# Supporting Teacher Assessment in Chinese Language Learning Using Textual and Tonal Features

Blinded for Review[1], Blinded for Review[1], Blinded for Review[2], Blinded for Review[1], and Blinded for Review[1]

[1] Blinded for Review {Blinded for Review}
[2] Blinded for Review
{Blinded for Review}

**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 a model designed to grade student audio assignments with the purpose of incorporating such a model into tools focused on helping both teachers and students in real classrooms. Building upon our prior work which explored features extracted from audio, the goal of this work is to explore additional features derived from tone and speech recognition models to help assess students on two outcomes commonly observed in language learning classes: fluency and accuracy of speech. In addition to the exploration of features, this work explores the application of Siamese deep learning models for this assessment task. 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.

**Keywords:** Audio Processing · Natural Language Processing · Language learning.

## 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[1–3].

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.

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.

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.

## 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 [4] 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 [5]. Other works have previously explored the similarity and differences between aspects of speech. In [6], 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 [7] 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 [8].

Many approaches having been explored in such works, starting from Hidden Markov Models [9] to more recently applying deep learning methods [6] 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 [10] [11][12] [9] [13] [14] [15]. 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 [16]. In the last few years, pre-trained vectors such as GLoVe[17] 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 [18] [19]. 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.

## 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 305 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 Feature Extraction

### 4.1 Pre-processing

The audio files submitted by students were of varying formats including mp3, m4a. The ffmpeg [20] python library is used to convert these audio inputs into a raw .wav format required by the speech models. Once converted to a standardized format the audio channels are reduced to 1 (reducing the 'left' and 'right' stereo channels to a single 'mono' channel) and the sampling frequency is set to 16000. The resulting audio file is then used to extract additional features.

While prior work [5] explored much of this process of processing and extracting features from audio, the current work intends to expand upon this prior work by additionally introducing textual and tonal features into our grading model. The next two sections detail the feature extraction process for audio, tonal. and textual features using the processed dataset.

### 4.2 Audio Features

Audio feature extraction is required to obtain components of the audio signal which can be used to represent acoustic characteristics in a way a model can understand. The audio files are converted to vectors which can capture the various properties of the audio data recorded. The audio features were extracted using an openly-available python library [21] that breaks each recording into 50 millisecond clips, offsetting each clip by 25 milliseconds of the start of the preceding clip (creating a sliding window to generate aggregated temporal features using the observed frequencies of the audio wave). A total of 34 audio-based features, including Mel-frequency cepstral coefficients (MFCC), Chroma features, and Energy related features, were generated as in previous work [5].

### 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 [22] 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 use another open library[23] to 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)[17], 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[25] 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 Google Speech to text API is reported to exhibit a mean Word Error Rate (WER) of 9% [26]. 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) [27]). 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 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, 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 [28] 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 spectrogram (image respresentation of an audio) of each of these samples to train the model. The model uses a convolutional neural network and multi-layer perceptron 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% [28]. 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.

### 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; 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.

## 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.

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 [5]. These models consist of a decision tree (Using the CART algorithm [29]) 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 4.5 for use in the Siamese network described in the next section.

### 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 and 1D Convolution Neural Network(CNN). There has been prior research showing the benefits of using CNNs on sequential data [30] [31]. We wanted to explore their performance on our data. We report the results for 1D CNN in this work.

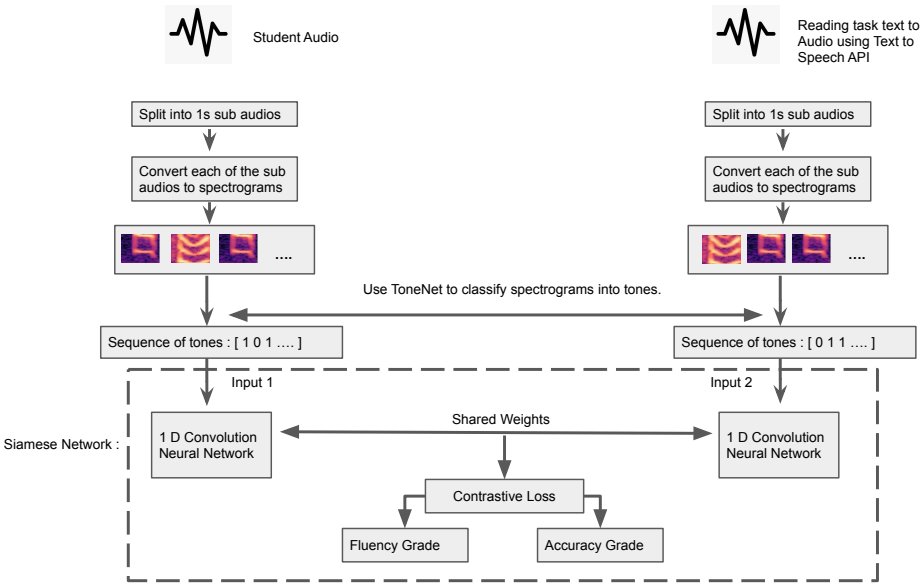


**Fig. 1.** The figure shows the steps involved in transforming audio to sequence of tones and feeding into the Siamese Network

