Speech error recognition based on broad learning

Yicheng Yan¹², Liwen Wang¹², Wei Guo¹²

¹College of Information Engineering, Yancheng Teachers University, Yancheng, Jiangsu, 224001, China
²College of Computer Science and Technology, Nanjing Tech University, Nanjing, Jiangsu, 211800, China
*Corresponding author’s e-mail: 201961227006@njtech.edu.cn

Abstract. In recent years, with the increasing popularity of Chinese in the world, more foreigners have begun to learn Chinese. In this regard, this paper studies the phonetic errors in a single sentence, and designs a phonetic error classification model based on the broad learning network. First, the Laplace wavelet convolutional layer is designed as the first layer of the traditional CNN, and the Morlet-CNN model is used to extract the speech signal. Then, the extracted speech signal is input into broad learning network for training. Finally, the speech with different error types is tested. By comparing with the experimental results of a single speech amplitude feature and MFCC feature, the classification accuracy of the proposed model has been significantly improved, which verifies the effectiveness of the model.

1. Introduction
Not only do most foreign students rarely have the opportunity to practice oral Chinese, but they also face difficulties in oral Chinese, pronunciation errors, and psychological resistance. Therefore, foreign Chinese learners need a suitable system to help them correct errors in the process of learning Chinese.

Many scholars have conducted research on speech recognition. In the voice interactive language training system developed by the International Stanford Research Society[1] in 1996, the Hidden Markov Model (HMM) was selected to combine the duration, speed and accuracy of the speech segment to classify and score the speech signal. Researchers such as Jiang-CC use Viterbi decoding to separate syllables one by one, combine the probability obtained by HMM and the tone recognition of Gaussian Mixed Model (GMM), and use neural network structure to recognize speech signals[2]. Liu ZN, Luo YZ and others proposed to extract the pitch trajectory and Mel-Frequency Cepstral Coefficients (MFCCs) characteristics of the voice to be tested and the standard voice, and then use the Dynamic Time Warping (DTW) algorithm for matching comparison, the classification results are obtained from a subjective point of view according to the degree of similarity[3].

At present, most speech recognition algorithms are based on deep learning[4][5][6], and the deep structure of neural networks has achieved satisfactory classification accuracy. However, due to its complex structure and more hyperparameter settings, the time cost is high and the requirements for computer hardware equipment are high.

Professor C. L. Philip et al. proposed the Broad Learning System[7][8]. Broad learning system network is flat, including only the feature layer, enhancement layer and output layer. It has the advantages of simple network structure and few parameters. Broad learning system uses ridge regression to calculate the weight of the network, avoiding complicated iterative calculations, and achieving a higher accuracy rate in a shorter time.
In order to solve the problem of long operation time of traditional deep network, the text is based on the broad learning system, combined with Morlet wavelet convolutional neural network. While improving the network speed, the accuracy of speech error recognition is further improved.

2. Broad Learning System
Broad learning takes the mapping feature of the input data as the feature node of the network, and then enhances it to an enhanced node through the activation function. The mapping feature and enhancement node are directly connected to the output terminal. The network model of the broad learning system is shown in Figure 1.

\[
\begin{align*}
Y &= \phi(Xw_i + \beta_i), i = 1 \ldots n, \\
\phi((Z_1 \ldots Z_m)w_i + \beta_i), j = 1 \ldots m
\end{align*}
\]

Figure 1 Illustration of Broad Learning System.

Assuming that the input data is \( X \), use \( \phi(Xw_i + \beta_i) \) to project the input data to obtain the \( i \)-th mapping feature \( Z_i \),

\[
Z_i = \phi(Xw_i + \beta_i)
\]

Where \( w_i \) is the input weight and \( \beta_i \) is the deviation. Concatenate the first \( i \) groups of mapping features to obtain the mapping layer, which is denoted as \( Z = (Z_1 \ldots Z_i) \).

The mapping feature is non-linearly mapped by the activation function to obtain an enhanced node, and the \( j \)-th group of enhanced nodes is represented by \( H_j \).

\[
H_j = \xi(Z_1w_j + \beta_j)
\]

Among them, \( w_j \) represents the input weight of the \( j \)-th group of enhanced nodes generated randomly, and \( \beta_j \) is the deviation of the enhanced nodes. The connection of all enhanced nodes in the first group \( j \) is denoted as \( W = (H_1 \ldots H_j) \).

