International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 5, No. 2, 2022

English Speech Scoring System Based on Computer Neural Network

Xianxian Wu[1, *], Yan Zhang[2]

[1] School of Foreign Languages, Taishan University, Taian, Shandong 271000, China
[2] School of Information Science and Technology, Taishan University, Taian, Shandong 271021, China
[*] **Corresponding author**: Xianxian Wu (Email: wuxianxian1980@163.com)

Abstract: In English phonetics teaching, in order to improve students' English phonetics quality, a computer neural network based English phonetics scoring method is proposed. First, the frequency domain spectrogram is used as the data input to construct a convolutional neural network model at the word and phoneme levels to detect speech similarity. Then the original sound time domain waveform is used as the data input, which is converted into text through neural network to detect the text difference. Finally, we combine the two with the assigned weight to give a relatively objective comprehensive pronunciation score. The simulation results show that the method is accurate and practical, and can promote the standardization of students' English pronunciation.

Keywords: Evaluation of english pronunciation quality, CNN, Fourier transform, English education.

1. Introduction

With the development of artificial intelligence recognition technology, intelligent human-computer interaction models and methods are gradually applied to various industries, and computer aided language teaching has received extensive attention. Especially in improving English pronunciation ability [1], it is of great significance to provide valuable guidance, evaluation and feedback to non-native English students and catch pronunciation errors [2]. Therefore, we need a software system based on artificial intelligence automatic scoring system of neural network to evaluate students' pronunciation accuracy, pronunciation quality and fluency [3, 4].

The speech scoring system consists of two parts: word phoneme speech matching and text similarity after automatic speech recognition (ASR) [5, 6]. Speech matching compares the characteristics of the test voice and the original voice, and gives scores according to the similarity [7]. It is usually difficult to manually propose the pronunciation characteristics of words or phonemes from the digital data of sound. In this paper, the method of using convolutional neural network (CNN) [8, 9] to extract spectrogram is discussed.

Automatic speech recognition generates time aligned word sequences for input speech, which is a technology to convert human voice into text. At present, the research on the technology of speech text transfer mainly focuses on the feature extraction and recognition of speech signals. Speech signal feature extraction focuses on the extraction of native speech signal features through signal analysis technology. Speech feature recognition focuses on the training of speech signal features, and uses machine learning to recognize semantics and convert them into characters. After the speech is recognized as a text, it is compared with the original text information to calculate the similarity of the two paragraphs of text [10], and the results are presented by scores.

The objective score that can reflect the quality of pronunciation can be obtained by weighted summation of the testee's speech fit and the text similarity after automatic speech recognition..

2. English Speech Recognition

2.1. Speech Feature Extraction

The original voice signal read from the audio file or microphone is a one-dimensional array. The length of the array is determined by the audio length and the sampling rate. The value represents the amplitude of the sound, according to which the sound waveform can be drawn. As shown in Figure 1, the above figure shows the sound waveform of the word "eight" with the adoption rate of 16000HZ. Fourier transform can transform time-domain signal into frequency-domain signal [11, 12]. After the original signal is divided into frames and windowed, many frames can be obtained. Fast Fourier transform is performed on each frame to convert time-domain signal into frequency-domain signal. The frequency domain signals of each frame can be stacked in time to obtain the acoustic spectrogram. As shown in the middle of Figure 1. In this way, one-dimensional data can be converted to two-dimensional data. The bottom figure in Figure 1 is to convert the power spectrogram (amplitude square) into decibel (dB) units.

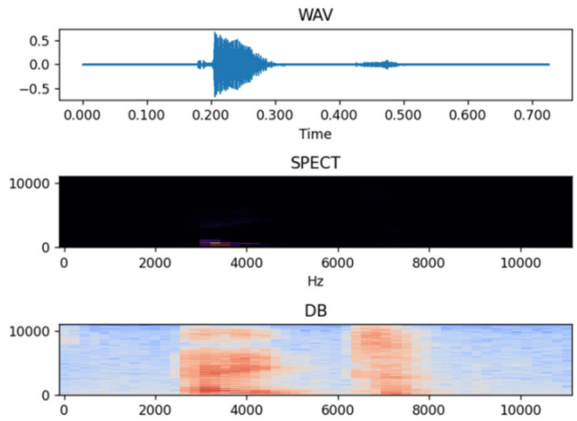

Figure 1. Time Domain and Frequency Domain Comparison of Pronunciation

2.2. Convolutional Neural Network

Convolutional neural network (CNN) shows a strong ability in image processing [13, 14], so this study uses the spectral characteristics of audio signal to train speech recognition model with convolutional neural network. The convolution neural network structure is shown in Figure 2.

