Hǎo Fāyīn: Developing Automated Audio Assessment Tools for a Chinese Language Course

Ashvini Varatharaj
Worcester Polytechnic Institute
100, Institute Rd
Worcester,MA
avaratharaj@wpi.edu

Anthony F. Botelho
Worcester Polytechnic Institute
100, Institute Rd
Worcester,MA
abotelho@wpi.edu

Xiwen Lu
Worcester Polytechnic Institute
100, Institute Rd
Worcester,MA
oluxiwen@gmail.com

Neil Heffernan
Worcester Polytechnic Institute
100, Institute Rd
Worcester,MA
nth@wpi.edu

ABSTRACT

We present and evaluate a machine learning based system that automatically grades audios of students speaking a foreign language. The use of automated systems to aid the assessment of student performance holds great promise in augmenting the teacher's ability to provide meaningful feedback and instruction to students. Teachers spend a significant amount of time grading student work and the use of these tools can save teachers a significant amount of time on their grading. This additional time could be used to give personalized attention to each student. Significant prior research has focused on the grading of closed-form problems, open-ended essays and textual content. However, little research has focused on audio content that is much more prevalent in the language-study education. In this paper, we explore the development of automated assessment tools for audio responses in a college-level Chinese language-learning course. We analyze several challenges faced while working with data of this type as well as the generation and extraction of features for the purpose of building machine learning models to aid in the assessment of student language learning.

Keywords

Audio Analysis, Automatic Grading, Machine Learning

1. INTRODUCTION

Learning 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 out loud; these verbal tests are 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 [6]. Language learning is a common part of many educational systems, and language classes often involve assignments which are related to speaking tasks. This gives us, as researchers and developers of tools to aid in the teacher assessment of students, an opportunity to collect and utilise language students' audio responses. The audio data of language learners are rich data-sets which provide insights about how well students are acquiring language skills based on the pronunciations and quality of speech. These data-sets may also identify problem areas where a student may be in need of improvement. However, collecting this data can be challenging because the data is not commonly stored uniformly and all in one place. Since oral tests are given in class, recording and storing this data would add an overhead to the teacher's responsibilities. Additionally, this process can occur more efficiently. The students are all evaluated for on similar measures like the standardized exams that test pronunciation and fluency. By automating the process of grading students, 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). The grading systems which have been researched previously are usually for more closed-form responses. 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 [5].

This paper presents an exploratory analysis representing an initial step toward developing automated assessment tools for language learning audio responses. In this paper, we explore the development of automated graders of student audio responses from a Chinese language class. 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?

Ashvini Varatharaj, Anthony Botelho, Xiwen Lu and Neil Heffernan "Hao Fa Yin: Developing Automated Audio Assessment Tools for a Chinese Language Course" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 663 - 666

2. 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 of pronunciation quality (cf. [7]). With Deep Learning improving the performance of many tasks, it has been widely used in the Automatic Speech Recognition and Scoring tasks [4]. 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.

3. DATASET

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. All work 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 data-set, the teacher followed a rubric when assessing each student that is reported in Table 1

3.1 Pre-processing

A series of pre-processing steps were applied to the raw audio responses for use in this work. Among the 60 responses, in 12 cases, responses included multiple students speaking; it is likely that permission was 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 averaging the two scores. The other 9 cases were 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 took the average of the two channels of the stereo data.

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

Table 1: The grading rubric used to assess student audio responses.

4. METHODOLOGY

Our methodology includes feature extraction and building models for predicting teacher-provided grades for student audio responses. We compared two non-linear models to a simple baseline. While the grade labels were provided on an 11-point scale (0-10), this scaling is non-linear 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.

4.1 Feature Extraction

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. The features used 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 split 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 varied 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 2. For each time window, we extracted the 34 features, where several of the features described in Table 2 are described using a multi-valued numeric vector indicated through the Feature ID column.

