Bidirectional LSTM Classification and Unsupervised Visualization of Speech Accent

Ian Lawson Gazi Naven Tolga Aktas

University of Rochester, Department of Electrical and Computer Engineering

Abstract

Accent classification is an important problem in the realm of speech/speaker recognition. The number of acoustics elements that affect the perception of accent make machine learning an ideal solution to this problem. We propose a Bidirectional LSTM network for classification of accented English. We performed unsupervised learning on the output of the network to visualize how different accents are associated with each other.

Background

Speech recognition algorithms that have been trained to function on only one type of accent may have trouble when presented with another. Therefore accent classification can aid these algorithms by providing an initial analysis of what a given speaker may sound like [4].

Accent itself can be broken down into short-time phoneme based features and longer prosodic elements [1]. Classification algorithms should take both of these components into account when learning on a dataset [2].

Methods

We implemented two models for accent classification. The first is a four-layer bidirectional LSTM network displayed in Fig. 1. The second is a the same network with a subsequent attention layer to pinpoint the most important segments of an audio sequence, shown in Fig. 2. Using the output of the first model, we ran k-means clustering to create a visualization of the accent dataset.

We used the Common Voice dataset to train our models, which includes audio clips of 44000 spoken English sentences in 17 accents [3]. The distribution of audio clips for each accent is displayed in Table 1.

In order to evaluate the output of our system, we compute cross entropy loss for 17 classes.

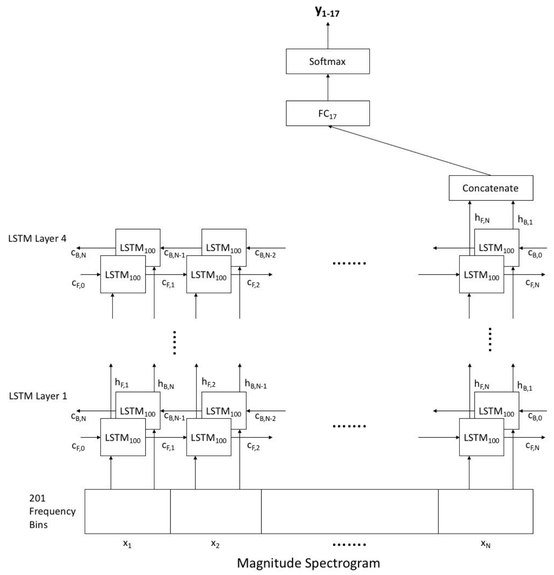

Fig. 1: Four Layer Bi-LSTM Model

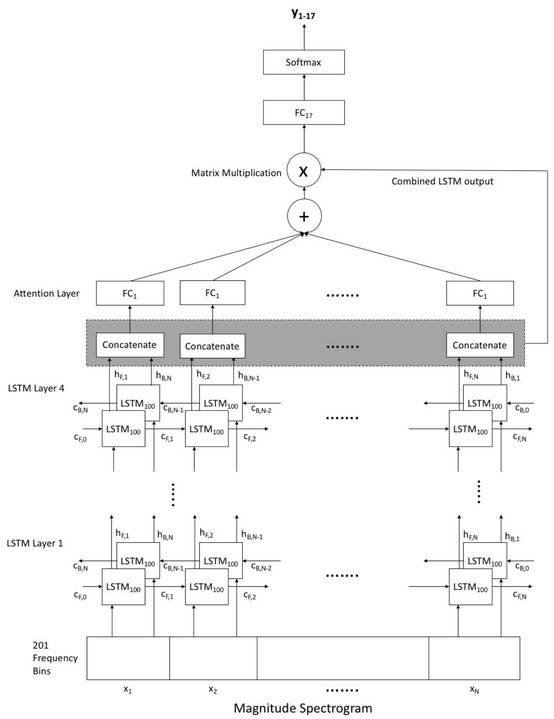

Fig. 2: Four Layer Bi-LSTM Model with Attention

Accent	Train	Valid	Test
US	24993	149637	630
England	5287	58607	154
Australia	4556	23966	290
Canada	3153	17586	58
NZ	585	6070	11
African	442	4089	25
Scotland	375	4382	12
Philippines	322	1330	10

