English Accent Recognition

Shi Lu, Li Zhu
University of Michigan School of Information

Objective

Non-native English speakers transform phonological rules of their mother tongue to English, resulting in systematic shifts of phones. Along with imbalanced training data, it has imposed a challenge to current speech recognition systems resulting in better performance for native English speakers. By correctly identifying the accent of a user, it is possible to modify the algorithm for better speech recognition performance. Therefore, we were inspired to build an accent classifier which could be used as a preliminary step in the speech recognition pipeline.

Method

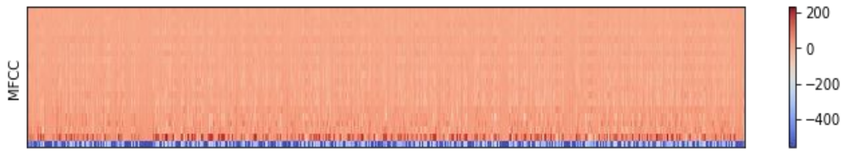
Dataset

- Wildcat Corpus of Native- and Foreign-Accented English[1]
- 40 pieces of Chinese Accented English recording
- 50 pieces of Native Accented English recording

Feature Extraction and Data Cleaning

Mel Frequency Cepstral Coefficients (MFCCs)

Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, which can be simply understood as the distribution of the energy of speech signals in different frequency ranges. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.



Data Manipulation

MFCC Extraction

- 1800 samples were created using the following procedure
- Remove silence
- Compress the audios to the same length
- Divided each audio into 20 pieces
- Extract MFCCs as a (13 * 300) matrix for each audio

Data Augmentation

- 1620 more samples were created, by randomly adding in noise in each matrix.

Model Training

- Support Vector Machine Classifier
- Multi-layer Perceptron classifier
- Gradient Boosting
- Convolutional Neural Network

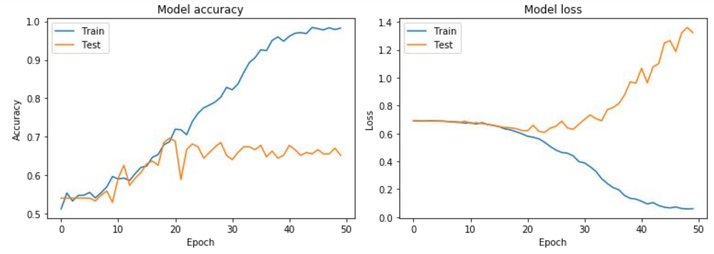
Layer (type)	Output Shape	Param #
conv1d_13 (Conv1D)	(None, 298, 128)	5120
max_pooling1d_13	(MaxPooling (None, 99, 128)	0
conv1d_14 (Conv1D)	(None, 97, 128)	49280
max_pooling1d_14	(MaxPooling (None, 32, 128)	0
dropout_10 (Dropout)	(None, 32, 128)	0
conv1d_15 (Conv1D)	(None, 30, 64)	24640
max_pooling1d_15	(MaxPooling (None, 10, 64)	0
conv1d_16 (Conv1D)	(None, 8, 64)	12352
max_pooling1d_16	(MaxPooling (None, 2, 64)	0
dropout_11 (Dropout)	(None, 2, 64)	0
flatten_4 (Flatten)	(None, 128)	0
dense_7 (Dense)	(None, 32)	4128
dropout_12 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 1)	33

Total params: 95,553
Trainable params: 95,553
Non-trainable params: 0

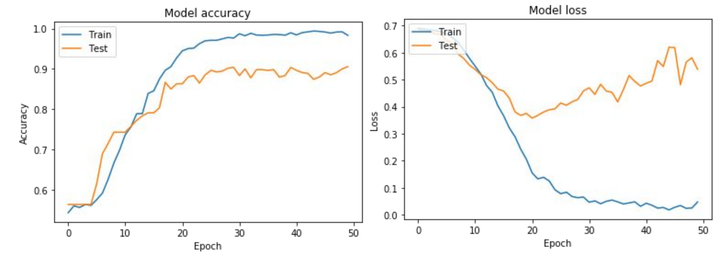
Result

- CNN with data augmentation reached 90% accuracy
- Data augmentation helps boost the performance for Gradient Boosting and CNN
- SVM model did not converge

Model	Data Augmentation	Accuracy	F1 Score
SVM	No	0.55	0.39
SVM	Yes	0.54	0.38
MLP	No	0.58	0.55
MLP	Yes	0.54	0.38
Gradient Boosting	No	0.59	0.60
Gradient Boosting	Yes	0.86	0.86
CNN	No	0.63	0.64
CNN	Yes	0.90	0.90



CNN Accuracy CNN Loss



CNN with data augmentation Accuracy CNN with data augmentation Loss

Discussion

1. Only MFCCs are extracted as features for accent prediction. Audios employ rich characteristics of human voices, which is not well presented by a few features. Features such as FBank, Delta, LPC could be extracted and experimented with in future work
2. Different methods of data augmentation should be experimented to improve model performance
3. Models for predicting accents from more languages would be a task for future works to improve the speech recognition system.
4. Feifei Li's team and Stanford University Music and Acoustics Computer Research Center jointly proposed a new method for sentence coding based on time convolutional network TCN, which improves the performance of speech recognition and emotion recognition by simulating the architecture of advanced acoustics[2]. Looking into works about TCN could be a future task

Reference

[1]http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/

Haque, Albert, Michelle Guo, Prateek Verma, and Li Fei-Fei. [2]"Audio-Linguistic Embeddings for Spoken Sentences." ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.