Developing Automated Audio Assessment Tools for a Chinese Language Course

by

Ashvini Varatharaj

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

August 2020

APPROVED:

Professor Neil T. Heffernan, Master Thesis Advisor

Professor Jacob R. Whitehill, Thesis Reader

Professor Craig E. Wills, Head of Department

Abstract

Assessment in the context of foreign language learning can be difficult and time-consuming for instructors. Distinctive from other domains, language learning often requires teachers to assess each student's ability to speak the language, making this process even more time-consuming in large classrooms which are particularly common in post-secondary settings;considering that language instructors often assess students through assignments requiring recorded audio, a lack of tools to support such teachers makes providing individual feedback even more challenging. In this work, we seek to explore the development of tools to automatically assess audio responses within a college-level Chinese language-learning course. We build models designed to grade student audio assignments with the purpose of incorporating such models into tools focused on helping both teachers and students in real classrooms. This work also includes exploring various features - audio, tonal and text to help assess students on two outcomes commonly observed in language learning classes: fluency and accuracy of speech. We find that models utilizing tonal features exhibit higher predictive performance of student fluency while text-based features derived from speech recognition models exhibit higher predictive performance of student accuracy of speech.

Acknowledgements

I would like to thank my advisor, Professor Neil Heffernan for his guidance and mentorship. I also want to thank him for the opportunity to explore my interest in language audio domain. I want to thank my thesis reader, Professor Jacob Whitehill, for giving his feedback to help improve my thesis. I would also like to thank Dr.Anthony Botelho for his continued support, iterative feedback and mentoring he provided during my time at WPI. I would also take this opportunity to thank my ASSISTments team, lab mates and our sister lab mates at FL314 and SL133 for their constant encouragement and feedback. Lastly, I would like to thank my family and friends who have been my pillars support and motivation, without which this would not have been possible.

Contents

1 Introduction 1

1.1 Challenges 3

1.2 Outline 4

2 Background and Related Work 6

3 Part 1 : Exploratory Analysis 8

3.1 Dataset and Rubrics 8

3.2 Building model to predict grades 9

3.2.1 Pre-processing 9

3.3 Methodology 11

3.3.1 Feature Extraction 11

3.3.2 Deep Learning Model 12

3.3.3 Decision Tree Model 13

3.3.4 Baseline Method 13

3.4 Results 13

3.5 Ablation Study 14

3.6 Discussion 17

3.7 Conclusion 19

4 Part 2 : Exploring Tonal and Textual Features 20

4.1 Introduction 20

4.2 Related Work 21

4.3 Dataset 22

4.4 Feature Extraction 23

4.4.1 Pre-processing 23

4.4.2 Audio Features 23

4.4.3 Text Features 23

4.4.4 Tonal Features 25

4.4.5 True Audio : Google Text to Speech API 27

4.5 Models 28

4.5.1 Siamese Network 30

4.5.2 Multi-Task Learning and Ensembling 31

4.6 Results 32

4.7 Overlap Analysis 33

4.8 Data Augmentation 34

4.9 Bias Analysis 34

4.9.1 Gender Prediction 36

4.9.2 T-test 37

4.10 Discussion 37

5 Designing Tool 39

5.1 Framework 39

5.2 Handling Input 40

5.3 Student Options 40

6 Conclusion and Future Work 42

List of Figures

3.1 The Waveform from an audio of a student answer along with the predictions by the LSTM model at each window. 18

4.1 The figure shows the 4 tones in mel-spectrograms 26

4.2 ToneNet Architecture 27

4.3 The figure shows the steps involved in transforming audio to text features and feeding into the Siamese Network. Note: This is not a multitask model. Two individual models are used to each metric. . . 28

4.4 The figure shows the steps involved in transforming audio to sequence of tones and feeding into the Siamese Network. Note: This is not a multitask model. Two individual models are used to each metric. . . 29

4.5 Adding Noise 35

4.6 Pitch Shift 36

5.1 The Prototype of the grading tool 40

5.2 Framework for the web tool 41

List of Tables

3.1 The grading rubric used to assess student audio responses. 9

3.2 The 34 Features extracted from the audio along with its description. . 10

3.3 Audio Grade prediction: average 5-fold RMSE and R2 score for the models 15

3.4 Ablation Study results from the LSTM and Decision Tree model . . . 15

4.1 Results for the different models. 31

4.2 Multitasking and Ensembled Siamese models 32

4.3 Result of Tone Overlap Analysis 34

4.4 Augmentation results 37

Chapter 1

Introduction

When learning a new language, it is important to be able to assess proficiency in skills pertaining to both reading and speaking; this is true for instructors but also for students to understand where improvement is needed. The ability to read requires an ability to identify the characters and words correctly, while successful speech requires correct pronunciation and, in many languages, correctness of tone. For these reasons, reading tasks are considered an integral part of any Standardized language testing system for the syntactic, semantic, and phonological understanding that is required to perform the task well [32, 35, 1].

Learning and improving upon proper pronunciation is an essential aspect of learning to speak a new language. Almost all standardized language tests involve a section where the person being evaluated is expected to speak a few sentences about a given topic or answer a question; this is used to assess student skill and knowledge of pronunciation, fluency, and the correct usage of vocabulary. In the previous years, computer-assisted pronunciation teaching (CAPT) has gained attention and has been commercialized such as Pearson, SRI, and ETS [38].

While a notable amount of research has been conducted in the area of automating grading of reading tasks by a number of organizations (cf. the Educational Testing Service's (ETS) and Test of English as a Foreign Language (TOEFL)), the majority of assessment of student reading and speech is not taking place in standardized testing centers, but rather in classrooms. It is here that better tools are most needed to support both teachers and students in these assessment tasks. In the current classroom paradigm, it is not unreasonable to estimate that the teacher takes hours to listen to the recorded audios and grade them; a class of 20 students providing audio recordings of just 3 minutes each, for example, requires an hour for the teacher to listen, and this does not include the necessary time to provide feedback to students. If given the right set of tools, this time could instead be used to provide directed attention to students in need. The grading systems which have been researched are usually for more closed-form responses, such as multiple-choices or fill in the blank responses in which there is a single or small number of known correct answers. Open ended responses, such as essays or explanatory, answers are a more challenging task to automatically grade, but there has been growing research on developing automated assessment tools for such tasks[37] [33].

By automating the process of grading students on their pronunciation in language learning assignments, we can both help learners self-evaluate their progress and provide tools to teachers who traditionally grade students by listening to audio files (a task that can be very time consuming considering the number of potential students a teacher may have in a single class). In the past few years, automatic grading has become a tool in many learning systems. By giving one of the most time consuming task of the teacher to automatic computations, we allow teachers to direct their time to one-to-one discussion with students and give them feedback based on their performances and mistakes. Thus, automatic grading has an enormous impact and has gained attention over the last 30 years [10].

1.1 Challenges

However, collecting this data has many challenges since not all the data is being stored in one place uniformly like these platforms. Since oral tests happen in class, recording and storing this data adds an overhead to the teacher's responsibilities. Therefore, collecting such data is a hard task in itself.

During these oral exams, the students are evaluated for similar measures like the standardized exams such as pronunciation and fluency. This aspect of learning a second language is particularly important in the context of learning Mandarin Chinese. Given that Chinese (Mandarin Chinese) is a tonal language, the way the words are pronounced could change the entire meaning of the sentence, highlighting the importance of assessing student speech (through recordings or otherwise) an important aspect of understanding a student's proficiency in the language. Small variations in tone or misunderstandings of characters could result in unintended meaning. For example, the 'a' in 'yan' is pronounced as 'yen', like the currency in Japan. While the 'a' in 'Yang' does not have that 'e' sound [14]. An instructor would want to attend to these subtle differences to ensure that students fully comprehend these important aspects of the language.

Review and feedback is an important part of second language learning. Many online applications are available as sources to practice reading (texts) and writing. However, there is no tool suitable for classroom set-up especially having the review and feedback feature; this being crucial for classroom setups.

1.2 Outline

This work consists of two parts. In the first part: Chapter 2, we seek to explore the development of automated assessment tools for audio responses within a college-level Chinese language-learning course. We focus on exploring several challenges faced while working with data of this type as well as the generation and extraction of audio features for the purpose of building machine learning models to aid in the assessment of student language learning. In this part we present an exploratory analysis representing an initial step toward the goal of developing automated assessment tools for language learning audio responses. Toward this goal, we seek to address the following research questions:

1. By employing models of varying complexity, are we able to automatically grade student audio responses better than a simple majority class baseline model?