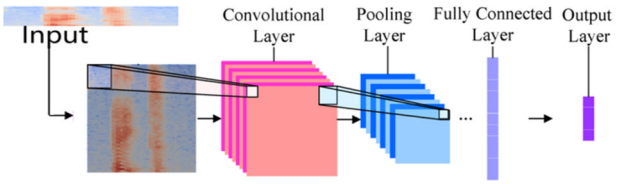

Figure 2. Structural Diagram of Convolutional Neural Network

2.2.1. Input Layer

Convolutional neural network is suitable for image processing, but in this study, the original sound file is one-dimensional data, so it is first converted into a spectrum map through fast Fourier transform, and then reduced to the 56X14 image matrix as the input layer of the neural network.

2.2.2. Convolution Layer

Convolution layer is an important part of CNN. It extracts the characteristics of data by convolving the data input from the previous layer. In the process of convolution movement, the eigenvalue will be calculated by convolution with the convolution kernel, and the result will be non-linear by ReLU() function.

2.2.3. Input Layer

The purpose of pooling is to filter the feature information extracted from the previous convolution layer, reduce the amount of data, remove the interference noise, and reduce the calculation amount of neural network. Therefore, after the data is processed by pooling layer, the data volume will shrink.

2.2.4. Input Layer

The role of the full connection layer is to integrate the data features after the previous convolution and pooling, and transform them by using affine transformation and nonlinear mapping. Finally, the classifier is selected to predict the sample category.

2.3. CNN English Speech Recognition Process

First, the speech signal is preprocessed. As mentioned earlier, the transformation of oscillogram into spectrogram is the key operation of preprocessing, so the parameters of FFT need to be determined. English speech is composed of factors. Generally, the duration of phonemes is more than 50 milliseconds, so the length of the frame should be less than 50 milliseconds to ensure the stability of the signal within the frame [15]. At the same time, it is required to ensure that there are enough vibration cycles in the frame to express the frequency information, so the length of the frame should not be less than 20 milliseconds. This is because the fundamental frequency of human speech is about 100 Hz. After recognition, the spectrogram is scaled, and the input layer is input into the CNN network for training to obtain a stable English speech recognition model. The main process is shown in Figure 3.

3. Pronunciation Quality Score

The assessment of English pronunciation quality can be divided into subjective assessment and objective assessment. The subjective evaluation of English pronunciation quality refers to the evaluation activities carried out by language expression experts or teachers on the accuracy of students' English pronunciation according to their accumulated professional pronunciation self-restraint. Under normal circumstances, the subjective evaluation follows the evaluator's own cognitive impression, compares and evaluates the students' English pronunciation according to their accumulated pronunciation cognition, and gives a comprehensive evaluation score.

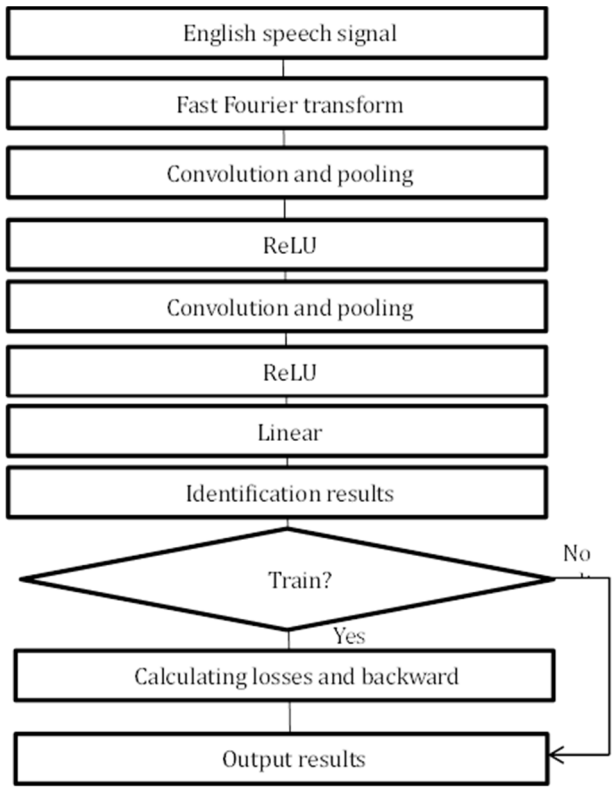

Figure 3. CNN English Speech Recognition Process

The objective evaluation needs the help of computer evaluation system in the form of artificial intelligence. Compared with the frequency of judgment errors in subjective evaluation, objective evaluation can more deeply reflect the fairness of pronunciation quality audit. The objective evaluation of English pronunciation consists of phoneme, word level evaluation and English sentence level evaluation. According to the teaching requirements, the two parts are multiplied by the weight coefficient and added to get the final score.

