CSI5180 WINTER 2022
PROJECT PROPOSAL

Jiacheng Hou
jhou013@uottawa.ca

1 Project Title

Deep Learning-based Models for English Accent Classification

2 Motivations

Voice assistant (VA) consists of six main modules: wake word detection, automatic speech recognition (ASR), intent detection, action/fulfillment, answer generation, and speech synthesis. The author believes that the most critical module is ASR, as it is at the front step of the pipeline. However, accented English speech recognition remains challenging due to speaker-specific stress, tone and duration.

With the wide variety of English speakers with accents using VA, it is essential to improve the performance of ASR regarding English speech with accents. If the ASR module can know the speaker's accent in advance, it can extract the primary features of speech to improve its performance. Therefore, the author aims to build an accent classification module to predict the speaker's accent automatically, and the accent information can be fed into the ASR module later.

In addition, accent information can help VA to provide a customized regional user experience. For example, if a speaker asks VA to tell him or her a joke, VA can tell some funny stories based on the speaker's origin.

Furthermore, VA could even add an automatic pronunciation correction (APR) feature in the future. If VA can recognize what the speaker is saying, even if there is an accent, VA can help the speaker pronounce it correctly if the speaker turns on the APR function. The author desires to have such a VA to correct pronunciation at any time.

3 New Knowledge and Prior Knowledge

On the one hand, the author will learn to use the Google Cloud Speech-to-Text API [1] to show the importance of the accent information for the ASR module. Google's Speech-to-Text API accepts a parameter, languageCode, which indicates the dialect of the audio input. The author plans to input a correct languageCode and an incorrect languageCode for the same audio and compare the generated text output based on word-level confidence.

Furthermore, this is the first time for the author to handle a speech classification problem. Thus the author will learn the audio data preprocessing library librosa [2] in order to extract Mel-frequency cepstral coefficients (MFCCs) features. Besides, the author will learn to feed the time-series data into the Transformer Encoder architecture proposed in paper [3].

On the other hand, the author has some experiences with PyTorch and TensorFlow, such as deep learning models, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-term Memory (LSTM) and Graph Neural Network (GNN).

4 Final Deliverable

The project will deliver two deep learning models to classify English speech accents. These models can be deployed on the VA, and the prediction, accent information, can be fed into the ASR module to improve its performance.

5 Project Goals and Limitations

The project goals are presented in the following:

- Show the importance of accent information for the ASR module using the Google Cloud Speech-to-Text API [1].
- Build and train a baseline MLP model.
- Build and train a Transformer model [3]. The author will only build the Transformer Encoder architecture because this is a classification problem.
- Test the previous two models using the same test dataset and compare their performance.

The main limitation of this project is that the author will not apply a state-of-the-art model to the accented English classification problem.

6 Software Platform and Dataset

The primary programming language of the project is python. In order to show the importance of the accent information for the ASR module, the author will use the Google Cloud Speech-to-Text API [1]. For data preprocessing and deep learning models, the author plans to use the librosa [2], and PyTorch libraries [4], respectively.

The author will use the UT-Podcast corpus [5] to train and test deep learning models.

7 Activity Table

Table 1 shows the activity table for this project.

Table 1: Activity Table for the Accented English Classification Project

Activity	Why	Time Planned	Deliverable
Read articles and search materials	Gather knowledge about accent classification	5h	
Explore dataset	For training and testing deep learning models	3h	The UT-Podcast corpus
Call the Google Cloud Speech-to-Text API	Present the accent information is important	6h	Google's ASR response regarding word-level confidence
Data preprocessing	Required to make data into the proper format to feed into deep learning models	3h	MFCCs features of the audio input
Develop a baseline model	To have a simple model to compare with	3h	A MLP architecture
Develop a complicated model	To have a complicated model to compare with	5h	A Transformer Encoder architecture
Evaluate the baseline and complicated models	To evaluate and analyze performances of the two models	5h	Evaluation metrics and results

References

[1] Google cloud speech-to-text api. https://cloud.google.com/speech-to-text/docs/basics. Accessed: 2022-03-13.

[2] Librosa library. https://librosa.org/doc/main/index.html#. Accessed: 2022-03-13.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Pytorch library. https://pytorch.org/. Accessed: 2022-03-13.

[5] John HL Hansen and Gang Liu. Unsupervised accent classification for deep data fusion of accent and language information. *Speech Communication*, 78:19–33, 2016.