Child Speech Recognition

Athithan Elamaran, Aditya Singampalli, Advaith Anand

Background

- AI works by detecting patterns on its training data to be able to predict the output of testing data
- Automatic speech recognition for children is significantly worse than for adults due to the differences in their speech
- **Pitch:** Children's voices have a higher fundamental frequency (pitch) due to smaller vocal cords.
- **Formants:** Formants, which are resonance frequencies of the vocal tract, differ due to the smaller vocal tract size in children.
- **Articulation**: Children may mispronounce words and poorly articulate their words, especially when compared to adults
- **Foundation ASR Models** are speech recognition model trained by companies such as Open AI and Facebook on hundreds of thousands of hours of data and have very good speech recognition performance
- **VTLN** is a method which attempts to simulate lengthening the length of the vocal tract of a child, thereby making them sound more like an adult

Research Question

What methods can be used to improve the performance of Foundation Models on Automatic Speech Recognition for children?

Hypothesis

- We predict that **providing more speech data from children** for the model to train on will lead to better speech detection for children, because ML algorithms train by making patterns, so more training on child data will improve the testing results in child data
- We predict that **applying VTLN on children's speech** before transcribing the speech with Whisper will improve performance as it should sound more like adult speech, which is what foundation models are mostly trained on.

Methods

- Used Whisper-Tiny as a Foundation Model
- Fine Tuned Whisper-Tiny using OGI Dataset, which contains speech data from children of age 5 to 16
- Applied VTLN by using the VTLP function from the NLP-Aug library

Results

Note: To get Word Error Rate in percentage, multiply below values by 100

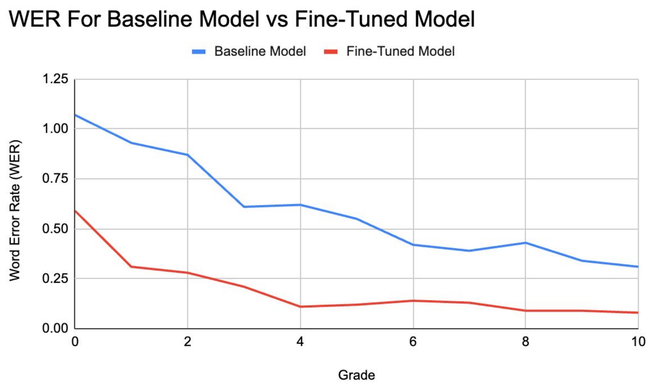

Figure 1

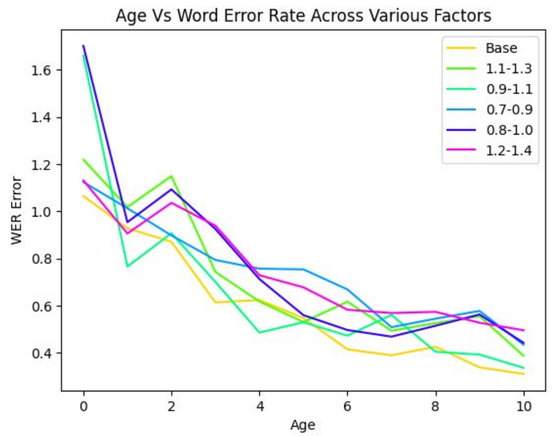

Figure 2

Table 1: Word Error Rate for Baseline Whisper-Tiny Model vs Fine Tuned Whisper-Tiny Model

Grade	0	1	2	3	4	5	6	7	8	9	10
Baseline WER	1.07	0.93	0.87	0.61	0.62	0.55	0.42	0.39	0.43	0.34	0.31
Fine-Tuned Model WER	0.59	0.31	0.28	0.21	0.11	0.12	0.14	0.13	0.09	0.09	0.08

Table 2: Word Error Rate for Baseline Whisper-Tiny Model vs Models trained on Data applied with VTLP

Grade	0	1	2	3	4	5	6	7	8	9	10
Base	1.065	0.928	0.871	0.614	0.624	0.548	0.415	0.39	0.426	0.339	0.311
VLTP Factor (1.1, 1.3)	1.219	1.018	1.149	0.743	0.618	0.531	0.617	0.494	0.526	0.556	0.388
VLTP Factor (0.9, 1.1)	1.66	0.766	0.907	0.701	0.486	0.53	0.473	0.561	0.405	0.393	0.337
VLTP Factor (0.7, 0.9)	1.124	1.013	0.898	0.794	0.757	0.754	0.669	0.509	0.545	0.578	0.435
VLTP Factor (0.8, 1.0)	1.701	0.954	1.093	0.925	0.713	0.56	0.497	0.469	0.515	0.563	0.443
VLTP Factor (1.2, 1.4)	1.131	0.906	1.036	0.939	0.729	0.678	0.583	0.569	0.574	0.528	0.496

Discussions

- Figure 1 and Table 1 show that children's ASR can significantly be improved by training foundation models on data specifically from children. However, one limitation is that all the training and testing speech for the fine tuning was done using only scripted speech. Spontaneous speech is a much more difficult challenge due to the way that words can connect. Another limitation is that we were not able to test different hyperparameters for the model due to limited time and computational units.
- Figure 2 and Table 2 show that the results are very mixed for VTLP-based synthetic adult data, and for the most part, the baseline model performs better. This is likely due to how we implemented our VTLP method, which operates on the waveform, converts it to a spectrogram, performs calculations, and then converts it back to the normal waveform. Whisper, however, takes in these spectrograms directly, leading to potential artifacts from repeated conversions.

Applications

- Improves accuracy of voice-activated devices and aids in education to boost efficiency of schooling for young children, which is helpful as most children can't type fluently until 3rd grade or later. If the machine is able to better identify the speech of these children, it can give more relevant responses, enhancing the education experience as a whole.
- Fine tune speech recognition softwares (Alexa, Siri, etc) to understand younger children more accurately. As these devices become more integrated into our lives, so will their use among children, making this very important.

Future Exploration

- Create Augmented Child data, by starting with adult data and applying VTLP to make it sound like child data so the Foundation model can be fine tuned on this augmented data, on top of whatever limited child data exists
- Repeat the experiment using other existing Foundation models
- Change the hyperparameters and use a validation dataset during training, which we weren't able to do due to limited compute units

References

- https://arxiv.org/pdf/2406.10507
- https://huggingface.co/blog/fine-tune-whisper
- https://www.semanticscholar.org/paper/Vocal-Tract-Length-Perturbation-(VTLP)-improves-Jaitly-Hinton/f79174a79b0391b6c75035abe1ebc7f5d52445f6