For the evaluation of words and phoneme levels, this research trains the standard voice through convolutional neural network algorithm, extracts features to create a neural network model, and then compares the testee's voice information with the standard voice after the operation of the neural network model, obtains the similarity with the standard voice features, and converts the absolute score according to the similarity. The relative ranking score can also be calculated based on the absolute score and the total number of people tested.

For the evaluation of English sentence level, by training the automatic English speech recognition neural network, or using the existing mature automatic speech recognition model, the standard speech content to be tested is recognized as text through the model, and then the testee's voice is also

recognized as text through the model, and the similarity between the two texts is compared to give an objective score.

4. Example Simulation

4.1. Experimental Environment and Data Set

The experimental environment of this paper is a desktop computer, the system is Windows 11, and the computer hardware is Core I7 2.5GHz processor, with 16GB memory. The experiment is implemented in Python 3.8 programming language under Pycharm development environment. The version of pytorch used is 1.21. The adopted dataset is speech commands dataset, with a total of 30 words and 64721 voice data. The length of each audio data segment is about 1 second.

In this experiment, 10 words from zero to nine and 23666 voices are selected to build a data set, 70% of which is used as the training set to train the model, and the remaining 30% of which is used as the test set to evaluate the performance. Each voice data is converted into a 56X14 pixels spectrogram after fast Fourier transforms. 24 sample examples are shown in Figure 4.

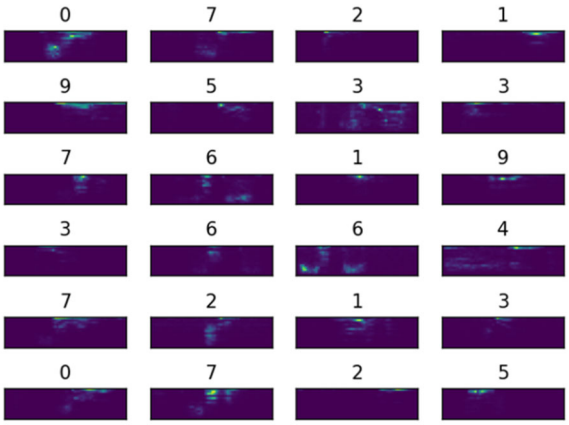

Figure 4. Word Pronunciation Samples Data

4.2. Word and Phoneme Level Evaluation Model Structure

The word and phoneme level evaluation model uses two 2d convolution layers, one fully connected linear layer. Each convolution layer is activated using the activation function ReLu function, followed by a maximum pooling layer. The basic process of the training model is as follows:

Input-->Conv2d-->Relu-->MaxPool2d-->
Conv2d-->Relu-->MaxPool2d-->Linear-->Output

After model training, use torch Save (net, path) to a file for future use.

The use process of the model is as follows:

(1) Load Model

```
model= torch.load(path)
```

(2) Turn on the microphone and record sound data or load waveform file data.

```
audio = pyaudio.PyAudio()
stream = audio.open(format=pyaudio.paFloat32,
channels=1, rate=16000, input=True,
frames_per_buffer=1600)
```

(3) Perform fast Fourier transform on sound data and convert it into acoustic spectrogram

```
spect=np.abs(librosa.stft(wav,n_fft=512,hop_length=256,
win_length=512))
```

(4) Adjust the spectrogram to standard size

(5) Use the model to process the data and output the classification tensor

```
output = model(input)
```

(6) Record the specific data value of the output classification results. The larger the value is, the higher the degree of fit with the model words. Score is obtained according to the degree of fit.

4.3. English Sentence Level Evaluation Model

The pre-trained Speech To Text Silero model is used for English sentence level evaluation in this experiment [16]. The model is robust to noise and low sampling rate. The basic use process of the model is as follows:

(1) Load the model.

```
model = torch.jit.load("./en_asr.jit")
model.eval()
```

(2) Load standard speech and use the model to recognize it as Text A.

```
def wav2txt(path):
wav, sr = torchaudio.load(path)
wav=torchaudio.transforms.Resample(orig_freq=sr,new_f
req=16000)(wav)
wav=wav.reshape(1,-1)
output = model(wav)
mx = torch.argmax(output[0].cpu(), axis=1)
ss=ls=''
for i in mx:
ll=model.labels[i.item()]
if (ll!='_') and (ll!=ls):
ss+=ll
ls=ll
return ss
txtA=wav2txt("standard.wav")
```

(3) Turn on the microphone and record the voice data of the tester, or load the recorded waveform file.

(4) Use the model to recognize voice data and convert it to text B.

```
txtB=wav2txt("test.wav")
```

(5) Compare the character differences between Text A and B and calculate the score [17].

```
difflib. SequenceMatcher(None, txtA,txtB).quick_
ratio()*100
```

4.4. Test results

The experimental results show that this method can effectively capture and separate English speech features with high accuracy and practicability. This machine scoring is helpful to measure the voice level of the tested.

