Bidirectional LSTM Classification and Unsupervised Visualization of Speech Accent

Ian Lawson
University of Rochester
ilawson@ur.rochester.edu

Gazi Naven
University of Rochester
gnaven@u.rochester.edu

Tolga Aktas
University of Rochester
taktas@u.rochester.edu

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.

1. INTRODUCTION

Along with gender and age, accent is one of the main differentiators between speakers. Speech recognition algorithms that have been trained to function on only one type of accent may be inept when presented with another. Therefore accent classification can aid these algorithms by providing an initial analysis of what a given speaker may sound like [9].

A number of systems have been proposed that focus on multi-accented speech recognition [8][12][13]. Vergyri et. al. used Gaussian Mixture Models to classify training data for speech recognition based on country-of-origin. Once classified, the data was run through accent-adapted models dependent on the accent of each speaker. Improvements are seen in the word error rate (WER) for speech in each accent when using this model [12]. This implies that initial classification of accent is a promising direction in the development of successful speech recognition algorithms.

2. BACKGROUND

The accent classification problem involves identification of a pattern of speech within spoken audio data that is indicative of a certain region of the world. Each language can have innumerable accents associated with it dependent on whether it is a speaker's first language or if it was learned later in life. Each of these accents may have features that resemble the first language of the speaker, while also taking on patterns that are more related to the process of learning as a second language.

Amongst native speakers there is also a large amount of variation. Country-of-origin, as well as region-of-origin may have a large effect on pronunciation and tempo of spoken word in the same language. For example, speakers of English from the south of the United States have a vastly different accent from those in more northern regions, and both share differences with English speakers in the United Kingdom.

Differences between accents are comprised of a number of components. Non-native speakers of a language may display defective articulation of certain phonemes, which can be seen as additions, distortions, omissions or substitutions [1]. These phoneme based modifications are visible in the spectral characteristics of speech waveforms over instantaneous or short time periods.

Prosodic elements of speech also show variation between accents. The meter of one's speech as well as intonation and pitch contour patterns all have an effect on perceived accent. Timing of pauses can be especially useful in determining the origin of a speaker [1].

While this mix of spectral and prosodic elements can be good indicators for accent, they can be difficult to pinpoint. Therefore, for over two decades now researchers have been studying methods for accent classification outside of acoustic feature detection.

Hidden Markov Model (HMM) based approaches were one of the earliest attempts at accent classification in a probabilistic manner [1] [4] [5] [6]. L. Arslan and J. Hansen tested three models with different levels of a priori knowledge and found that a system in which HMM accent recognizers are trained on a predefined small vocabulary could perform at approximately 93% accuracy in accent classification given four accent classes. While this is a promising performance, the system only functions on the highly controlled set of words selected for training. Additionally, their evaluation of HMM approaches with less a priori information showed significantly worse performance, ranging from 60-70% accuracy [1].

More recently researchers have been applying artificial neural networks to accent classification as they can be powerful tools for training large sets of speech data [2] [7]. Y. Jiao et. al. implemented a fused Deep Neural Network (DNN) and Recurrent Neural Network (RNN) system that trains on long-term and short-term features respectively [2]. By factoring in both prosodic and articulative components of speech, this system saw increased performance in accent classification over systems that only considered one or the other. Overall accuracy of the system was around 52.5% for speech data only labeled by accent. Improvement is still possible in

this realm but the use of long short-term memory (LSTM) networks for accent classification is promising.

We decided to build upon some of the proposed accent classification networks to implement our own. The rest of this paper is organized into sections describing this implementation. Section three covers the methodology of our system. Section four explains how we evaluated the performance of our network. Section five presents the results of running our model. Finally section 6 concludes our paper.

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
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

3. METHODS

3.1 Dataset

We train our model with the Mozilla Common Voice Dataset. The training dataset contains over 44000 spoken english sentences. These sentences contain accent labels as well. Each sentences contain ~20 second long speeches. These sentences are spoken by people from a mix of 17 different origins, distributed as shown in Table 1. These origins are used as the accent labels for speeches [3]. We apply a 201 dimensional short term Fourier transformation (STFT) with block-size of 400 and hop-size of 200 on each of the speech sound files to convert them into spectrograms. We use the frequency level dimensions as the features for training our model.

3.2 Network Architecture

In our network shown in Figure 1 the model architecture is comprised of four Bidirectional LSTM layers. Each Bidirectional LSTM cell takes in 201 dimensions of inputs from each segment of the spectrogram and passes a 100 hidden state of 100 dimensions to the next cell. The final hidden layer, which is a concatenation of the hidden state from both directions, from the Bidirectional LSTM cell then passes through a fully connected network. Softmax function is applied to this layer giving probability for each of the 17 categories of accent labels.

Bidirectional LSTM can accept the information only in the forward direction, due to the temporality of the sequential data [11]. We employ BiLSTM cells in our model to let the network input both past and future information into current computation, thus learn important features from the past and future information, based on the neighborhood of the current state, rather than the specific ordering of the information in the sequence.

