Supervised Learning Models for Classifying Tones in Mandarin Chinese

Taylor D. Burke[1], **Bridgette Maurer**[1]

[1] Department of Computer Science – Wellesley College
Wellesley, MA

tburke3@wellesley.edu, bmaurer2@wellesley.edu

Abstract. *This open sourced project is focused on creating auditory feedback by effectively categorizing user speech into one of the four tones of Mandarin Chinese. This will enhance language learning with real-time pronunciation feedback. There already exists a plethora of sources to aid in memorizing vocabulary and understanding grammar, yet few for real-time pronunciation without the addition of an instructor or large price tag. This serves as a supplementary source that can provide quick and informative auditory feedback for tonal monosyllabic words.*

1. Introduction

When learning a new spoken language, the biggest issue can be the absence of an essential component: auditory feedback. Having a tutor or native speaker present to practice with is vastly more beneficial to a student than without. The goal of this project is to create a program that will fill this absence and assist language learning for those without access to native speakers. The scope of the project was narrowed to focus on Mandarin Chinese not only for the difficulty for native non-tonal language speakers in learning the language but also its popularity. For Chinese, there are 921.5 million first language speakers and 1.120 billion total speakers, making it a widely useful language.

For our project to be successful it needs to classify tonal data. Many existing projects we can employ demonstrate very high accuracy for a given tonal dataset. The most relevant was Madarin Tone Machine Learning Pro-ject by Alice Xue. Her project trained a Convolutional Neural Network (CNN) to classify the Tone Perfect dataset on either male, female, or combined data. For the data, 60 mel-frequency cepstral coefficients (MFCC), mel-spectrograms, or pitch contours were extracted from the audio files and fed as input features into the CNN. Her model's highest accuracy achieved was a 99.8% when data was separated based on gender. Thus indicating that a high accuracy is achievable.

In addition, real-time feedback requires real-time interpretation of a user speaking. Speech recognition research frequently utilizes Recurrent Neural Network models (RNNs). More specifically, Long Short Term Memory networks – "LSTMs" – which are a type of Recurrent Neural Network (RNN) that are specifically designed to avoid the long-term dependency problem and capable of recalling information over long periods of time. Each LSTM layer contains memory blocks which are a set of recurrently connected blocks. The blocks contain at least one recurrently connected memory cells and input, output and forget gates which indicate the operation of a given cell.

What makes our project unique is we combine the classification of tonal data and real-time auditory feedback to substitute the need for a human native speaker for anyone learning Mandarin Chinese. In addition, our model builds upon previous work by eliminating the need for gendered data sets, optimizing using hyper-parameters, and reducing the complexity of the model.

2. Methods

2.1. Collecting and Pre-Processing Data

Access was gained to the large dataset: Tone Perfect: Multimodal Database for Mandarin Chinese. This data set includes utterances of 410 monosyllabic words spoken in each of the four tones by six individuals (three male and three female) – 9,860 audio files total. The data was ultimately transformed from .mp3 formatted files into two saved .npy formatted files included in the project folder: one for labeling tones and one for 100 Mel-ceptrum coefficients across time for each audio file. The number for Mel-ceptrum coefficients was initially chosen to be large enough for hyperparameter testing later on. Feature extraction was focused on Mel-frequency coefficients (MFCCs). Extraction was done using a python audio analysis tool, librosa. In order to control for variance in audio file duration, the time dimension for each audio file was appropriately padded with zeros. The data was randomly split between training, testing, and validation data, respectively 60, 20, and 20% of the original data.

2.2. The Model

Given that the data is non-linear, sequential across time, and labeled, we chose to use a RNN, more specifically, a Long Short Term Memory model (LSTM). As supervised learning algorithms, these are specifically useful for unsegmented, continuous speech. The LSTM model architecture that was used is shown in Figure 1.

The model was trained on training data and validation data. Initial parameter testings showed that a fine-tuning of the hyper-parameters was necessary for full optimization.

