotional_State

[10] Mozilla. (2024). Common Voice. Accessed: Feb. 18, 2025. [Online.] Available: https://commonvoice.mozilla.org/en/datasets

[11] Schwoebel, J. (2021). Voice Datasets. Accessed: Feb. 18, 2025. [Online.] Available: https://commonvoice.mozilla.org/en/datasets

[12] Seek, D. & Nitz, C. (2020). Voice Pitch Analyzer. Accessed: Feb. 18, 2025. [Online.] Available: https://voicepitchanalyzer.app