Our modified model, as shown in Figure 2, further employs attention mechanisms to learn the mappings between a subset of hidden states, such that it maximizes the use of more relevant information from past hidden states. LSTM layers create a context of the past information via hidden states. This is achieved by generating weights for each segment hidden state outputs from the BiLSTM cells via a separate single layered fully connected mini networks. As shown in Equation (1), this allows the network to build a context from past information selectively, maximizing the flow of more relevant information for the prediction at the current timestep. We use a soft attention mechanism that learns a weight vector $w \varepsilon (0, 1]$ using softmax function, and computes the selective context vector as a weighted sum over all hidden variables: [10]

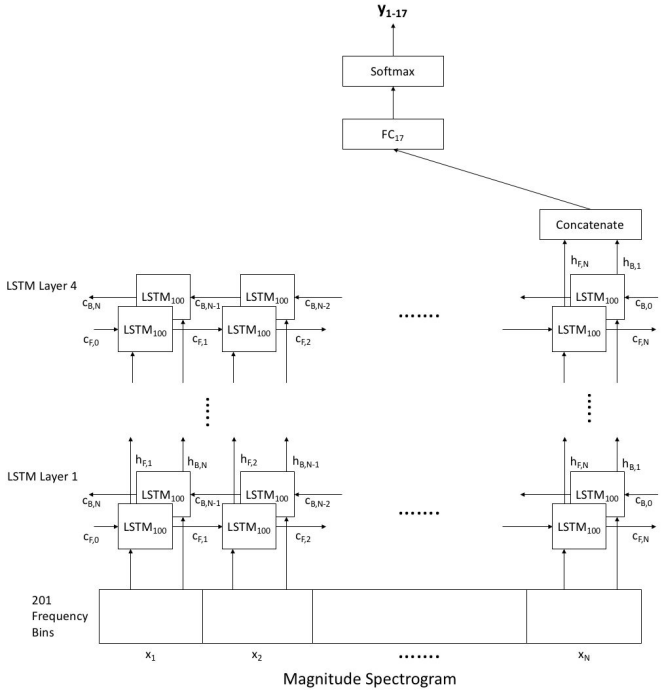

Figure 1. Four Layer LSTM Model

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \text{ ; where } a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \quad (1)$$

3.3 Clustering

We also attempt to understand the similarity of each of the accent in an unsupervised fashion. From the LSTM model without attention we collect the tensors of the last layer of the fully connected network for each of the speeches after the model has been fully trained. K-means clustering has been applied to this by treating the tensors as data. K value of 17 has been used as our dataset contains 17 distinct accent. We then map the accents for each of the clusters to understand the proportion of accent labels in each of the clusters. This has been visualized in the feature space to understand and analyze the similarity in each of the accents.

For ease of visualization, we visualize the data through a dimensionality reduction method called TSNE plot. This essentially squashes all the 400 feature dimensions into two dimensions. The true labels for each data are shown via color codes of the data points. The different colored contours in the background represent the clustered labels. We also label the TSNE plot with the proportion of true labels in each cluster.

4. Evaluation

We compute Cross Entropy Loss for 17 classes, using a softmax classifier on negative log likelihoods of the last activation layer.

$$CrossEntropyLoss = -\sum_{c=1}^{M=17} y_{o,c} \cdot log(p_{o,c}) \quad (2)$$

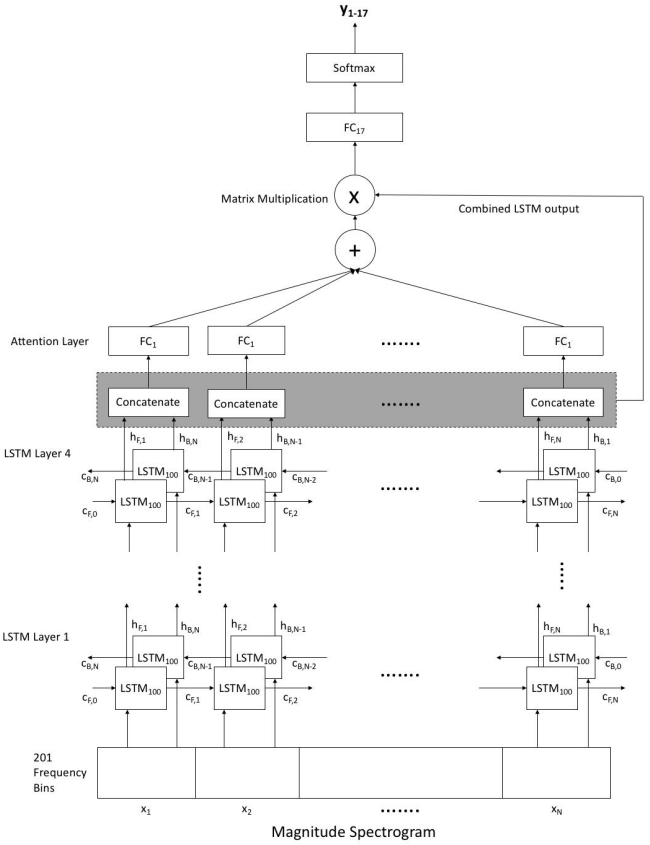

Figure 2. Four Layer LSTM Model with Attention

where c is the label, o is the binary indicator I (1,0) and M is the total number of classes (17 accents).

