Machine Learning for Speech Error Detection

Joy Liu, 2019 Science Internship Program at Univ of California Santa Cruz

ABSTRACT

This report summarizes my summer project at UC Santa Cruz Science Internship Program (SIP), joint work with fellow student Davin Nguyen, under the guidance of mentors Breanna Baltaxe-Admony (PhD candidate) and Prof. Sri Kurniawan. In this project, I have worked on building a Machine Learning method to segment speech signals to detect speech errors. This effort is part of SpokeIt, a speech therapy software to help children with cleft problems.

1 BACKGROUND

Cleft lip and cleft palate are birth defect caused by mouth developing improperly during pregnancy. Tissues from side of mouth do not fuse properly in the center, causing a split or opening in the lip or palate. Cleft lip/palate are fairly common, affecting 1 in 700 babies globally. Children with cleft may suffer multiple problems, one of which is the difficulty in speech. Voices get distorted and take on a nasal sound, making speech difficult to understand.

Treatment of cleft lip/palate is a lengthy and expensive process. It requires multiple surgeries spanning from infancy to early adulthood. Many children have to live with cleft for years while their families save up money for treatment. UC Santa Cruz's ASSIST Lab provides SpokeIt (https://spokeitthegame.com), an online speech therapy software teaching cleft children to speak. It is an interactive game, in which a child is prompted to make a certain sound, and speech recognition technologies are used to decide whether the sound is correct or incorrect. By providing the real-time feedback, it guides the child to make adjustments towards more comprehensible speech. This offers a solution for the children to live a normal life, at least on the speech side.

2 PROBLEM STATEMENT

SpokeIt's main focus is on speech error detection. For cleft speeches, a few types of error are common. For instance, high pressure sounds such as "/s" are usually hard to make. A more prominent error is resonance – the soft palate may not close properly, resulting in hyper-nasality sound. Currently SpokeIt's error detection mechanism works by first feeding the speech signal to a speech-to-text mechanism and then checking the resulting text for misspellings. While this is a quick-to-implement solution, it leaves some space for accuracy improvement.

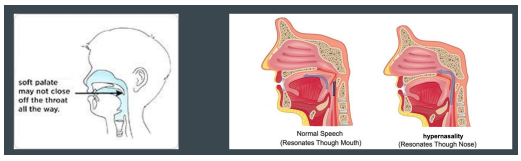

Figure 1: Resonance errors: hypernasality and hyponasality

My project is a piece fitting into SpokeIt, focusing on speech segmentation. This problem is due to the nature of the speech signal itself. Fig. 2 shows an example recording, color-coded into different segments: red is the signal from the therapist, blue is from the child, green is background noise, and gray is silence. Obviously we only want to detect error in the child speech, and skip the therapist, background noise, and silence periods. This calls for a machine learning scheme to separate out the child voice segments.

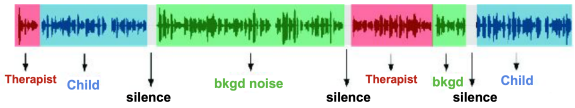

Figure 2: Speech signal (an example)

3 METHOD

For the speech segmentation problem, we work on a data set containing 150,000 speech recordings collected primarily from India and the US, in collaboration with Smile Train, the world's largest cleft organization.

Rather than building our segmentation scheme from scratch, we would like to reuse state-of-the-art tools (e.g., python ML packages) as much as possible. To start with, we have used `librosa` to label silence periods (the grey chunks in the figure). Background noise is also relatively easy to separate out because background noise has very different spectrum characteristics than human voice. The real challenge is the segmentation of therapist voice and child voice. They are both human voice, making them hard to separate. For this task, we use the python machine learning package `sklearn` and its K-means algorithm.

K-means is a method to cluster samples into clusters. K-means builds the cluster via an iterative process:

- At the initiation stage, we choose the number of clusters, and randomly pick the cluster centroids.
- In each iteration, a sample is associated to the nearest cluster centroid.
- In each iteration, each cluster recomputes its cluster centroid.

Through the iterations, samples form tighter and tighter clusters, and similar samples are grouped together. In our project, the samples are speech signal spectrum features (output of `librosa`), and the clusters are corresponding to therapist speech and child speech respectively.

We have also experimented with the deep learning package `tensorflow`. Performance is similar to that of K-means in tt sklearn.

4 LESSONS LEARNED

I really enjoy this research experience at SIP. I have been working in the lab environment, and learned a lot from PhD students, professors, and other fellow SIP students. It is eye-opening. I also enjoy learning new concepts such as spectrum analysis, K-means clustering, and deep learning, and apply them to the cleft problem with real-world impacts. Had we got more time, I'd continue to work to build a classifier for hypernasality. It would get to the heart of the cleft speech error detection problem.