2. Does a recurrent deep learning model outperform a static decision tree model in regard in predicting student grades?

3. Which features extracted from student audio responses provide the greatest impact on model performance in predicting student grades?

In the second part: Chapter 4, we focus on more intuitive features and different model architectures. We use the additional new data collected in form of recorded student audio from reading assignments with the goal of developing models to better support teachers and students in assessing proficiency in both fluency, a measure of the coherence of speech, and accuracy, a measure of lexical correctness. We present a set of analyses to compare models built with audio, textual, and tonal features derived from openly available speech-to-text tools to predict both fluency and accuracy grades provided by a Chinese language instructor.

1. How can we utilize the text prompt information which is already known?

2. How can we obtain and use tone related features?

3. Can we use other model architectures suitable for our task?

Chapter 2

Background and Related Work

Automatic speech recognition systems have been the focus of several works over the past three decades. Some of this notable research has been conducted on the use of Automatic Speech Recognizer (ASR) technology for automatic speech scoring and the evaluation pronunciation quality [11] [4] [3] [39], as well as detecting mispronunciation [36]. To our knowledge, the majority, if not all, of this previous research has observed records collected through standardized language tests, leaving little work having been done to grade audio responses from students submitted as a part of a course or a class in schools or colleges as is observed in this current work.

As described, we are certainly not the first to explore aspects of audio data. Multiple approaches have been used by researchers in various domains to assess and understand speech. Bernstein, for example, developed a system which processes the scripted recordings of a speaker to evaluate language proficiency [4]. Their automatic grading system evaluated the pronunciation of student using Hidden Markov Models trained on pronunciations of native speakers' data.

Speech recognition is a great tool to help language learning and grading audio. As a part of their Voice Interactive Language Training System (VILTS) project, SRI attempted to improve upon automated audio evaluation methods by developing a system that did not rely on a scripted response [26]. That work utilized the "Decipher" continuous speech recognition system which generates phonetic segmentation to produce estimates of evaluative scores.

Other previous works to evaluate pronunciation have used ASR models combined with confidence metrics such as the Goodness of Pronunciation (GOP) [25] [23] [27]. There have been a number of prior works that have focused on the grading of read-aloud and writing tasks using Automatic Speech Recognition using interactive Spoken Dialogue Systems (SDS) [20], phonetic segmentation [5][11], as well as classification [36] and regression tree (CART) based methodologies[4][39]. Some speech recognition systems have used language-specific attributes, such as tonal features, to improve their model performances [31] [9].

With Deep Learning improving the performance of many tasks, it has been widely used in the Automatic Speech Recognition and Scoring tasks. In [19], an alternative to model-based phone distances in the form of a tunable Siamese network feature extractor was used to extract distance metrics directly from the audio frame sequence to predict score.

Many states have begun to use computer-based English Learning Proficiency (ELP) assessments, which has led to a growing interest in automated scoring of spoken responses to increase the efficiency of scoring. However, there is a dearth of systems which grades foreign language learners. In this paper we plan to focus on Chinese learning Undergraduate students.

Chapter 3

Part 1 : Exploratory Analysis

This chapter, like we mentioned previously is about an exploratory analysis representing an initial step toward the goal of developing automated assessment tools for language learning audio responses. In particular we look at low-level audio features and the strength of each of these features using an Ablation study.

3.1 Dataset and Rubrics

The data set is collected from a Chinese language class for undergraduate students in a university located in the North-East region of the United States. The data collected consists of 63 distinct students from 3 classes run between 2017 and 2019. All of these classes were taught by the same instructor. The data is comprised of 60 audio recordings which includes audio responses from 3 different prompts. The average length of the audio is approximately 2.5 minutes.

As part of the class, students were regularly given assignments containing prompts which required students to respond by recording themselves speaking and uploading their recordings alongside additional written responses to additional problems; this work focuses solely on the audio recordings submitted by students. All work

Grading Components	**Percentage**
Rich Content	40%
Grammar and word usage	20%
Accuracy of tones and pronunciation	20%
Fluency	20%

Table 3.1: The grading rubric used to assess student audio responses.

is submitted through an institute-hosted learning management system from which the teacher then downloads all the files in order to listen and grade each student's work. In our dataset, the teacher followed a rubric when assessing each student that is reported in Table 3.1. From this rubric, we can say that the content of each responses represents 40% percent of the grade while grammar, pronunciation, and fluency each represent 20% each.

3.2 Building model to predict grades

3.2.1 Pre-processing

A series of pre-processing steps were applied to the raw audio responses for use in this work. The data processing and cleaning steps is important here as often the case in any data mining or machine learning task; we describe these steps in detail here to improve the replicability of our work in the future.

Among the 60 distinct responses contained within our dataset, 29 responses are in the form of a video while the remaining 31 responses are purely audio across the 3 prompts. As students were not assessed based on any visual aspects of their responses and the goal of this work is to develop assessment tools of student audio, we first extracted the audio track from each video and standardized all files to use the same audio file format (.wav).

In 12 cases, student responses included multiple students speaking; it is likely

Feature ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Table 3.2: The 34 Features extracted from the audio along with its description.

that permission were given to these students to work collaboratively on the assignment. In 2 of these cases, a separate grade was given to each student present in the recording while in the remaining 9 cases, the students involved were given the same score. We observed that the grades in these 2 cases where the score differed did not vary more than 1 point (on a scale of 11, from 0-10 inclusively), and therefore we aggregated each of the responses and grades into a single instance by simply averaging the two scores. In this way, we represented these 2 cases with a single averaged score. The other 9 cases were then left as individual samples , leaving us with the aforementioned 60 distinct responses for use in our models.

For the feature extraction step (described in the next section), we further need to convert our audio data in a mono-channel format. The data we collected contains stereo data (i.e. it contains a separated right and a left channel). To convert each response to a mono signal, we simply take the average of the two channels of the stereo data.

3.3 Methodology

In this section we describe our methodology of extracting features and building models for the purpose of predicting teacher-provided grades for student audio responses. We compare two non-linear models to a simple baseline. While the grade labels are provided on an 11-point scale (0-10), this scaling is non-linear in nature due to non-uniformity across the grades (the average grade was 7.41). Since few of the assignments were graded on different scales, a min-max scaling was used to transform all the scores into a scale of 0-10.

3.3.1 Feature Extraction

The first step of our methodology is to extract features that we believe will best measure aspects of the audio signal that will be useful in identifying the linguistic information for assessment. Audio feature extraction is performed to transform the audio signals recorded in the wav files into a representation which can be used for machine learning. We try to characterize the signal with values capturing commonly-measured properties describing the sound. The features we use in the analysis are extracted using the *PyAudioAnalysis* library; this is a python-based library which is exclusively used for audio data feature extraction.

We use the default parameters of 50 milliseconds and 25 milliseconds for the window size and step amount respectively. With these parameters the feature extraction function splits the input signal into short windows (frames), leading to a sequence of short-term feature vectors at regular intervals within the signal. The number of windows varies for each of the signals based on the length of the audio file. The features extracted for use in this work are briefly described in Table 3.2. For each time window, we extracted the 34 features, where several of the features described in Table 3.2 are described using a multi-valued numeric vector indicated through the Feature ID column.

3.3.2 Deep Learning Model

The sequential and temporal aspects of audio data makes the application of a recurrent deep learning model an appropriate choice for developing automated assessment tools. Specifically, we utilize a Long Short Term Memory (LSTM) [15] network, as it is designed to model complex temporal relationships within sequential data. The deep learning model observes a sequence of time steps (e.g. frames of audio as described in the Feature Extraction Section) and is trained to produce a single value corresponding to the estimated grade for the student response.

Due to the number of features and length of each student response, we opted to utilize a network structure with 3 hidden layers. The input to the model is represented as a sequence of 34-valued vectors corresponding with the extracted features, which is then passed to a LSTM hidden layer of 50 nodes, before being passed through 2 additional fully connected non-recurrent layers of 100 units each. An output layer of a single node is used corresponding with the grade of the student, treated as a regression as opposed to a classification task. We chose this particular structure to help prevent the network from overfitting due to the long sequences (i.e. by using a smaller LSTM layer), but providing enough depth in the model to learn feature representations from each sequence; we explore additional model structures in the next section.

