|  |
| --- |
| Long Short-Term Memory Networks for  Classification of English Vowel Phonemes  Alejandro Roman Martinez  Dartmouth College  Hanover, New Hampshire  alejandro.r.martinez.20@darmouth.edu  https://github.com/amartinez2020/SpeechRecognition |
|  |
|  |
|  |
|  |
|  |
|  |

Abstract

Long short-term memory networks (LSTMs) have taken over the field of speech recognition through the introduction of constant error carrousels. These carrousels allow the network to learn long-term relationships while also mitigating the risks of vanishing gradients. When applied to audio signals, LSTMs are able to learn valuable dependencies between frequency coefficients and forget invaluable information, making them ideal networks for speech recognition. In this study, we chose to apply these networks to the challenge of English vowel phoneme classification. When tested on the University of Western Michigan Vowel Database, our network achieved an accuracy of 91.37% on our validation set. With the promising results obtained, we anticipate that this network can be applied to vowel phoneme classification for other less prominent languages in efforts of recovery.

Introduction

Phoneme classification is a field of Computational linguistics research that focuses on the capacity to accurately classify phonemes. This field is an important aspect in understanding a given language’s phonetical structure. Much of previous work on phoneme classification is done through the use of Support Vector Machines (SVMs) (Amami et al 2012; Li et al 2005). While support vector machines provide simplicity to a network, in an effort to avoid overfitting and achieve worthy generalization, they are not optimal when dealing with sequential audio data. In this study we present the use of Long short-term memory networks (LSTMs) on such data, due to their exceptional ability to learn valuable long-term dependencies and forget invaluable dependencies (Hochreiter & Schmidhuber 1997; Graves et al 2005).

We specifically focus on English vowel phonemes in this study as a preliminary experimentation. As the chosen language in this study seems somewhat mundane (in the age of widely available English speech recognition programs) the true purpose of this study is to provide optimality to phone classification with LSTMs which may be able to generalize to less prominent languages that face the risk of extinction. Therefore, the results reflect future performance on these more imperative use cases.

This paper is structured as follows: In section 2 the methodology of the experiment is described including the dataset, feature extraction, the architecture of the network and its training strategy; the results of our network’s performance are presented in section 3; in section 4 we discuss the results as well as the anticipations and possible applications for this study.

Methodology

**Data**: The dataset we utilized was obtained from the University of Western Michigan Vowel Database:

<https://homepages.wmich.edu/~hillenbr/voweldata.html>

The dataset was available online and comprises of audio recordings of every English vowel phoneme from 150 men, women, and children (see figure 1 for English vowel phonemes).

A screenshot of a cell phone

Description automatically generated

**Fig. 1:** English Vowel Phonemes

**Feature Extraction:** After obtaining the data we transformed each audio file into Mel Frequency Cepstral Coefficients (MFCC). We chose to utilize MFCCs because of their inherent ability to accurately model human auditory perception. As humans have trouble distinguishing between high frequency sounds, MFCCs accurately model this quality by performing a log transformation on frequency data (see figure 2 for MFCC spectrogram). Our MFCC feature set contains a total of 12 MFCC parameters. The sampling rate was set to the audio files default of 16kHz.

Because the network we utilized was supervised we needed to provide training labels for our input. These labels were formatted as one hot encoded vectors that corresponded to the appropriate vowel phonemes.

A picture containing screenshot

Description automatically generated

**Fig. 2:** Mel Frequency Cepstral Coefficients on an Audio Signal

**Network Architecture:** As mentioned above, we chose to utilize an LSTM network to process each audio file (see figure 3 for LSTM architecture). The model was adapted from Keras, an open source machine learning library. As our goal was to classify a signal as one of 12 English vowel phonemes, we chose to implement an LSTM layer, took as input a sequence of MFCCs, followed by a fully connected dense layer that outputs a 12-dimensional vector. The LSTM layer contained 100 memory units and the dense layer utilized a sigmoid activation function.

The compilation of the model utilized a binary cross entropy loss function and a stochastic gradient optimizer because of the nature of the one hot encoded training vector.

A drawing of a person

Description automatically generated

Fig. 3: Long short-term memory network architecture

**Training:** We trained the network with a batch size of 32 for a total of 15 epochs. Because of the small size of our dataset, our network did not need many training iterations in order to reach convergence.

Results

Utilizing Keras’ model metric of ‘accuracy’, we obtained a loss of 0.3365 and an accuracy of 91.39% on our training set and an accuracy of 91.27% on our validation set.

Although there is room to improve, the similarity in performance between the training and validation set indicates that the model was able to avoid overfitting.

Discussion

As discussed in the introduction, the purpose of this study is to provide optimality to phone classification with LSTMs. In comparison with a study done by Amami et al 2012, who utilized SVMs for Phoneme Recognition, our LSTM model achieved a superior accuracy (91.37% vs. 52%). These results confirm the notion that LSTMs are optimal for classification on sequential audio data due to their ability to remember valuable long-term dependencies and forget invaluable dependencies.

The future for this study is twofold: increase the dataset and apply the network to less prominent languages at risk for extinction. As referenced in the methodology section our dataset did not need many training iterations to reach convergence. This is both beneficial and limiting. It is beneficial because we can achieve similar accuracy on languages that are at risk and cannot afford to produce large datasets due to a shortage of speakers. On the other hand, it is limiting because the early stopping prevents the network from obtaining a better accuracy rating that could be obtained by training on a larger dataset. This step for improvement must be dormant until we can find large datasets for at risk languages.

Conclusions

Acknowledgments

I would like to thank Rolando Coto-Solano for providing me with helpful guidance on network architecture as well as general mentorship.

References