Wordification — Dialectical Speech Synthesis Using Machine Learning Systems (Milestone Report 1)

Shashank Comandur
Clay Crews

October 5, 2023

1 Introduction & Problem Statement

Wordification is a research project at the University of South Carolina (mentored by Dr. Stanley Dubinsky and Dr. Homayoun Valafar) that both team members are currently developers on. Its goal is to gamify the teaching of the English language to young children, utilizing the linguistic properties of words to instruct in more effective ways. Computerized spelling instruction has proven to be far more effective for young children than rote memorization of words, and this is especially true for speakers of regional dialects, such as African American English (AAE), or Southern White English (SWE).

2 Technical Approach

The games, in their current form, involve the use of spoken audio of English words. The goal of the project is to address how machine learning systems can be used to create the game audio of spoken words in different dialects, such as AAE or SWE. There are already tools that exist to achieve this goal: see resemble.ai, or speechify.com. However, these tools are proprietary and generally cost a fair share to generate audio at the scale we wish — the database of words utilized by the app contains ∼850 entries, and we wish to generate audio of each word, as well as 2-4 sample sentences for each, thus necessitating the use of a custom model. We plan to read up on how the algorithms behind the aforementioned tools work to see what we can learn, and have also spent time examining the existing published literature surrounding machine learning systems for audio generation.

3 Preliminary Results

After searching on GitHub for existing repositories to aid in the development of our own model, we have come across a few that seem to be promising. We have attempted to run code from these repositories on the class provided cloud compute resource, Chameleon, but we have run into issues regarding the leasing functionality and reserving GPU resources necessary to train the model. As a result, we have tried to use Google Colab in addition to Chameleon, which has yielded more promising results.

One of the more interesting repositories we have explored and run on Colab is an advanced TTS generation library: Coqui-TTS. We sourced some test audio from YouTube to evaluate its performance. For the Southern White English (SWE) dialect, we chose audio from "Southern Fried True Crime", a podcast with a narrator who speaks the dialect. Podcasts are prime candidates for test audio samples in speech synthesis, because they are highly available on the Internet and contain concentrated audio samples of a speaker, typically in a high-fidelity recording environment. The generated audio is attached to our GitHub repository, and we have performed a cursory evaluation on it.

The audio contains blemishes, and is not as natural as we would like it to be, but we believe that this can be attributed to the inputs that we fed the model. The audio itself was not vetted all the way through, and we believe we could obtain better results by cherry picking for certain words and trying to modify the audio to exclude noise (background / environment noises, music, etc.). Because plan to sit down and record the final input audio samples with speakers ourselves, this will not be a problem for our final results, but we wish to see how polished an output the test models can produce.

Additionally, we seek to perform a final analysis of several different implementations, compare their methodologies, and their outputs to evaluate what kind of algorithms are best for dialectical speech synthesis. We will perform a comparative analysis of the model output against the test data and evaluate its effectiveness for use in the games.