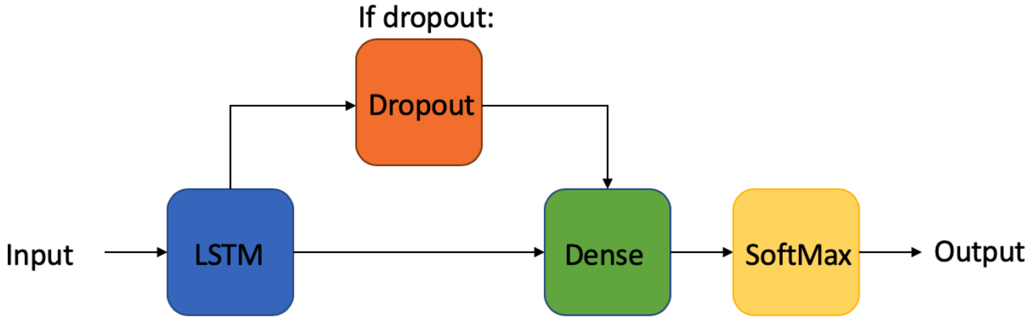
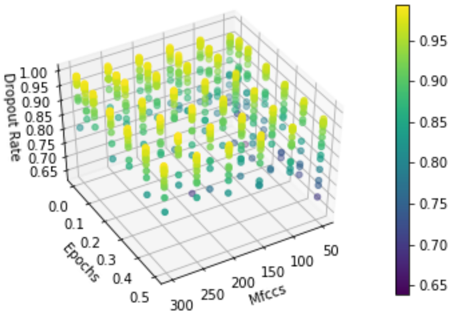
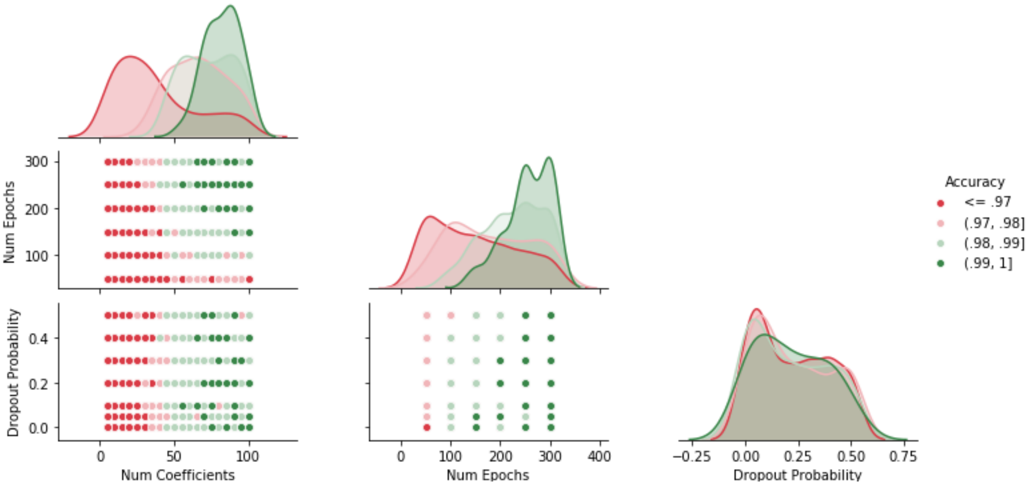


Figura 1. The Long Short Term Memory (LSTM) model architecture used for classifying the four tones of Mandarin Chinese. This model uses a model generator with a batch size of 24 for both the validation and training data. In the LSTM layer, only the output from the last cell at the last time stamp is passed on. The hidden layer size was pre-set to be 500. If a dropout value was specified above 0, a dropout layer was initialized. The model uses a SoftMax activation function in order to achieve the probability of classifications.

2.2.1. Optimization

Adjustments to these hyper-parameters were done to optimize the model: number of MFCCs, number of epochs, and dropout rate. To better understand the relationship between all of these and their effect on accuracy, the following graphs were made.





The top sets of hyper-parameters were chosen based their accuracy scores when models were trained and tested on validation data. The top performers, for which there were 5, were selected by setting an exclusive threshold at 0.9935. As seen in the figures above, all the top scoring models had more than 250 epochs and 60 coefficients. The top performing models were selected to move onto the next stage: 10-fold cross-validation. Cross validation was done on the whole dataset, for each of the 5 top performers. The set of hyper-parameters that scored the highest accuracy after cross validation was chosen as the optimal model.

3. Results

The optimal model had a mean accuracy of 0.981 during cross validation. This model's hyper-parameters were: 80.0 coefficients, 250.0 epochs, and a 0.3 dropout rate. The optimal model was then retrained on the data and its weights saved to a JSON file for future use.

3.1. User-Interaction

Users can interact with the project under the ”User Interaction”section. They will load in the optimal model that was saved to a JSON file. Then, users are able to speak into their computer microphone for 1.4 seconds to capture their utterances. The program will then display the user-recorded audio file and its predicted tone with a percent accuracy. They may rerun these two functions over and over again so as to increase their accuracy. This enables a user to verify their pronunciation and continue to practice with words of the same tone.

4. Future Work

Though the final product is not as robust as a launched app or deployed website, it runs with decent accuracy for immediate auditory feedback. An extension would be developing a user friendly GUI, though that was not the scope of this project.

Related to the limitations of this project, filler audio – "UM"s, "UH"s, slurred speech, ect. – are not taking into account for user input. Future work of the project would be to account for filler words.