The LSTM looks at each frame and provides an estimated grade for it, but is only updated and evaluated on the final frame of the sequence. We applied a 5-fold cross validation on the dataset and measure performance using RMSE and Spearman correlation.

3.3.3 Decision Tree Model

To contrast the deep learning network, we also compare a simple decision tree model. Such a model still has the capacity to learn non-linear relationships between the features and teacher-provided grades, but does so in a non-sequential manner. As such, we needed to aggregate the audio features into a single vector that describes the response as a whole as input to the model. We simply take the average across each of the 34 features across each response and use this to predict the teacher-provided grade.

We apply a decision tree regressor using the CART algorithm [7]. Similar to the deep learning model, we evaluate the model using a 5-fold cross validation using measures of RMSE and Spearman correlation. Within each training fold we optimized it for it's depth within each of the 5 folds using the training data.

3.3.4 Baseline Method

We compare the LSTM model and decision tree model to a simple baseline using a majority class model. For this method, we simply take the average grade provided by the teacher and use this as a prediction for every sample. As this is a simple baseline, it can be used to evaluate how well our LSTM and decision tree models perform in comparison to a model that incorporates no audio information. Failure to outperform this baseline would indicate that such a model was unable to effectively learn from the data.

3.4 Results

We report model performance using measures of RMSE, a measure of prediction error, and Spearman correlation (Rho); lower values of RMSE below the baseline model are indicative of high model performance, while higher values of Rho (above zero) are indicative of high model performance.

The performance of each of our models in predicting the grade of student audio responses is reported in Table 3.3. From this, it can be seen that both models outperform the baseline model in regard to RMSE, but only the LSTM model exhibited positive correlation. For this reason, it is apparent that the LSTM model is the superior model, although the values suggest that there is still room for improvement. As the grade labels followed a 10 point scale, the best RMSE of 2.728 exhibited by the LSTM suggests that it is, on average, over- or under-predicting the true grade of the student by just under 3 grade points.

Despite this room for improvement, the results do suggest that both the LSTM and decision tree models are learning from the data in potentially different ways. For this reason, we further explore each of these models through an ablation study.

3.5 Ablation Study

In this experiment we perform an ablation study where we run a model with all the features and iteratively remove each to observe impacts to model performance. Changes in model performance as a result of removing a feature can be used then as a measure of feature importance in determining the grade of the student. In this way we try to identify and rank the importance of each feature in evaluating student audio responses for each the LSTM and decision tree models.

Table 3.4 reports the results of this study across both the LSTM and Decision Tree models. The rows are sorted to reflect the features of highest impact found in regard to changes in RMSE for the LSTM model as this was the highest performing model across both metrics.

Model	**RMSE**	**Rho**
Majority Class	3.323	-
Decision Tree	2.807	-0.076
LSTM	2.728	0.163

Table 3.3: Audio Grade prediction: average 5-fold RMSE and R2 score for the models

	LSTM		**Decision Tree**	
Feature Removed	**Delta RMSE**	**Rho**	**Delta RMSE**	**Rho**
Chroma	**0.118**	0.172	0.217	**0.162**
Entropy of Energy	**0.076**	0.18	0.143	0.074
Zero Crossing Rate	**0.042**	0.149	0.132	0.121
Spectral Centroid	0.042	0.148	**0.226**	0.072
Spectral Spread	0.04	0.189	0.205	0.085
Spectral Rolloff	0.028	0.206	0.136	0.107
Spectral Entropy	0.025	**0.217**	0.132	**0.121**
Chroma Deviation	0.018	**0.208**	0.059	0.076
Energy	-0.011	0.19	**0.361**	-0.025
Spectral Flux	-0.016	**0.242**	0.151	0.11
MFCC	-0.18	0.161	**0.314**	**0.132**

Table 3.4: Ablation Study results from the LSTM and Decision Tree model

In regard to the decision tree model, the 3 features which cause the largest drop in RMSE are Energy of the wave, MFCC features, and Spectral Centroid. The RMSE for Energy increases from 2.807 to 3.168 giving an increase of 0.361 which is high compared to the rest of the features. This is followed by the removal of the 12 MFCC features which cause an increase from 2.807 to 3.121 which is a 0.314 increase. The spectral centroid is the third feature whose RMSE increases to 3.012 from the original model giving an increase of 0.226. With respect to the Spearman corelation metric, the top three correlated features are the Chroma features followed by the Zero Crossing Rate and MFCC features.

In regard to the LSTM model, The 3 features which cause the largest decrease in the RMSE are the twelve Chroma features, the Energy Entropy, and the Zero Crossing Rate feature. The removal of the Chroma feature shows the maximum increase in the RMSE of value 0.118. It is then followed by the Energy Entropy feature which increases the RMSE by 0.076. The zero crossing rate is the third highest feature which causes an increase of 0.042 in the RMSE value. On the other hand removal of MFCC features causes a decrease in the RMSE by 0.180 which reduces the RMSE of the model to 2.548. Similarly, the removal of Spectral Flux and Energy feature also causes a decrease in the RMSE by 0.016 and 0.011 respectively. However comparing the Spearman's correlation measure (rho) does not follow the same trend as the RMSE. Considering the LSTM model,the model with all features has a correlation of 0.163 with the true label given by the teacher. While removing certain features shows a higher correlation value. For example, the removal of Spectral Flux value produces a model which has the highest correlation value of 0.242. This is followed by the Spectral Entropy and Spectral Rolloff features which give a correlation value of 0.217 and 0.206 respectively. In the case of LSTM most of the models have a rho value more than the model with all the features. This may be suggestive of either overfitting occuring within the model, particularly as some of the features similarly lead to improvements in RMSE when removed.

3.6 Discussion

As we saw from the decision tree model as well as the LSTM model, the energy related features(Entropy of Energy and Energy) seems to be having a significant impact on the evaluation of the pronunciation. In [12] the idea of 'formants' are discussed, which are bands of acoustic energy around a particular frequency which characterizes different resonances of the vocal tract. It is mentioned in the same study that by measuring these differences in the mouth/tongue by instrumentally finding how the formant structure may help understand pronunciation across different vowels. The second feature which seemed to have an impact in the decision tree is the MFCC feature. The important task of understanding speech and pronunciation is that the sounds produced by humans are filtered by the shape of the vocal tract which determines what sound is emitted. By accurately defining the shape, we could get representation of the phoneme being produced. The MFCC feature accurately characterizes the envelope of the short time power spectrum which in turn manifests the shape of the vocal tract. Hence it makes sense that it influences the pronunciation score. However, the LSTM model seems to understand MFCCs differently and hence, the removal of these MFCC features seem to having a improvement in the model. Further analysis on the MFCC features can help us understand why this is the case. The Chroma feature is usually used for music data. It provides information related to the twelve different pitches which are used to analyze musical data. However we chose to explore how the pitches in human voice can be analysed using these features. These Chroma features seem to influence the LSTM model

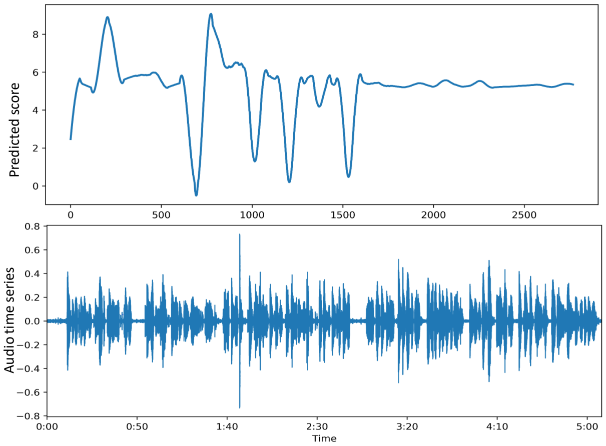


Figure 3.1: The Waveform from an audio of a student answer along with the predictions by the LSTM model at each window.

with their removal producing a high increase in the RMSE. Even the decision tree model seems to show an increase in the RMSE by 0.217 when these features are removed.

As mentioned earlier, one of the goals of this paper was to understand if we can provide valuable feedback to the students based on their audio submission. A benefit of the sequential structure of the LSTM model is its ability to illustrate the development of its grading estimates over the audio response. From moment-to-moment, a grade estimate can help to indicate sections of the audio response that are suggestive of a high grade (e.g. well-pronounced words) and sections suggestive of a low grade (e.g. poor pronunciation or areas of silence); an example of this is illustrated in Figure 3.1. In that figure, the bottom image depicts the wave form of a student audio response while the top figure illustrates the LSTM estimate over the length of the response. Such a report could help teachers to identify sections of audio where the student may be in need of additional aid by observing where the model predicts a low score.

