Speech Accent Classification

Corey Shih
ctshih@stanford.edu

Abstract and Motivation

Automatic speech recognition systems are becoming increasingly common, and the ability to distinguish between the accents of different speakers can provide useful information about the speaker's nationality and heritage. To this end, I trained a long short-term memory (LSTM) neural network to classify accents from English audio samples. When classifying amongst 4 different accents, the model achieves a test accuracy of 52.27%, which is a significant improvement over random guessing, but still leaves much to be desired.

Data and Features

Data was taken from the Speech Accent Archive, which provides audio clips of different people speaking the same English paragraph and information about their native language.[1] Samples were taken from 4 of the most common accents (British, Spanish, French, and Mandarin) for a total of 430 examples, split into 386 training examples and 44 test examples.

Because people speak at different rates, the audio signals were resampled to be roughly the same length. The first 13 mel-frequency cepstral coefficients (MFCCs) of the signals were calculated for each time step.[2] The data was trimmed to 500 time steps, resulting in feature matrices for each example of dimension 13 × 500.

Figure 1. MFCCs for English Training Example 1.

Model

LSTM: LSTMs are a type of recurrent neural network capable of learning long-term temporal dependencies. They are commonly used for natural language processing and automatic speech recognition.

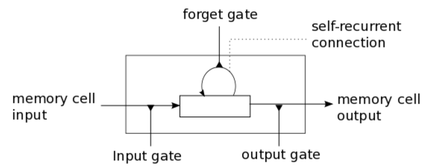

Figure 2. Schematic of LSTM cell. Image from deeplearning.net.[3]

LSTMs use sigmoidal and tanh activation functions. The network trained in this project contains a single LSTM layer and utilizes the softmax function as the activation function for the output node.

Results

Table 1. Training and test error for LSTM model.

Model	Training Error	Test Error
LSTM	20.98%	47.73%

References

1) Weinberger, S. H. (2015). Speech Accent Archive. http://accent.gmu.edu/
2) Ellis, D. (2012). PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m. https://labrosa.ee.columbia.edu/matlab/rastamat/
3) Carrier, P. L.; Cho, K. (2017). LSTM Networks for Sentiment Analysis. http://deeplearning.net/tutorial/lstm.html

Discussion

The model performs significantly better than randomly guessing amongst 4 categories, but there is still much room for improvement. The accuracy of the model could be negatively impacted by the fact that not all speakers listed under a specific language in the Speech Accent Archive actually have an accent. One would have to listen to all the samples and select only those with heavy accents. Additionally, some speakers in the audio clips stutter, which could throw off the model. Additional data may also be required for improved accuracy, as the dataset is very small compared to other datasets used for machine learning.

In general, classifying accents seems to be a difficult task, as picking up slight tonal differences in pronunciation is more difficult that identifying the word being spoken itself.

Future Work

To improve the performance of the model, further preprocessing of the data may be required. Dynamic time warping could be used to sync the audio signals as closely to each other as possible, or the audio sample could be cut into individual words to be used as separate features.

The dataset used has all speakers saying the same paragraph, but ideally, an accent classifier would be able to classify accents properly regardless of the English words being spoken. To this extent, the model would need to be trained on a much larger dataset of people speaking different phrases with various accents.