### 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 [32] 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.

## 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 non-linear relationships between the prediction and the labels (while continuous, the

| 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 1.** Results for the different models.

labels do not necessarily follow a normal distribution as students were more likely to receive higher grades).

Table 1 illustrates the model performance when comparing models utilizing each of the three described sets of features (See Section 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 2.** Multitasking and Ensembled Siamese models

Table 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. The slight improve-

ment 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.

## 7 Discussion and Future Work

In [5], it was found that the use of audio features helped predict fluency and accuracy scores better than a simple baseline. In this paper 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 exploration of fairness among the models. It was described that the ToneNet model used training samples from men, but not women; as with any assessment tool, it is important to fully explore any potential sources of bias that exist in the input data that may be perpetuated through the model's predictions. 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 the fairness 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.

## 8 Acknowledgements

Blinded for Review

## References


1. Harry Singer and Robert B Ruddell. Theoretical models and processes of reading. 1970.
2. Richard K Wagner, Christopher Schatschneider, and Caroline Phythian-Sence. *Beyond decoding: The behavioral and biological foundations of reading comprehension*. Guilford Press, 2009.
3. Ahmed Shakir AlKilabi. The place of reading comprehension in second language acquisition. *Journal of the College of Languages (JCL)*, (31):1–23, 2015.
4. Silke M Witt. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. IS ADEPT*, 6, 2012.
5. Blinded for review.
6. Konstantinos Kyriakopoulos, Kate M Knill, and Mark JF 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.
7. Michael Pucher, Andreas Türk, Jitendra Ajmera, and Natalie Fecher. Phonetic distance measures for speech recognition vocabulary and grammar optimization. In *Proc. the 3rd Congress of the Alps Adria Acoustics Association*, 2007.
8. X Lu, KS Ostrow, and NT Heffernan. Understanding the complexities of chinese word acquisition within an online learning platform. In *11th International Conference on Computer Supported Education*, 2019.
9. Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. Automatic evaluation and training in english pronunciation. In *First International Conference on Spoken Language Processing*, 1990.
10. Diane Litman, Helmer Strik, and Gad S. Lim. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309, 2018.
11. Jared C Bernstein. Computer scoring of spoken responses. *The Encyclopedia of Applied Linguistics*, 2012.
12. Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(2-3):121–130, 2000.
13. J Bernstein. Automatic grading of english spoken by japanese students. *SRI International Internal Reports Project*, 2417, 1992.
14. Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009.
15. Si Wei, Guoping Hu, Yu Hu, and Ren-Hua Wang. A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 51(10):896–905, 2009.
16. Anastassia Loukina, Nitin Madnani, and Aoife 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, September 2017. Association for Computational Linguistics.
17. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
18. Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan. Highly accurate mandarin tone classification in the absence of pitch information. In *Proceedings of Speech Prosody*, volume 7, 2014.

19. C Julian Chen, Ramesh A Gopinath, Michael D Monkowski, Michael A Picheny, and Katherine Shen. New methods in continuous mandarin speech recognition. In *Fifth European Conference on Speech Communication and Technology*, 1997.
20. Download ffmpeg. http://ffmpeg.org/download.html. (Accessed on 10/02/2019).
21. Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015.
22. Anthony Zhang. Speechrecognition · pypi. https://pypi.org/project/SpeechRecognition/. (Accessed on 10/02/2019).
23. Junyi Sun. jieba · pypi. https://pypi.org/project/jieba/. (Accessed on 10/02/2019).
24. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
25. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
26. Joshua Y Kim, Chunfeng Liu, Rafael A Calvo, Kathryn McCabe, Silas CR Taylor, Björn W Schuller, and Kaihang Wu. A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*, 2019.
27. Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.
28. Qiang Gao, Shutao Sun, and Yaping Yang. Tonenet: A cnn model of tone classification of mandarin chinese. *Proc. Interspeech 2019*, pages 3367–3371, 2019.
29. Leo Breiman. *Classification and regression trees*. Routledge, 2017.
30. Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
31. Weijie Huang and Jun Wang. Character-level convolutional network for text classification applied to chinese corpus. *arXiv preprint arXiv:1611.04358*, 2016.
32. Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.