Team: Josh Chen, Will Chen, Robert Lo, Suhas Raja

**Key Challenge**

Reducing input size made the model very fast, but also sensitive to other optimization techniques. As a result, pruning and quantization cannot further reduce the latency without damaging EER.

**Potential Future Directions**

I think this model is suitable for applying NAS to reduce the number of parameters.

In addition, it might be worth investigating what the EER will look like if we re-train the model using 4kHz input.

Team: Chenzhun Huang, Tinglong Liao, Wenqing Cao

**Key Challenge**

Since GPT2 receives and produces subword units, the model must be run multiple times to output a complete word. This causes the inference time to increase significantly. Meanwhile, this increased computation cost doesn’t really improve the accuracy compared to much simpler models, such as n-gram.

**Potential Future Directions**

I believe GPT2 architecture might be too complicated for this application. Maybe it’s a good idea to create a smaller transformer model and train from scratch on word-level tokens. Maybe we can use BERT style training instead of autoregressive, in case user is typing in a middle of a sentence.