After every epoch the model is evaluated on the validation set which contains over 3000 voice clips. We were able to run our model completely through the training set for 30 epochs due to time and computing constraints. We use the average loss on the validation set to select the best model that gives the lowest validation loss. This model is then tested on the test set which contains over 2000 speeches in the similar format. The accuracy is computed on the basis of what proportion of the accent labels the model was able to predict right.

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

Table 2: Model Accuracies

5. Results and Discussions

Upon evaluating our results on the trained model, the test accuracy percentage and clustering analysis, a number of observations and proposals can be made.

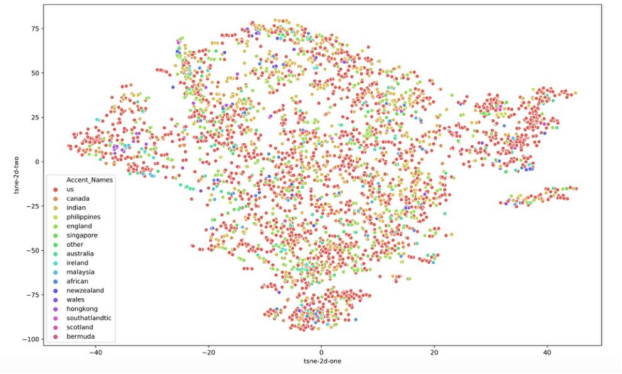

Figure 3. TSNE plot with ground truth accent labels

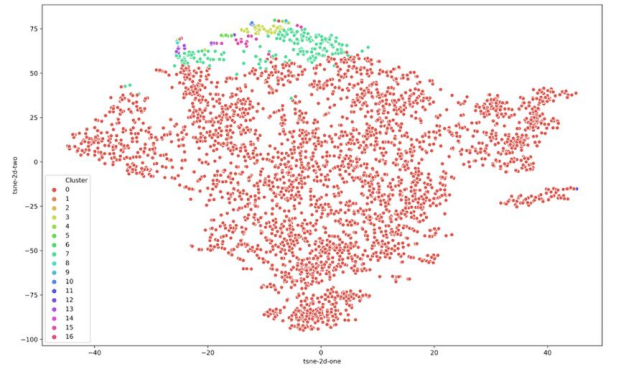

Figure 4. TSNE plot with clustering from network output

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. This most likely led our model to predict the class with most data all the time.

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 an overall better-functioning network is something that could be studied in the future.

6. CONCLUSION

The fact that our best model performs with a 54% test accuracy, beating the pure-chance baseline, implies promising subsequent works. An interesting future direction of research would be to employ generative training schemes to learn a variational approximate distribution of each accent with additional conditioning on the gender and the word-embeddings. Integrating language and pronunciation models can allow the network to better learn fine-grain features from the building block units (graphemes, phonemes etc.) chosen to represent lower-level input. Generative training can provide a prior and a regularizer for network to limits its parameter space search to more meaningful subspaces. A discriminative training following this generative modelling may yield more accurate predictions for having a better understanding of the data distribution of the different accents.

7. REFERENCES

[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. H. Tang and A. Ghorbani: "Accent Classification Using Support Vector Machine and Hidden Markov Model," *Lecture Notes in Computer Science*, Vol. 2671, pp. 629-631, 2003.

[5. K. Kumpf and R.W. King: "Automatic Accent Classification of Foreign Accented Australian English Speech," *ICSLP '96*, pp. 1740-1743, 1996.

[6. C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," *ICSLP '96*, pp. 1784–1787, 1996.

[7. A. Siddhant, P Jyothi, and S. Ganapathy. "Leveraging Native Language Speech for Accent Identification Using Deep Siamese Networks," *ASRA 2017*, pp 621-628, 2017.

[8. K. Rao and H. Sak: "Multi-Accent Speech Recognition with Hierarchical Grapheme Based Models," *ICASSP*, pp. 4815-4819, 2017.

[9. 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.

[10. D. Bahdanau, K. Cho,Y. Bengio: "Neural machine translation by jointly learning to align and translate," International Proceedings. International Conference on Learning Representations

[11. M. Schuster and K.K. Paliwal: "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[12. D. Vergyri, L. Lamel, and J. Gauvain: "Automatic Speech Recognition of Multiple Accented English Data," *INTERSPEECH 2010*, pp. 1652-1655, 2010.

[13. T. Fraga-Silva, J. Gauvain, and L. Lamel: "Speech Recognition of Multiple Accented English Data Using Acoustic Model Interpolation," *EUSIPCO*, pp. 1781-1785, 2014.