3.7 Conclusion

This work represents an initial step toward the development of automated assessment tools designed to aid language learning students and teachers. Toward this, we have compared the decision tree and a LSTM model to predict the scores of the audio. With the ablation study on these two models, we see that certain features did correlate to the scores given by the teacher. However with different features found to be important across both models and across metrics, further in-depth analyses of different combinations of the features could help to better understand these relationships. This led us to our next part of exploring additional features. In part-2 we look at reading tasks in specific where students read-out aloud the given prompt. We dive into looking at features related to text and tones extracted and understand how they affect these grading models.

Chapter 4

Part 2 : Exploring Tonal and Textual Features

4.1 Introduction

This work is the continuation of part 1. Part 1 provided insight whether a model could learn to grade the audio assignments based on the low-level features. However, the study does not take the text spoken as input. In this current study, we look into 'reading task' assignment which provides us the advantage of having the text read out by the students. This work observes student data collected from a college-level Chinese language learning course. We use data collected in form of recorded student audio from reading assignments with the goal of developing models to better support teachers and students in assessing proficiency in both fluency, a measure of the coherence of speech, and accuracy, a measure of lexical correctness. We present a set of analyses to compare models built with audio, textual, and tonal features derived from openly available speech-to-text tools to predict both fluency and accuracy grades provided by a Chinese language instructor.

4.2 Related Work

There has been little work done on developing tools to support the automatic assessment of speaking skills in a classroom setting, particularly in foreign language courses. However, a number of approaches have been applied in studying audio assessment in non-classroom contexts. Pronunciation instruction through computer-assistance tools has received attention by several of the standardized language testing organizations including ETS, SRI, and Pearson [38] in the context of such standardized tests; much of this work is similarly focused on English as a second language learners.

In developing models that are able to assess fluency and accuracy of speech from audio, it is vital for such models to utilize the right set of representative features. Previous work conducted in the area of Chinese language learning, of which this work is building upon, explored a number of commonly-applied features of audio including spectral features, audio frequency statistics, as well as others [34]. Other works have previously explored the similarity and differences between aspects of speech. In [19], phone distance has been used , which is defined relative to the other phones rather than characterising each individual phone to analyse non-native English learners. Work has also been done on analysing different phonetic distances used to analyse [30] speech recognition of vocabulary and grammar optimization. Research has shown that assignments that provide sound-based connections are more beneficial to language learners and thus meaning-based connections should be provided as secondary to these [22].

Many approaches having been explored in such works, starting from Hidden Markov Models [4] to more recently applying deep learning methods [19] to predict scores assessing speaking skills. Others have utilized speech recognition techniques for audio assessment. There have been a number of prior works that have focused on the grading of reading and writing tasks using Automatic Speech Recognition [20] [5][11] [4] [3] [39] [36]. As there has been seemingly more research conducted in the area of natural language processing, such an approach as to convert the spoken audio to text is plausibly useful in understanding the weak points of the speaker. Recent works have combined automatic speech recognition with natural language processing to build grading models for English Language [21]. In the last few years, pre-trained vectors such as GLoVe[29] have gained importance in the field of natural language processing. Having the text to be read out provides the advantage of comparing the speaker's audio converted to text to it with the help of pre-trained vector representation of words to help predict the grades.

Speech recognition systems have used tonal features to improve their model performances [31] [9]. Since tones are an important component of pronunciation in Chinese language learning, we also consider the use of tonal features in this work for the task of predicting teacher-provided scores.

4.3 Dataset

The data set used in this work was obtained from an undergraduate Chinese Language class taught to non-native speakers. The data was collected from multiple classes with the same instructor. The data is comprised of assignments requiring students to submit an audio recording of them reading a predetermined prompt as well as answering open-ended questions. For this work, we focus only on the reading part of the assignment, observing the audio recordings in conjunction with the provided text prompt of the assignment.

For the reading part of the assignment, students were presented with 4 reading tasks which were meant to be read out loud and recorded by the student and submitted through the course's learning management system. Each reading task consists of one or two sentences about general topics. The instructor downloaded such audio files, listened to each, and assessed students based on two separate grades pertaining to fluency and accuracy of the spoken text. Each of these grades are represented as a continuous-valued measure between 0-5; decimal values are allowed such that a grade of 2.5 is the equivalent of a grade of 50% on a particular outcome measure. This dataset contains 304 audio files from 128 distinct students over four distinct sentence reading tasks. Each audio is taken as a separate data point, so each student has one to four audio files. Each sample includes one of the reading tasks read by the student along with the intended text of the reading prompt.

4.4 Feature Extraction

4.4.1 Pre-processing

Refer to Section 3.2.1 in Part 1.

4.4.2 Audio Features

Audio Features have already been explained in Part 1. See section Audio feature extraction Section 3.3.1.

4.4.3 Text Features

From the audio data, we also generate a character-representation of the interpreted audio file using openly-available speech recognition tools. The goal of this feature extraction step is to use speech recognition to transcribe the words spoken by each student to text that can be compared to the corresponding reading prompt using natural language processing techniques; the intuition here is that the closeness of what the Google speech-to-text model is able to interpret to the actual prompt should be an indication of how well the given text was spoken. Since building speech recognition models is not the goal of this paper, we used an off-the-shelf module for this task. Specifically, the SpeechRecognition library [40] in python provides a coding interface to the Google Web Speech API and also supports several languages including Mandarin Chinese. The API is built using deep-learning models and allows text transcription in real-time, promoting its usage for deployment in classroom settings. While, to the authors' knowledge, there is no detailed documentation describing the precise training procedure for Google's speech recognition model, it is presumably a deep learning model trained on a sizeable dataset; it is this later aspect, the presumably large number of training samples, that we believe may prove some benefit to our application. Given that we have a relatively small dataset, the use of pre-trained models such as those supplied openly through Google, may be able to provide additional predictive power to models utilizing such features.

With the Google-transcribed string, we segment the text into character-level components and then convert them into numeric vector representations for use in the models. In applications of natural language processing, the use of pre-trained word embeddings has become more common due to the large corpuses of data on which they were trained. Pre-trained models of word2vec[24] and Global Vectors for Word Representation (GloVe)[29], for example, have been widely cited in applications of natural language processing. By training on large datasets, these embeddings are believed to capture meaningful syntactic and semantic relationships between words through meaning. Similar to these methods, FastText[6] is a library created by Facebook's AI Research lab which provides pre-trained word embeddings for Chinese language. Each character or word is represented in the form of a 300 dimensional numeric vector. Once the segmented characters are obtained, these embeddings are used to convert each component to its vector space representation.

The embedding process results in a character level representation, but what is needed is a representation of the entire sentence. As such, once the embeddings are applied, all characters are concatenated together to form a large vector representing the entire sentence. The pipeline for text feature extraction is shown in Figure 4.3.

The Google Speech to text API is reported to exhibit a mean Word Error Rate (WER) of 9% [17]. To analyse the performance of Google's Speech to Text API on our dataset of student recordings, we randomly selected 15 student audio files from our dataset across all the four reading tasks and then transcribed them using the open tool. We then created a survey that was answered by the Teacher and 2 Teaching Assistants of the observed Chinese language class. The survey first required the participants to listen to each audio and then asked each to rate the accuracy of the corresponding transcribed text on a 10-point integer scale. The Intraclass Correlation Coefficient (ICC) was used to measure the strength of inter-rater agreement, finding a correlation of 0.8 (c.f. ICC(2,k) [18]). While this small study illustrates that the Google API exhibits some degree of error, we argue that it is reliable enough to be used for comparison in this work.

