Au Wing Hing 3035199367
So Chi On 3035195359

Detailed Project Plan

Project Background

Recognizing speech and giving response are both human nature but not for machine, therefore human started to investigate a new technology that enable computer to capture the words uttered by human, which is speech recognition.

The concept of Speech Recognition was initialized in 1940s and the development has grown rapidly in 1980s, the traditional recognition models such as Hidden Markov Models (HMM) and neural network has successfully bridge human-computer interaction with great improvement. However, most of the research are based on the speech in dialogue but not in the tone of singing, these two sounds familiar but they are different in terms of the phoneme and pitch, such difference has made a huge difficulty to develop and train the model in this area.

The difference would be more significant in Cantonese songs whose lyrics would change the intonation of the words. Moreover, there are only a few Cantonese speech corpus in public, not to mention the non-existing corpus for Cantonese song. In light of this, we hope to initial a project to replenish this incomplete part of speech recognition development.

In the beginning of the stage, we consider to adopt HMM, one of the traditional yet effective model as our core part of the library. Meanwhile, since sequence-to-sequence models, especially attention-based model, outperform the traditional speech recognition models, we have done researches on several models that are possible to be implemented as part of the library.

Hidden Markov Models

It is a statistical model that can be used for modeling sequential data. Unlike the Markov Model that produce data underlying the clear processes, the state sequence(s) through that the model passes has/have been hidden and only the observational data can be known. It can be described as a probability distribution that can be used for classification. For speech recognition, the model can identify the given audio in spectrogram form and predict the correlated words and character based on the short sequence of stationary signal. It was commonly used as an acoustic model for speech recognition. The other parts of a speech recognition are preprocessing, pronunciation model and language model.

Au Wing Hing 3035199367
So Chi On 3035195359

Neural Network

Neural network has been treated as an efficient approach for ASR-related projects in different aspects including speech recognition. Compared to HMM, neural network requires fewer assumptions on statistical properties thus it is suitable on acoustic model that requires phoneme classification, preprocessing like spectrogram transformation.

Normally neural network underperforms under the continuous recognition model because of the inability on modelling temporal dependencies which is a crucial part for speech recognition, the recent models like Recurrent Neural Network and Time Delay Neural Network have tacked such issue and they have been implemented on end-to-end model.

End-to-End Models

There are two major forms of end-to-end model, which aim at jointly training acoustic, pronunciation and language models in order to simplify the whole process. It is a sequence-to-sequence model, which directly predicts the character or word from the given audio (spectrogram).

This first form is Connectionist Temporal Classification system which was first developed in 2014. It combined acoustic and pronunciation models and could map speech signals to English characters. However, due to the assumption of conditional independence, it was not able to learn language model and could make some spelling mistakes.

The second form is Attention-Based system which combined acoustic, pronunciation and language models since it does not have any assumption of conditional independence.
It is easier to build compared to traditional approach which requires separate models for different parts. The model basically consists of an encoder (consists of multiple recurrent neural network layers modeling the acoustics), a decoder (consists of multiple recurrent neural network layers predicting the output text sequence) and an attention layer between them to select frames in encoder for decoder to attend to.

Project Objective

The objective of the project is to develop a library for identify the Cantonese song lyrics.

Au Wing Hing 3035199367
So Chi On 3035195359

When the user input an audio file of Cantonese song, the library would process and response text-lyrics corresponding to the song.
The library would be support by a speech recognition model which consists of layers of models: acoustic model, pronunciation model and language model sequentially. These models would be trained by different datasets, the merged model would also be context-dependent.

Project Methodology

Preparation of Training Data Source

Since there are merely no ASR-corpus for Cantonese Songs, it is more expensive and difficult to train the models using songs. Although we may do that by ourselves, we plan to use existing Cantonese speech data to train our models first. Currently, the database we have found that would be possibly useful are: CU2C, Hong Kong Cantonese Speech Recognition Database, HKCanCor. Other than normal speech data, we also planned to create some data from songs which can be obtained from our own CD or maybe some music streaming platforms (such as Spotify, KKBox, etc). This should be time-consuming task because we need to extract syllables one by one from the song by segmentation and then map them with the corresponding Chinese character or pinyin by ourselves so this task will probably be handled after finishing the risky acoustic model. Pinyin of characters can be found in CUHK Chinese Character Database. However, since the tone of a character would vary due to different songs, singers, environment and other factors, tone of pinyin will not be considered.
We are not sure about the data size but we hope to gather at least 6 samples (half for male & half for female) for each of the 5000 common Chinese words.

Preprocessing

Since the songs have background music or other sounds produced by instruments, which would affect the accuracy of prediction of the model, and the tone and pitch the singers sing are quite different from normal speech, we propose extra preprocessing on the songs. Basically, we want to separate the vocal or human voice from the songs, which may be done by existing trained open source models. We also want to search for a way to turn singing voice to be more-likely speaking voice in fit with our model trained from normal Cantonese speech.
We plan to evaluate the performance of two kinds of outputs (inputs for acoustic model) which are spectrogram and MFCC (mel-frequency cepstrum coefficients). Spectrogram is a 3D graph with time, frequency and amplitude as axes while MFCC is a short-term power spectrum of the signal based on nonlinear mel scale of frequency.

Au Wing Hing 3035199367
So Chi On 3035195359

Acoustic Model

Since we are not sure able whether traditional or modern approach would work better on Cantonese songs, performance of 3 kinds of acoustic models, which are phonemes-based GMM-HMM (Gaussian Mixture Model-Hidden Markov Model), phonemes-based Attention-based Model and character-based Attention-based Model, will be evaluated. Due to the different modeling units (phonemes & characters), different labeling on training data is required.

Pronunciation & Language Model

If the output of acoustic model is Chinese pinyin, a pronunciation model is necessary to map the pinyin to a Chinese character, which can be done by searching in a pinyin-character mapping. Since there may be many Chinese characters with the same pinyin (even more cases due to missing tone), a language model is required to find the best Chinese character to fit in the sentence.

If the acoustic model is character-based Attention-based Model, although end-to-end ASR does not need a language model to predict speech theoretically, it has been proved that with an additional language model, the error rate would be improved. For Cantonese, there are even higher chance to predict a wrong word that has the same pronunciation with the correct word so an extra language model to adjust the prediction according to vocabulary nature and grammar may be needed.

Two kinds of common language models, n-gram model and neural language model, will be evaluated.

Evaluation

There are different combinations of output of preprocessing, acoustic model and pronunciation & language model. Each of them will be evaluated by CER (character error rate) and WER (word error rate) which are commonly used in ASR-related research papers.

Au Wing Hing 3035199367
So Chi On 3035195359

Project Timeline

Stage	Time	Name	Deliverables
1	16/9 – 30/9, 2018	Proposal	• Model Researches • Project plan
2	1/10 – 15/11, 2018	Requirement Analysis Data Preparation	• Drafted Data Preprocessing Model • Raw Training Data
3	16/11 – 31/12, 2018	Model Design	• Drafted Acoustic Model • Adjusted Data Preprocessing Model • Processed Data
4	1/1 – 20/1, 2019	Elaboration	• Detailed interim report
5	21/1 – 23/2, 2019	Construction (I)	• Drafted Pronunciation and Language Model • Trained acoustic Model
6	24/2 – 13/2, 2019	Construction (II)	• Trained Pronunciation and Language Model • Drafted Merged Model
7	14/3 - 14/4, 2019	Testing, Evaluation and adjustment	• Finalized tested Model • Final report