Accent	Train	Valid	Test
Singapore	294	702	4
Ireland	257	3424	23
Malaysia	114	843	11
Other	113	10341	33
Hong Kong	20	1181	11
Wales	3.0	1128	4
Bermuda	0	449	10
South Atlantic	0	212	3

Table 1: Dataset distribution by accent and gender

Results

	LSTM	LSTM + Attention
Train Accuracy	57%	60%
Validation Accuracy	55%	55%
Test Accuracy	52%	54%

Table 2: Model Accuracies

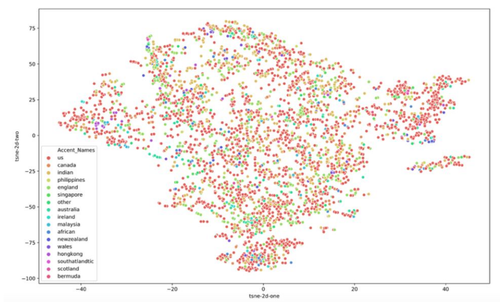

Fig. 3: TSNE Plot with ground truth accent labels

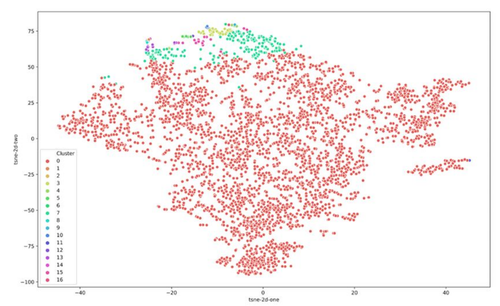

Fig. 4: TSNE Plot with clustering from network output

Discussion/Conclusion

Upon evaluating our results on the trained model, the test accuracy percentage and clustering analysis, a number of observations and proposals can be made.
While our trained model produces a test accuracy of 54 %, beating a pure-chance baseline, it is yet to yield reliable results on the accent classification task.

Ultimately our system had trouble with accurately classifying accents. As table 2 shows, neither model runs with a test accuracy over 54%. Clustering the output of the LSTM model, as displayed in Figure 4, led to one large cluster with smaller clusters at the top of the graph that still display inaccuracies. A comparative analysis of the two t-SNE clusters may indicate a source of problem due to the severe imbalance of distribution in the dataset, where the majority of the input samples come from certain accents, biasing the network to falsely classify towards more frequently appearing accents.

Another source of error can be attributed to the audio quality of the dataset. Our empirical studies have shown that many samples were fairly noisy, which might make it difficult for the network to properly learn good high-level representations of given accent. The level of variability in the quality of audio samples could also account for the lack of performance. In our training and inference, we used magnitude response from the STFT of the audio signal. This may not be the best representation of the raw input to learn features from, using more salient input representations such as the Mel-Frequency Cepstral Coefficients could yield better accuracy results.

The addition of the attention layer to the LSTM network appeared to only slightly improve the accuracy of the model. Whether this level of improvement would increase on a overall better-functioning network is something that could be studied in the future.

The fact that our best model performs with a 54% test accuracy, beating the pure-chance baseline, implies promising subsequent

Future Work

An interesting future direction of research is to employ generative training schemes to learn a variational approximate distribution of each accent with additional conditioning on the gender and the word-embeddings. Additionally, clustering could be utilized to classify regional accents within an unlabeled dataset of speech from a single country.

Citations

1. L Levent and J. Hansen: “Language accent classification in American English,” *Speech Communication*, Vol. 18, pp 353-367, 1996.
2. Y. Jiao, M. Tu, V. Berisha, and J. Liss: “Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features,” *Interspeech*, pp. 2388-2392, 2016.
3. “Common Voice,” *Mozilla*, 2019.
4. C. Huang, T. Chen, and E. Chang: “Accent Issues in Large Vocabulary Continuous Speech Recognition,” *International Journal of Speech Technology*, Vol. 7, No. 2-3, pp. 141-153, 2004.