Deep Learning Approach To Accent Classification

Leon Mak An Sheng, Mok Wei Xiong Edmund {leonmak, edmundmk}@stanford.edu

Problem

Dialect or accent identification appears to be one of the bottlenecks affecting the fluency of automatic speech recognition systems.[1] Our goal was to apply deep learning methods to classify an English speaker's native language from the associated recordings in accented english. We built several neural networks with *Keras* to achieve reasonable classification accuracy on three different accents.

Dataset

We obtained our data from the Wildcat Corpus of Native- and Foreign-Accented English[2], in which participants clearly enunciate a list of words, one at a time. This made it easy to segment individual words pronounced by a speaker using *pydub*, an audio processing library for Python, for further feature extraction. Among the 84 scripted and unscripted recordings: 24 were of native English speakers of and 60 non-native speakers of English. As speakers' native language was predominantly English, Chinese or Korean, we chose these native languages as our classes.

Features

We used the Mel-frequency cepstrum coefficients (MFCCs), widely used in speech machine learning tasks.. After splitting each clip into utterances, we extracted 50 MFCC bands from each utterance and padded (or trimmed) them to be of the same fixed length of 1 second, then normalized it by subtracting the mean and dividing by the standard deviation. The input data is an *m* x 50 x *n* tensor, where *m* is the total number of utterances, and n is the number of frames sampled at 22050 Hz.

Models

We experimented with both traditional machine learning techniques such as SVM, as well as several deep learning architectures such as multi-layer perceptron (MLP), convolutional neural network (CNN) and LSTM recurrent neural networks (RNN) using the *Sequential Model* in *Keras*. We used categorical cross entropy as the loss function for neural networks and softmax activation function for the final layer, and Adam optimizer. Initial results showed overfitting for all neural networks, so we added dropout layers and applied L2 regularization to reduce overfitting. We used CNN with 3 convolutional layers with max pooling and an RNN with 3 LSTM layers. Reducing the learning rate also helped the models to converge to better accuracies.

Cross entropy equation $-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$

Discussion

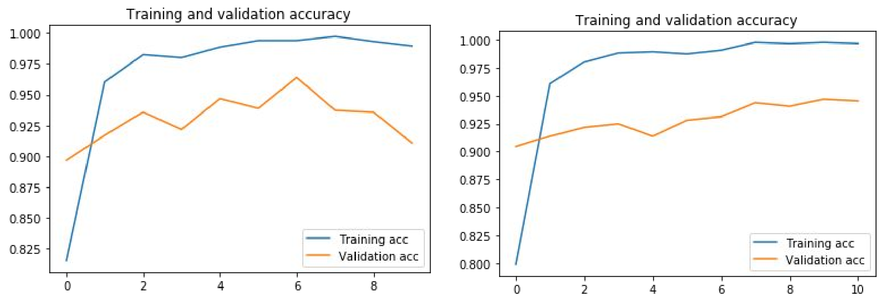
Our initial data was based on file level (full speech sentence) sampling at a fixed length, but we were unable to obtain reasonable performance (best test accuracy 40% with 5 classes on a different dataset, but similar preprocessing). This difference could be due to the fact that at the word level rhythmical characteristics except intonation is captured and can be used to distinguish english accent[3]. We also observed that traditional machine learning models like random forest and gradient boosting (implemented with *Scikit-Learn*) already performed particularly well, reaching around 92% test accuracy. Neural nets still outperformed them, despite our lack of data with only 20 speakers per class, since they are known to need massive amounts of data. We also observed that data augmentation by adding Gaussian noise to the MFCCs did give a minor boost to the accuracies (around 2%).

Results

Train Size: 9585 (88%), Dev Size: 639(6%), Test Size: 639 (6%)

Model	Train Error (%)	Validation Error (%)	Test Error (%)
MLP	0.96	7.04	9.79 / 7.66*
CNN	0.37	6.73	5.63 / 4.14*
RNN	0	4.38	5.59 / 3.57*

*Before / After Data Augmentation



Training and validation metrics for CNN (left) and RNN (right)

Future Work

Results show that classification works well on three classes, but we can get more classes to see if our model is able to discern the classes with more subtle variations. We could also extract other types of audio features like MFCC n-order derivatives (deltas) and mel-spectrograms. We can also extend our model to handle a wider range of more common words and use their phonetic features.

References

1. Ma, B., Yang, F. and Zhou, W. (2017) Accent Identification and Speech Recognition for Non-Native Spoken English
2. "Wildcat Corpus of Native- and Foreign-Accented English" [Online]. Available: http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/content.html
3. J. C. Wells, Accents of English. Cambridge, U.K.: Cambridge University Press, 1982, vol. I, II, III