4.2 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) [3] network, as it is designed to model complex temporal relationships within sequential data. This 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 chose a network structure with 3 hidden layers. The input of 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 rather than a classification task. We chose this structure to 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; exploring additional model structures is planned for future work. 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 data-set and measure performance using RMSE and Spearman correlation.

4.3 Decision Tree Model

To contrast the deep learning network, we also compare a decision tree model which has the capacity to learn non-linear

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

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 took the average across each of the 34 features across each response and used it to predict the teacher-provided grade. We applied a decision tree regressor using the CART algorithm [1]. Similar to the deep learning model, we evaluated 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 using the training data.

4.4 Baseline Method

We compare each the LSTM model and decision tree model to a simple baseline using a majority class model. For this method, we take the average grade provided by the teacher and use this as a prediction for every sample. It can be used to evaluate how well our models perform in comparison to a model that incorporates no audio information.

5. RESULTS

We report model performance using measures of RMSE, a measure of prediction error, and Spearman correlation (Rho). The performance of each of our models in predicting the audio response grades is reported in Table 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. These results demonstrate that the LSTM is the superior model, although the values suggest that there is still room for improvement. As the grade labels follow 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. We further explore each of these models through an ablation study.

6. 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. Table 4 reports the results of this study across both the LSTM and Decision Tree models. The rows in the table 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: Audio Grade prediction: average 5-fold RMSE and R2 score for the models

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. 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 the LSTM model, the 3 features which cause the largest decrease in the RMSE are the 12 Chroma features, the Energy Entropy, and the Zero Crossing Rate feature. However, comparing the Spearman's correlation measure (rho) does not follow the same trend as the RMSE. In the case of LSTM, most of the models have a rho value more than the model with all the features. This may suggest overfitting within the model, particularly as some of the features similarly lead to improvements in RMSE when removed.

7. DISCUSSION

As can be seen from the decision tree model as well as the LSTM model, the energy related features has a significant impact on the evaluation of the pronunciation. In [2] it was shown that 'formants' are bands of energy around a particular frequency which characterizes different resonances of the vocal tract and it helps understand pronunciation of vow-

	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 4: Ablation Study results from the LSTM and Decision Tree model

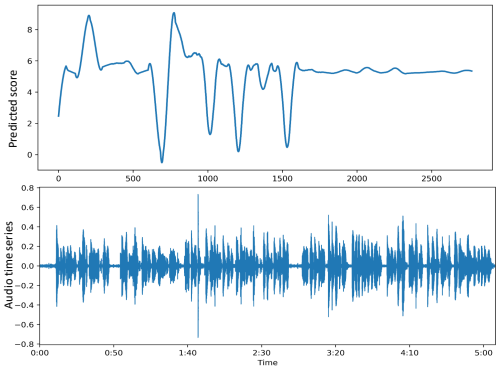

Figure 1: The Waveform from an audio of a student answer along with the predictions by the LSTM model at each window.

els. The second feature that has an impact in the decision tree is the MFCC feature. The MFCC features accurately characterizes the envelope of the short time power spectrum which 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 provides information related to the 12 musical octaves. Both the LSTM and the decision tree model seem to show an increase in the RMSE when these features are removed.

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 suggest a high grade (e.g. well-pronounced words) and sections that suggest a low grade (e.g. poor pronunciation or areas of silence); an example of this is illustrated in Figure 1. In this 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.

8. CONCLUSION AND FUTURE WORK

Based on the results we plan to follow the following steps for the future. First, exploring additional model architectures may lead to more accurate assessment tools. Second, we plan on incorporating contextual knowledge since 40 percent of the grade includes content. Converting the audio to text and extracting text-related features could potentially provide more understanding of the evaluation of the audio. And finally, using LSTMs the scores for each segment of audio can be graded and we plan on using it to aid students in self-assessment and to help teachers learn where their students need further aid. This work is an initial step toward the development of automated assessment tools designed to aid language learning students and teachers. We hope that further in-depth analyses of different combinations of the features will better help understand these relationships.

9. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.