Improving Speech Recognition Performance using Synthetic Data

NYU Center for Data Science

Group members: Ilias Arvanitakis, Kristin Mullaney, Alexandre Vives
Mentors: Brenden Lake, Michael Picheny, Bhuvana Ramabhadran

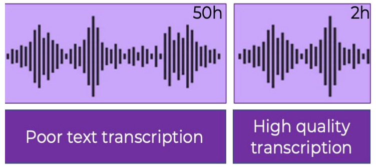
Abstract

Our speech recognition model will be using the SAYCam dataset, which consists of low-quality audio-transcription pairs of a parent speaking to their child. The footage for this data was collected by attaching a headcam to the child.

Previous approaches included training a traditional NLP algorithm that reached a 50% word error rate (WER), which was then brought down to 35% by a previous capstone group through the use of a transformer architecture. Our team generated several samples of synthetic recordings from different embeddings to enhance the current dataset and used them to train the model, achieving a 31.4% WER.

Background

The overall project focuses on childhood development, with the goal of understanding how children develop their language abilities through intaking speech from their surroundings. More precisely, we aim to build a model capable of accurately transcribing parent-to-child speech. The 2021 capstone team was able to achieve a word error rate (WER) of 35% using a transformer architecture, and our challenge is to assess whether expanding and/or enhancing the dataset using synthetic speech would decrease the WER further.



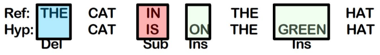
The dataset consists of 52 hours of a parent speaking to their child along with its transcription, where 50 hours were manually transcribed by several different people and therefore ended up being poorly transcribed (not properly aligned and having some wrong word transcriptions), and the remaining 2 hours were carefully transcribed by our mentor Michael Picheny.

Methods and models

Our team's objective is to use the transcription text of the poorly transcribed 50 hours of audio to generate a new 50 hours of audio that can substitute the old one in order to make sure that the text and the audio are aligned. Once that was completed, several models were trained using a combination between the 2 hours of properly transcribed data and a variable amount of synthetic data.

To evaluate our models, we used Word Error Rate (WER), which considers three types of errors: Substitutions, Deletions and insertions.

$$\text{WER} = \frac{\text{Substitutions + Deletions + Insertions}}{\text{Number of words}}$$



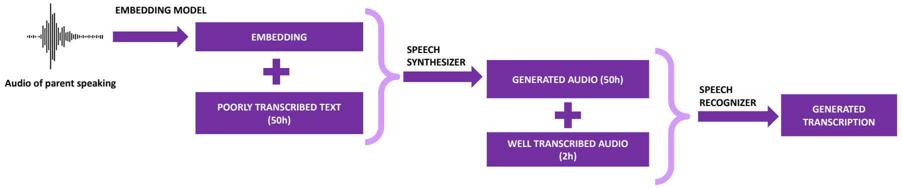
Error rate = 100 x (1 S + 1 D + 2 I) / 5 = 80%

This project required the use of three models in series.

First, a model that generated the embedding of the speaker's voice, taking a recording of the speaker's voice and returning a numeric vector representing that voice (by taking into account the pauses between words, how fast the person speaks, etc.).

Second, a speech synthesizer that generates the synthetic audio by receiving the embedding along with some text to be pronounced.

Third, a speech recognizer that generates the transcriptions of a given audio with a low WER.



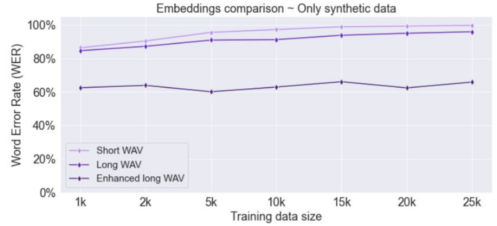
Experiments

Our team decided to use a wide variety of approaches utilizing the synthesized data. To create the embedding we used three different methods. Our first approach was to listen to the audios and pick a really high quality sound clip that we would then use as an embedding for our synthesizer. Our second approach was to combine multiple audio files to create a longer input for the embedding and the third approach would be to remove the background noise from the longer audio file to make an enhanced embedding. Initially we ran three models only with synthetic speech to see which embedding was the best. We also used two types of transcripts to train our model. The complete amount of transcripts and a reduced amount only containing those with more than 10 characters. Having two different transcript types and 3 embeddings we ended up with six different models.

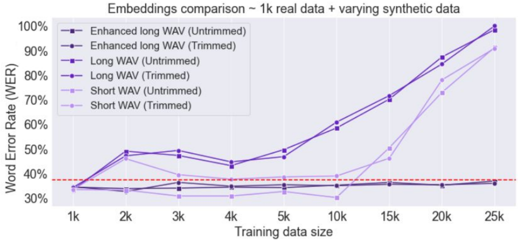
Results and discussion

Three different embeddings were designed to generate audio data: (1) Short input audio ~ 20 seconds, (2) Long input audio ~ 45 seconds and (3) Long enhanced audio ~ 45 seconds.

The model was trained with varying sizes of only synthetic data using each embedding in order to compare them and, as shown in the figure below, the enhanced long embedding had a much lower WER.



The final results can all be summarized under the figure below, where six models were trained on different synthetic data. Each colored pair was trained on model generated by a different embedding and the marker styles (circles and squares) represent whether the training data was trimmed (very short sentences were removed) or not.



Next steps

The enhancement of a dataset through the addition of synthetic data has proven useful in other areas such as image processing, so even though our methods managed to marginally decrease the initial WER from 35% to 31.4%, we believe there is potential to decrease it much further, for instance, by feeding a different synthetic dataset on every epoch. That is why we will attempt to carry on with the project to do so in the near future.