Cleft Palate: Data Generation Plan & Strategy

Background & Goal

Synthetic Data Generation can be used to introduce additional variability in the training data sample. In dealing with audio voice data, voices can be cloned/modified through GANs and Autoencoders. For the cleft palate project, the synthetically generated data should have significant variation to the original data, therefore reducing overfit by the model.

Two Approaches Identified

Approach 1: Using direct voice cloning of non-pathological and pathological voices separately, generate cloned voices. Then, have the cloned voices produce sentences of lengthy nature. Such would introduce additional variability as we would have more complete data on pathological/non-pathological voices. In holistic terms, this approach would have voices very similar to the original ones say more words.

Approach 2: Using voice cloning, fine-tuning, data augmentation, feature extraction generate different synthetic voices from the original data. The approach would require to have a model learn differing pathological tendencies present in the voices to recreate voices representative of a particular spectrum of pathological/non-pathological voices. Approach is more complex and may require a doctor to label the generated voices as pathological or not.

Plan

1. Attempt a model: GAN, variable autoencoder
2. Preprocess the data: separation, formatting,
3. *(Only for Approach 2)* Use an attention based neural network to identify particular patterns present in pathological/non-pathological speakers.
4. Train chosen model
5. Retrieve synthetic voices, have them labelled if necessary.
6. Collect summary statistics, voice metrics to verify the correctness of the results
7. Re-train the whisper model with the added synthetic training data and re-evaluate its performance.

Prior Research

While our study is novel in using Generative AI in speech pathology and applying synthetic data generation, prior studies show the success of machine learning models in tangential fields. I identified generative adversarial networks and variable autoencoders as being the most successful models in recent literature. The following literature review co-written by the department of Pathology of Stanford summarizes the information well: [Machine Learning for Synthetic Data Generation: A Review](https://ar5iv.labs.arxiv.org/html/2302.04062v6)