4.4.4 Tonal Features

Chinese is a tonal language. The same syllable can be pronounced with different tones which, in turn, changes the meaning of the content. To aid in our goal of predicting the teacher-supplied scores of fluency and accuracy, we decided to explore the observance of tonal features in our models. In the Mandarin Chinese language, there are four main tones. These tones represent changes of inflection (i.e. rising, falling,

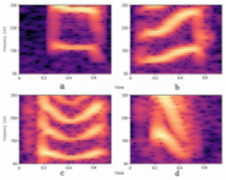


Figure 4.1: The figure shows the 4 tones in mel-spectrograms

or leveling) when pronouncing each syllable of a word or phrase. When asking Chinese Language teachers what are some of the features they look for while assessing student speech, tonal accuracy was one of the important characteristics identified. To extract the tones from the student's audio, we use the ToneNet [13] model which was trained on the Syllable Corpus of Standard Chinese Dataset (SCSC). The SCSC dataset consists of 1,275 monosyllabic Chinese characters, which are composed of 15 pronunciations of young men, totaling 19,125 example pronunciations of about 0.5 to 1 second in duration. The model uses a mel spectrogram (image respresentation of an audio in the mel-scale, see Figure 4.1) of each of these samples to train the model.The model uses a convolutional neural network and multi-layer perceptron (see Figure 4.2) to classify Chinese syllables in the form on images into one of the four tones. This model is reported to have an accuracy of 99.16% and f1-score of 99.11% [13]. To use the ToneNet on our student audio data, we first break the student audio into 1 second audio clips and convert them into spectrograms. We then feed these generated spectrograms to the ToneNet model to predict the tone present in each clip. The sequence of predicted tones is then used as features in our fluency and accuracy prediction models.The pipeline for tonal feature extraction is shown in Figure 4.4.

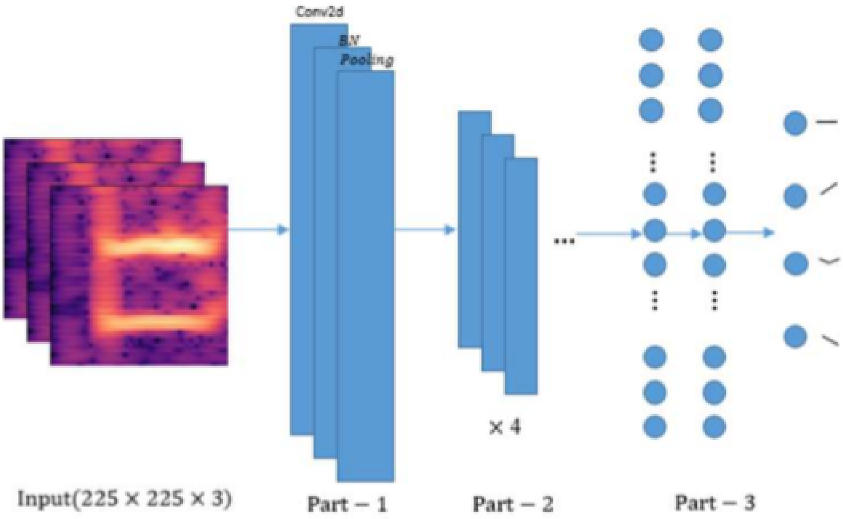


Figure 4.2: ToneNet Architecture

4.4.5 True Audio : Google Text to Speech API

As a final source of features for comparison in this work, we believed it may be useful to compare each student audio to that of an accepted “correct” pronunciation; however, no such recordings were present in our data, nor are they common to have in classroom settings for a given reading prompt. Given that we have audio data from students, and the text of each corresponding reading prompt, we we wanted to utilise Google’s text-to-speech API to produce a “true audio” - how Google would read the given sentence. Though this API is not equivalent to a native Chinese-speaking person, given that it is trained on large datasets, we believe it could help our models learn certain characteristics that differentiate between different grades;

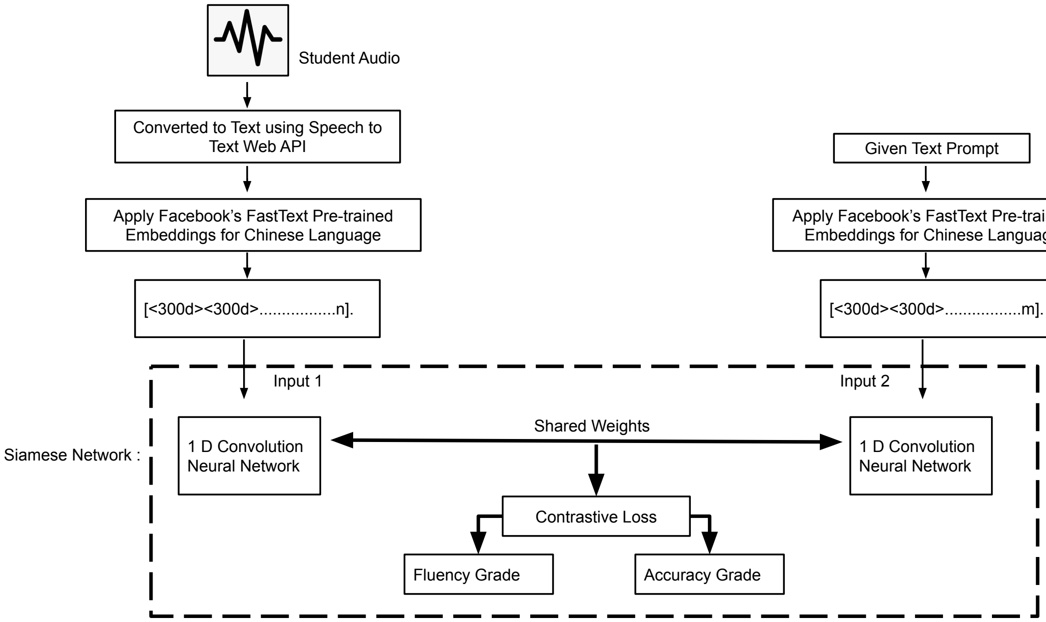

Student Audio
Converted to Text using Speech to Text Web API
Apply Facebook's FastText Pre-trained Embeddings for Chinese Language
[<300d><300d>.................n].
Given Text Prompt
Apply Facebook's FastText Pre-trained Embeddings for Chinese Language
[<300d><300d>.................m].
Input 1
Input 2
Siamese Network :
1 D Convolution Neural Network
Shared Weights
1 D Convolution Neural Network
Contrastive Loss
Fluency Grade
Accuracy Grade

Figure 4.3: The figure shows the steps involved in transforming audio to text features and feeding into the Siamese Network. Note: This is not a multitask model. Two individual models are used to each metric.

the features extracted from the true audio is particularly useful in training Siamese networks, as described in the next section, by providing a reasonable audio recording with which to compare each student response.

4.5 Models

In developing models to assess students based on the measures of accuracy and fluency, we compare three models of varying complexities and architectures (and one baseline model) using different feature sets described in the previous sections. Our baseline model consists of assigning the mean of the scores as the predicted value. We use a 5-fold cross validation for all model training.

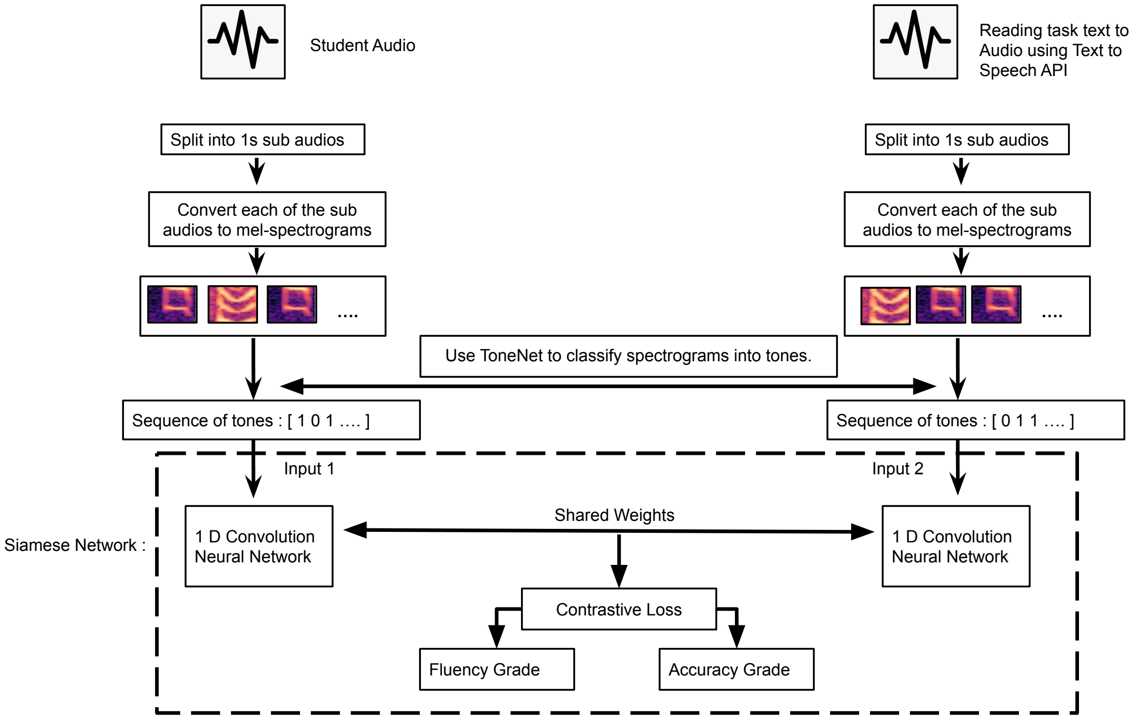

Student Audio
Reading task text to Audio using Text to Speech API
Split into 1s sub audios
Split into 1s sub audios
Convert each of the sub audios to mel-spectrograms
Convert each of the sub audios to mel-spectrograms
....
....
Use ToneNet to classify spectrograms into tones.
Sequence of tones : [1 0 1 ....]
Sequence of tones : [0 1 1 ....]
Input 1
Input 2
Siamese Network :
1 D Convolution Neural Network
Shared Weights
1 D Convolution Neural Network
Contrastive Loss
Fluency Grade
Accuracy Grade