5. Conclusion

Studying the evaluation mechanism and method of English pronunciation is of great significance in improving the effectiveness of English pronunciation teaching, and can grade students' pronunciation with a unified judgment basis, which can promote the standardization of students' English pronunciation. Based on the original time domain waveform data input, this paper gives two objective evaluation methods for English speech scoring, namely, the CNN model is used to detect the similarity of words and phoneme levels and the recognition rate of English sentence levels. Finally, the student's speech score is obtained.

In the case simulation part, using the existing data set, the proposed method is evaluated and verified at the code level,

the CNN model is constructed and trained, and the application of the model is tested. The results show that the proposed method is effective.

Acknowledgment

This work is supported by Research on the Integration of Information Technology, Chinese Culture and English Education in Primary and Secondary Schools under the Background of Double Reduction Policy (Grant No. TJk202106ZX037).

References

[1] Dalia Lisette Aguilar Vacacela, Maria Rossana Ramirez. "Self-awareness Strategy Using Podcasting to Improve Tense and Lax Vowel Pronunciation Sounds in Beginner EFL-Adult Learners," Journal of Foreign Language Teaching and Learning, vol. 5, no.1, pp. 79-98, 2020.

[2] Le Tian. "Research and Application of Interactive Teaching Strategies for College English Phonics in the Context of Internet plus," Modern English, no.16, pp. 103-105, 2021.

[3] S. Misirov, "The peculiarities of teaching English pronunciation in elementary classes (GRADES)[J]," Scientific Bulletin of Namangan State University, vol. 1, no. 2, p. 63, 2019.

[4] T. Isaacs and L. Harding, "Pronunciation assessment," Language Teaching, vol. 50, no. 3, pp. 347–366, 2017.

[5] CHEN Xiao-hong and TENG Hua, "Research on English speech recognition based on deep machine learning," Journal of Guiyan University Natural Sciences, vol. 16, no.3, pp.1-33, 2021.

[6] Wenjuan He , "Automatic Error Detection Method of Embedded English Speech Teaching Recognition System under the Background of Artificial Intelligence," Mobile Information Systems, 2022.

[7] Ming Liu and Lei Chen, "Similarity Calculation via Passage-Level Event Connection Graph, Applied Sciences," vol. 12, no. 9887, p.9887, 2022.

[8] Yu Zhao, "Control System and Speech Recognition of Exhibition Hall Digital Media Based on Computer Technology," Mobile Information Systems, vol. 2022, Article ID 7427899, 2022.

[9] Yuyuan Zhang, Wenjun Yan, Limin Zhang, Ling Ma, "Automatic Space-Time Block Code Recognition Using Convolutional Neural Network With Multi-Delay Features Fusion," IEEE Access, vol 9, pp. 79994 -80005, 2021.

[10] Yanmin Yu, Yongcai Lai, Ping Yan, Haiying Liu, "The Novel Sequence Distance Measuring Algorithm Based on Optimal Transport and Cross-Attention Mechanism," Shock and Vibration, vol. 2021, Article ID 3272119, 2021.

[11] Yu-Yi Lin, Wei-Zhong Zheng, Wei Chung Chu, Ji-Yan Han, "A Speech Command Control-Based Recognition System for Dysarthric Patients Based on Deep Learning Technology," Applied Sciences, vol. 11, no.2477, p. 2477, 2021.

[12] Wan-Ju Lin, Shih-Hsuan Lo, Hong-Tsu Young, Che-Lun Hung, "Evaluation of Deep Learning Neural Networks for Surface Roughness Prediction Using Vibration Signal Analysis," Applied Sciences, vol. 9, no. 7, p. 1462, Apr. 2019.

[13] Youhui Tian, "Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm," IEEE Access, vol. 8, pp. 125731-125744, Jan. 2021.

[14] X. Zhe, S. Chen, and H. Yan, "Directional statistics-based deep metriclearning for image classification and retrieval," Pattern Recognit., vol. 93, pp. 113–123, Sep. 2019.

[15] Yin Hui, Nadeu Climent, Hohmann Volker, "Pitch- and Formant-Based Order Adaptation of the Fractional Fourier Transform and Its Application to Speech Recognition," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2009, no. 1, p. 304579, Jan. 2009.

[16] Alexander Veysov, "Silero Models," 2022. Available: https://github.com/snakers4/silero -models

[17] Gonnella Giorgio, Kurtz Stefan, "Readjoiner: a fast and memory efficient string graph-based sequence assemble," BMC Bioinformatics, vol. 13, no. 1 , p. 82, May. 2012