18781 Term Project:Create ASR only from TTS data

Group member: Dantong Dong, Yunfan Hua, Yanjun Guo, Dingzhong Hu

1. Motivation:

The goal of this project is to build an Autonomous Speech Recognition (ASR) system only using readable text data without any direct speech recordings. Currently, ASR still remains a huge challenge in the research area of natural language processing (NLP). One of the many challenges is training an efficient and accurate ASR model requires a large amount of well-collected and labeled human speech data. Although there are several existing open-sourced dataset for research purposes, it would still be very beneficial if we are able to utilize not only speech recordings but also the much more abundant and variable text data for training ASR models.

2. Proposed Solution:

As a natural extension of the problem stated above, we propose to try to create ASR models only from TTS data. It could potentially be very beneficial to various NLP research due to the follow reasons:

1) There are more dataset from text data than speech data. Leveraging the existing TTS framework can potentially generate speech data in a more scalable and dynamic manner.
2) TTS data has the ability to generate speech from words that are uncommon in spoken english. Leveraging TTS in ASR would give the system an edge when recognizing those uncommon words.
3) ASR trained from TTS data could potentially achieve higher accuracy than ASR trained from traditional speech data in certain situations when formal speech is required, such as airport, subway, and other public transportations.

Even if reating ASR only from TTS data does not achieve the same level of accuracy as training from traditional speech data when coming to traditional, more day to day speech, we believe it would still be very beneficial because our proposed technique may have the potential to integrate with existing ASR and increase the robustness of its recognition capabilities. However, it will not be in the scope of this project.

3. Detailed Approach:

The approach is divided into two parts: 1. TTS data collection and 2. ASR model training

3.1 TTS data collection

For text data, we will be primarily using Librispeech dataset as our raw text input to generate our speech output. We will also be using some customizing raw text input to supplement the dataset to enhance text data's variance.

After we have preprocessed and cleaned out our raw text dataset, we will be using **ESPNet2**'s exposed TTS API to generate speech dataset that will be later fed to our ASR system. We will also be utilizing AWS to assist with the computation of this step. Ideally it would generate well organized .mp4 format speech data after this step

3.2 ASR model training

As for the ASR model, we will be primarily leveraging some mainstream open-sourced ASR model, such as ESPnet Model (based on https://github.com/espnet/espnet_model_zoo) and training by only using the TTS data generated above.

We will be using Librispeech's ASR as our baseline. We will make a comparison between different ASR system's recognition capabilities against both TTS data and natural spoken data. We will also try to finetune our parameters so that the model will be better suited for TTS data. For example, we will explore feature extraction methods and different learning parameters.

4. Milestone timelines:

1) **Project proposal submission - Monday, October 4, 2021**
2) Collect and preprocess raw text data from Librispeech - Monday, October 11th
3) Generate TTS data from raw text data collected from step 2) - Monday, October 25th
4) Benchmark our TTS data against baseline and finished training at least one ASR system - Monday, November 1, 2021
5) **Midterm project presentation - Friday, November 5, 2021**
6) training other existing ASR system(2-3) and compare test results - Monday, November 15, 2021
7) Based on training result, build a customized ASR system specifically for TTS data training and compare with existing ASR system - Monday, November 29, 2021
8) **Poster presentation - Friday, December 3, 2021**
9) Wrapping up project and provide discussion on TTS data generation and comparison between different ASR system - Friday, December 10, 2021
10) **Final term project report - Monday, December 13, 2021**

5. Special Requirement:

This project could potentially benefit from utilizing Google's TTS API as the initial TTS data generation step, but we should be able to proceed with ESPNet2's exposed TTS API.