Figure 4.4: The figure shows the steps involved in transforming audio to sequence of tones and feeding into the Siamese Network. Note: This is not a multitask model. Two individual models are used to each metric.

Aside from the baseline model, the first and second models explored in this work are the same as applied in previous research in developing models for assessing student accuracy and fluency in Chinese language learning [34]. These models consist of a decision tree (Using the CART algorithm [7]) and a Long Short Term Memory (LSTM) recurrent neural network. While previous work explored the use of audio features, labeled in this work as "PyAudio" features after the library used to generate them, this work is able to compare these additional textual and tonal feature sets. Similar to deep learning models, the decision tree model is able to learn non-linear relationships in the data, but also can be restricted in its complexity to avoid potential problems of overfitting. Conversely, the LSTM is able to learn temporal relationships from time series data as in the audio recordings observed in this work.

As in our prior research, a small amount of hyperparameter tuning was conducted on a subset of the data.

In addition to the three sets of features described, a fourth feature set, a cosine similarity measure, was explored in the decision tree model. This was calculated by taking the cosine similarity between the embedded student responses and the embedded reading prompts. This feature set was included as an alternative approach to the features described in Section 4.4.5 for use in the Siamese network described in the next section.

4.5.1 Siamese Network

The last type of model explored in this paper is a Siamese neural network. Siamese networks are able to learn representations and relationships in data by comparing similar examples. For instance, we have the audio of the student as well as the Google API-generated “true audio” that can be compared to learn features that may be useful in identifying how differences between these correlate with assigned fluency and accuracy scores. In this regard, the generated audio does not need to be correct to be useful in understanding how the different student audio recordings differ from each other and how these differences relate to their scores.

The network is comprised of two identical sub networks that share the same weights while working on two different inputs in tandem (e.g. the network observes the student audio data at the same time that it observes the generated audio data). The last layers of the two networks are fed into a contrastive loss function which calculates the similarity between the two audio recordings to predict the grades.

We experimented with different base networks within the Siamese architecture including a dense network, an LSTM, TCN(Temporal Convolution Network) and 1D Convolution Neural Network(CNN). However, the 1D CNN proved to be better

Model	**Features**	**Fluency**		**Accuracy**	
		Rho	MSE	Rho	MSE
Siamese Network	Text Features	0.073	0.665	**0.317**	**0.833**
	Tonal Features	**0.497**	**0.497**	-0.006	0.957
LSTM	PyAduio Features	0.072	1.139	-0.066	2.868
	Tonal Features	-0.096	0.648	0.042	0.929
	Text Features	-0.123	1.885	0.128	3.733
Decision Tree	PyAudio Features	-0.005	0.749	0.011	1.189
	Tonal features	0.285	0.649	0.107	0.960
	Text Features	0.090	0.794	0.261	0.998
	Cosine Similarity	0.037	0.674	0.162	0.984
	Baseline	-	0.636	-	0.932

Table 4.1: Results for the different models.

amongst all of them. There has been prior research showing the benefits of using CNNs on sequential data [2] [16]. We report the results for 1D CNN in this work.

4.5.2 Multi-Task Learning and Ensembling

Following the development of the decision tree, LSTM, and Siamese network models, we selected the two highest performing models across fluency and accuracy and ensembled their predictions using a simple regression model. As will be discussed in the next section, two Siamese network models (one observing textual features and the other the tonal feature set) exhibited the highest performance and were used in this process.

Our final comparison explores the usage of multitask learning [8] for the Siamese network. In this type of model, the weights of the network are optimized to predict both fluency and accuracy of speech simultaneously within a single model. Such a model may be able to take advantage of correlations between the labels to better learn distinctions between the assessment scores.

4.6 Results

In comparing the model results we use two measures to evaluate each model's ability to predict the Fluency and Accuracy grades: mean squared error (MSE) and Spearman correlation (Rho). A lower MSE value is indicative of superior model performance while higher Rho values are indicative of superior performance; this later metric is used to compare monotonic, though potentially nonlinear relationships between the prediction and the labels (while continuous, the labels do not necessarily follow a normal distribution as students were more likely to receive higher grades).

Table 4.1 illustrates the model performance when comparing models utilizing each of the three described sets of features (See Section 4.4). From the table, it can be seen that in terms of MSE and Rho, the Siamese network exhibits the best performance across both metrics. It is particularly interesting to note that the textual features are better at predicting the accuracy score (MSE=0.833, Rho=0.317), while the tonal features are better at predicting the fluency score (MSE=0.497, Rho=0.497).

Model	**Fluency**		**Accuracy**	
	Rho	**MSE**	**Rho**	**MSE**
Multitasking	0.129	0.603	0.313	0.915
Ensemble (Regression)	0.477	**0.490**	**0.34**	0.839

Table 4.2: Multitasking and Ensembled Siamese models

Table 4.2 shows the results for the multitasking and ensemble models. We see that the Siamese model with multitasking to predict both the fluency and accuracy scores do not perform better than the individual models predicting each. This suggests that the model is not able to learn as effectively when presented with both labels in our current dataset; it is possible that such a model would either need more data or a different architecture to improve. On the other hand, in a intutive sense the reason could also be that these two metrics are orthogonal and hence the model is not able to learn a generalized representation for both. The slight improvement in regard to Fluency MSE and Accuracy Rho exhibited by the ensemble model suggests that the learned features (i.e. the individual model predictions) are able to generalize to predict the other measure. The increase in Rho for accuracy is particularly interesting as the improvement suggests that the tonal features are similarly helpful in predicting accuracy when combined with the textual-based model.

4.7 Overlap Analysis

Since we break the student audio into continuous one second sub audios, it is possible that there may be loss of tonal information from one second to another. To improve upon our current method of tonal features, we explored the method of extracting features on overlapping windows. By overlap, we mean the next audio frame starts from a point in the previous frame so as to not miss out any transitional information. For example: For a 2 second audio, we break it into 2 consequent audios - each of 1s interval. When we do 50 percent overlap, the first audio is from 0 to 1s. The second audio is from 0.5 seconds to 1.5 seconds. The third audio is from 1 to 2s. Different percentages of overlap between two consecutive 1 second audio frames were experimented with - 25%, 50%, 75% . Our analysis shows that there is no improvement from the no overlap method (See Table 4.3).

	Fluency		**Accuracy**	
Method	**Rho**	**MSE**	**Rho**	**MSE**
No Overlap	0.497	0.497	-0.006	0.957
50% overlap	0.083	0.63	0.117	0.93
75% overlap	0.129	0.617	0.040	0.949
25% overlap	0.268	0.553	0.093	0.923

Table 4.3: Result of Tone Overlap Analysis

4.8 Data Augmentation

Recent advances in deep learning models have been largely attributed to the quantity and diversity of data gathered in recent years. Data augmentation is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data. In this work we show two types of augmentation methods to increase our data size (We tried other methods but failed to obtain significant results with them). The first method shown in Figure 4.5 involves injecting small amount white noise into the audio data such that the original content is still retained. The second augmentation method shown in Figure 4.6 involves shifting the pitch (but still retaining intelligibility). The results in table 4.4 show that augmentation, although provides us with more data, does not improve the model performance.

4.9 Bias Analysis

The dataset used to train the ToneNet model is the Syllable Corpus of Standard Chinese Dataset (SCSC). The creator is The Institute of Linguistics Chinese Academy

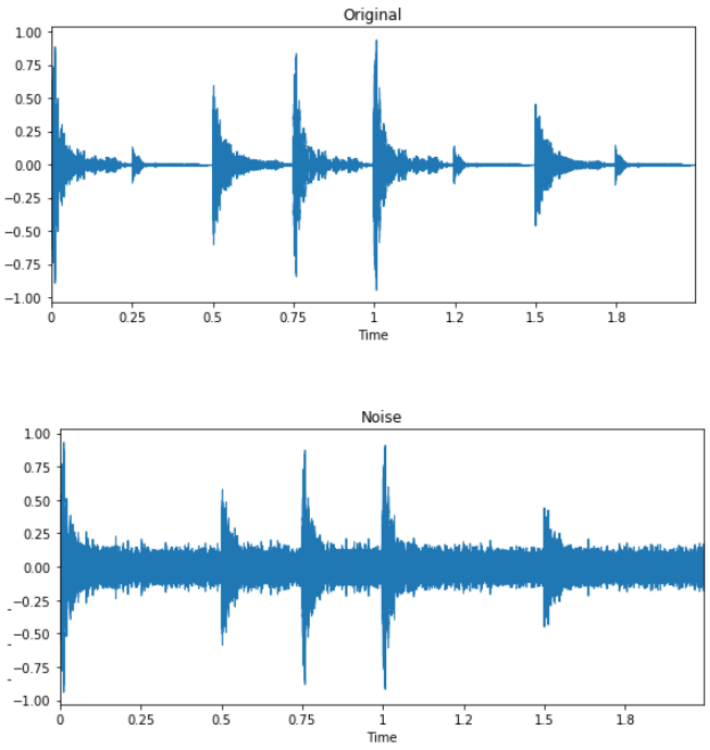


Figure 4.5: Adding Noise

of Social Sciences. Mandarin mono-syllable corpus is comprised by mono-syllable wave data, the list of mono-syllable and management software. The SCSC dataset consists of 1,275 monosyllabic Chinese characters, which are composed of 15 pronunciations of young men. Since the dataset consists of only audios trained on one gender, there exists a possibility of bias in the predictions. For example - the predictions for male students could be better than the predictions for female students. When building assessment tools, we do not want any biases existing in them. To make sure the system is unbiased, a t-test was performed on the results of the model based on gender.

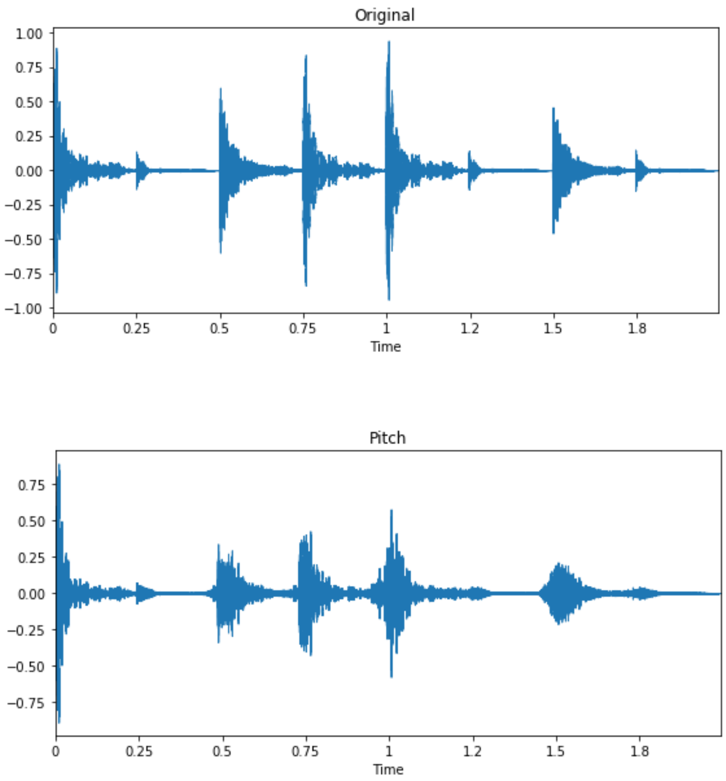


Figure 4.6: Pitch Shift

4.9.1 Gender Prediction

As a first step, the gender data of the students was required. Using the python library 'gender-guesser', the gender of the students based on their names was obtained. There are three possible values obtained from this tool for each name - male, female and unknown. Out of the 304 data points, we had gender value for 257 and unknown for the remaining. The number of female data points were 121 and the male were 136.

	Augmentation Type	Fluency		Accuracy	
		Rho	MSE	Rho	MSE
Text Siamese Model	**None**	0.073	0.665	**0.317**	**0.833**
	Adding White Noise	0.194	0.736	0.210	0.837
	Pitch Shift	0.270	0.701	0.217	0.877
Tone Siamese Model	**None**	**0497**	**0.497**	-0.006	0.957
	Adding White Noise	0.182	0.590	0.057	0.935
	Pitch Shift	0.020	0.624	0.102	0.929

Table 4.4: Augmentation results

4.9.2 T-test

The square of the errors for all these datapoints were calculated and seperated by gender. These were then used to calculate the t-test score. There was a not a significant difference in the scores for male (M=0.538, SD=0.807) and female (M=0.452, SD=0.8) gender. The results were not significant (p=0.396,statistic=-0.851) for 95% confidence interval.

4.10 Discussion

In chapter one, we saw that the use of audio features helped predict fluency and accuracy scores better than a simple baseline. In this study, the textual features and tonal features explored provide even better predictive power. A potential limitation of the current work is the scale of the data observed, and can be addressed by future research. The use of the pre-trained models may have provided additional predictive power for the tonal and textual features, but there may be additional ways to augment the audio-based features in a similar manner (i.e. either by using pre-trained models or other audio data sources). Similarly, audio augmentation methods may be utilized to help increase the size and diversity of dataset (e.g. even by simply adding random noise to samples).

Another potential limitation of the current work is in regard to the set of pre-trained speech recognition models provided through Google's APIs, additional performance biases may exist for speakers with different accents. Understanding the potential linguistic differences between language learners would be important in providing a feedback tool that is beneficial to a wider range of individuals. A deeper study into these properties of our assessment models would be needed before deploying within a classroom.

As Mandarin Chinese is a tonal language, the seeming importance and benefit of including tonal features makes intuitive sense. In both the tonal and textual feature sets, a pre-trained model was utilized which may also account for the increased predictive power over the audio features alone. As all libraries and methods used in this work are openly available, the methods and results described here present opportunities to develop such techniques into assessment and feedback tools to benefit teachers and students in real classrooms; in this regard, they also hold promise in expanding to other languages or other audio-based assignments and is a planned direction of future work.

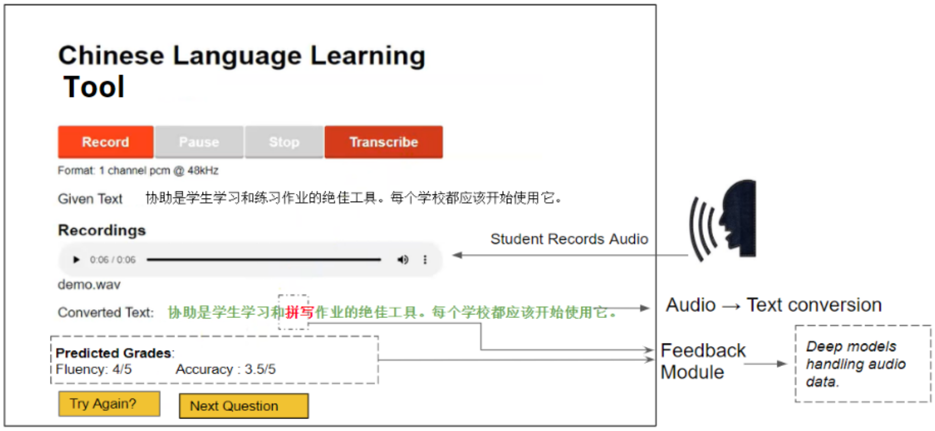
Chapter 5

Designing Tool

To incorporate the models to be used in a tool, a web application was built. As a fist version, the tool aims to transcribe the spoken text and display it to let the students know what they have spoken. It also aims to show the grades predicted by the system using our models. The main components of the web application are as follows.

5.1 Framework

The front end was developed with HTML and CSS. It mainly consists of the interface where the student views the reading task question and records her answer by clicking on record and stop buttons. Flask was used to connect the python elements to the front-end. Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. Figure 5.2 shows the framework of the tool built. The transcribe button, when pressed calls the python elements to perform the next steps.



Format: 1 channel pcm @ 48kHz

Given Text 协助是学生学习和练习作业的绝佳工具。每个学校都应该开始使用它。

demo.wav

Converted Text: 协助是学生学习和拼写作业的绝佳工具。每个学校都应该开始使用它。

Predicted Grades:
Fluency: 4/5 Accuracy : 3.5/5

Figure 5.1: The Prototype of the grading tool

5.2 Handling Input

The audio recorded then is passed to the python component. Here, the feature preprocessing occurs and the input is made ready for the pre-trained models to predict on. Once the predictions are made, the transcribed text and the grades are sent back to the front end to be displayed to the student.

5.3 Student Options

The student after viewing the scores has the option to try again or go to the next question.

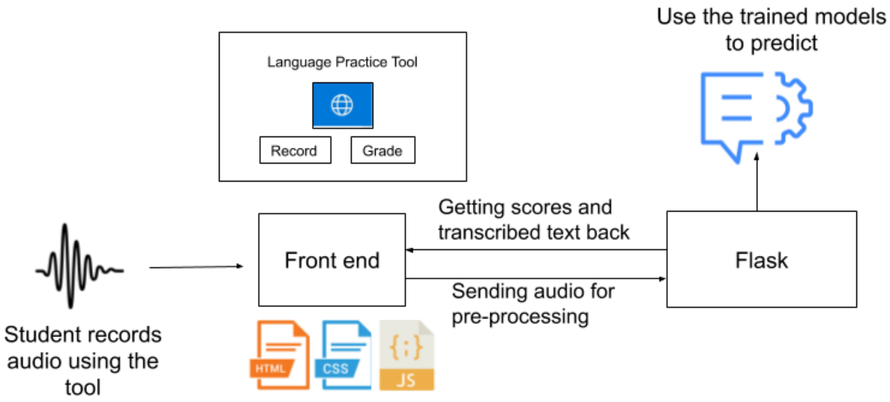

Language Practice Tool
Record
Grade
Use the trained models to predict
Getting scores and transcribed text back
Front end
Flask
Sending audio for pre-processing
Student records audio using the tool
HTML
CSS
JS

Figure 5.2: Framework for the web tool

Chapter 6

Conclusion and Future Work

To begin with, in chapter 3, we studied the different low-level audio features and their importance by performing an ablation study. From the study, the MFCC, Energy and the Chroma features stood out in terms of importance. We also build decision tree and LSTM models using all the audio features and observe that we are able to predict better than the baseline model. In chapter 4, we introduce tonal and text features, and their feature extraction process. Multple models were used on these features to predict for the scores. We see that the Siamese network using the tonal and text features are better at predicting the fluency and accuracy score respectively. Additional analyses including data augmentation, tone overlap models, and bias analyses were performed to explore different impact of our models. Though we see have the Siamese network performing with a rho of 0.497(fluency) and 0.317(Accuracy), the question is if they are good enough to give it to students and how the MSE error of the model affects the student scores. We acknowledge that there are chances that the model provides a score higher or lower than the true score. However, given that there is no such tool in the current setup, this stress-free practice tool could help students get an idea of their performance. We argue that having some feedback is better than having no feedback. Tutors in higher education are encouraged to ensure that feedback is provided to students on assessed work in a way that promotes learning and facilitates improvement[28]. By providing feedback with the transcribed text along with their scores, we give the students an idea of their performance and furthermore the ability to practice using the tool to improve their scores on the actual assignment.

Like mentioned before, there is a plethora of opportunity to work on such rich audio data set and bridge the gap in language learning in classroom setups by building suitable tools. This is a first step in that direction. Further work would involve conducting RCTs and the effect of such a tool in classrooms, and to study how students respond to it. In the computational aspect of it, further model architectures could be implemented to gain performance.

Bibliography

[1] A. S. AlKilabi. The place of reading comprehension in second language acquisition. *Journal of the College of Languages (JCL)*, (31):1–23, 2015.

[2] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[3] J. Bernstein. Automatic grading of english spoken by japanese students. *SRI International Internal Reports Project*, 2417, 1992.

[4] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. Automatic evaluation and training in english pronunciation. In *First International Conference on Spoken Language Processing*, 1990.

[5] J. C. Bernstein. Computer scoring of spoken responses. *The Encyclopedia of Applied Linguistics*, 2012.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[7] L. Breiman. *Classification and regression trees*. Routledge, 2017.

[8] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[9] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen. New methods in continuous mandarin speech recognition. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[10] A. T. Corbett, K. R. Koedinger, and J. R. Anderson. Intelligent tutoring systems. In *Handbook of human-computer interaction*, pages 849–874. Elsevier, 1997.

[11] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(2-3):121–130, 2000.

[12] V. Fridland, K. Bartlett, and R. Kreuz. Do you hear what i hear? experimental measurement of the perceptual salience of acoustically manipulated vowel variants by southern speakers in memphis, tn. *Language variation and change*, 16(1):1–16, 2004.

[13] Q. Gao, S. Sun, and Y. Yang. Tonenet: A cnn model of tone classification of mandarin chinese. *Proc. Interspeech 2019*, pages 3367–3371, 2019.

[14] A. Hake. The importance of learning mandarin chinese pronunciation, May 2016.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] W. Huang and J. Wang. Character-level convolutional network for text classification applied to chinese corpus. *arXiv preprint arXiv:1611.04358*, 2016.

[17] J. Y. Kim, C. Liu, R. A. Calvo, K. McCabe, S. C. Taylor, B. W. Schuller, and K. Wu. A comparison of online automatic speech recognition systems and the

nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*, 2019.

[18] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[19] K. Kyriakopoulos, K. M. Knill, and M. J. Gales. A deep learning approach to assessing non-native pronunciation of english using phone distances. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018, pages 1626–1630, 2018.

[20] D. Litman, H. Strik, and G. S. Lim. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309, 2018.

[21] A. Loukina, N. Madnani, and A. Cahill. Speech- and text-driven features for automated scoring of English speaking tasks. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 67–77, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[22] X. Lu, K. Ostrow, and N. Heffernan. Understanding the complexities of chinese word acquisition within an online learning platform. In *11th International Conference on Computer Supported Education*, 2019.

[23] A. Metallinou and J. Cheng. Using deep neural networks to improve proficiency assessment for children english language learners. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[25] N. Moustroufas and V. Digalakis. Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, 21(1):219–230, 2007.

[26] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1457–1460. IEEE, 1996.

[27] M. Nicolao, A. V. Beeston, and T. Hain. Automatic assessment of english learner pronunciation using discriminative classifiers. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5351–5355. IEEE, 2015.

[28] P. Orsmond, S. Merry, and K. Reiling. Biology students' utilization of tutors' formative feedback: a qualitative interview study. *Assessment & Evaluation in Higher Education*, 30(4):369–386, 2005.

[29] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[30] M. Pucher, A. Türk, J. Ajmera, and N. Fecher. Phonetic distance measures for speech recognition vocabulary and grammar optimization. In *Proc. the 3rd Congress of the Alps Adria Acoustics Association*, 2007.

[31] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan. Highly accurate mandarin tone classification in the absence of pitch information. In *Proceedings of Speech Prosody*, volume 7, 2014.

[32] H. Singer and R. B. Ruddell. Theoretical models and processes of reading. 1970.

[33] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, 2016.

[34] A. Varatharaj, A. F. Botelho, X. Lu, and N. T. Heffernan. Hao fayin: Developing automated audio assessment tools for a chinese language course. *The Twelfth International Conference on Educational Data Mining*, Jul 2019.

[35] R. K. Wagner, C. Schatschneider, and C. Phythian-Sence. *Beyond decoding: The behavioral and biological foundations of reading comprehension.* Guilford Press, 2009.

[36] S. Wei, G. Hu, Y. Hu, and R.-H. Wang. A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 51(10):896–905, 2009.

[37] D. M. Williamson. A framework for implementing automated scoring. In *Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA*. Citeseer, 2009.

[38] S. M. Witt. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. IS ADEPT*, 6, 2012.

[39] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009.

[40] A. Zhang. Speechrecognition · pypi. https://pypi.org/project/SpeechRecognition/. (Accessed on 10/02/2019).