Connecting the mapping node and the enhancement node to the output layer at the same time, the obtained broad learning system can be expressed as the following equation.

\[
Y = [Z_1 \ldots Z_m \mid H_1 \ldots H_m] W^m = [Z^* \mid H^*] W^m
\]

where the \( W^m \) are the connecting weights for the broad structure. It is optimized through the following formula.

\[
\arg\min_{W} \| Y - AW - E \cdot A \|_1
\]

3. Wavelet Convolutional Broad Learning Network
Adding constraints to the CNN model and optimizing the shape of the filter through effective parameterization can reduce network parameters, accelerate the fitting speed of the model, increase the interpretability of the neural network, and even improve the accuracy of model classification[9]. Since speech signals are non-stationary signals, traditional feature extraction methods cannot effectively extract non-stationary signals. Wavelet analysis, as a commonly used tool for processing sound, image and other signals, has shown excellent capabilities. So we chose the wavelet convolutional layer to
replace the first traditional convolutional layer of the neural network[10]. The mathematical expression formula of Morlet wavelet is as follows:

$$\psi(p) = \pi^\frac{1}{2} e^{-\frac{p^2}{2}} \cos(2\pi \ast p)$$

(5)

Use the expansion factor $a$ and the translation factor $b$ to replace the variable $P$ of the Morlet wavelet, and the replacement formula is as follows:

$$P = \frac{(t-b)^2}{a^2}$$

(6)

Use Morlet wavelet to replace the first layer of the convolutional neural network. Compared with the traditional convolutional layer, the wavelet convolution has fewer parameters and the model converges faster. The speech signal processed by Morlet wavelet transform and convolutional neural network is input into the broad learning network, and the wavelet convolution broad learning network is proposed. The algorithm process of the wavelet convolution broad learning network is as follows.

Figure 2 Illustration of the wavelet convolution broad learning network.

According to the algorithm flow, it can be seen that the input voice data network is processed successively through two pooling layers, a traditional convolutional layer, and two fully connected layers. The output result of the fully connected layer is used as the input data of the broad learning network, and the rest of the process is the same as the training process of the broad learning network.

4. Experimental design and analysis

4.1. Experimental data set and comparative experimental design

In a laboratory environment, using a sampling rate of 11.025KHz, a poem was recorded as an experimental voice. The length of each recorded audio was 5 seconds, with 55,125 sampling points, for a total of 279 voices. Among them, 195 voices are used as the training set, and the remaining 84 voices are used as the test set. The data set has a total of 5 categories, which are all right, one random wrong character, two characters, three characters, and dialect phonetics.

4.2. Amplitude and MFCC feature extraction

In order to compensate for the problem of energy loss during sound propagation, the voice signal can be pre-emphasized. At the same time, the pre-emphasis technology can reduce the influence of low-frequency interference and glottal pulses, and improve the signal-to-noise ratio. Figure 3 is a time-domain waveform diagram of the voice signal after pre-emphasis. By comparing with Figure 4, it can be seen that the low-frequency part of the voice signal after pre-emphasis has been effectively suppressed, and the high-frequency component has been strengthened.
At this time, the amplitude feature extraction can be performed on the voice signal after pre-emphasis. The amplitude feature dimension of each voice signal is $[1,55125]$. After pre-emphasis, the speech signal is subjected to framing, windowing, energy spectrum calculation, Mel filtering and logarithmic calculation, and then MFCC feature extraction is performed. Set the extraction parameter to 20. The voice feature dimension after extraction is $[1,20]$. 

4.3. Experimental results and analysis

The experiment was carried out under the windows platform, and the experiment parameters were set as follows. The learning rate is 0.001, and the pre-emphasis coefficient is 0.98. In the broad learning network, the number of feature layer nodes is 100, and the number of enhancement layer nodes is 900. In the same environment, experiments were conducted on the control group method and the proposed method respectively, and the results obtained are shown in the following table 1 and the mixed matrix of the classification results is shown in the figure 5.

| Feature type | Method | Test accuracy |
|-------------|--------|---------------|
| Amplitude   | BLS    | 83.33         |
| MFCC        | BLS    | 84.52         |
| Raw data    | CNN+BLS | 86.91        |
| Raw data    | Ours   | 88.10         |

It can be seen from Table 1 that compared with other methods, our method has achieved a higher accuracy of speech error classification. At the same time, it can be seen from the confusion matrix that when the number of errors is too large, the recognition accuracy is higher. Although our method has a low accuracy rate for dialect speech classification, it has a better effect on most speech scores, which verifies the effectiveness of Morlet wavelet convolution to extract features.

5. Conclusion

Aiming at the problem of speech misclassification, it is proposed to apply Morlet wavelet convolution to the convolutional neural network and combine it with the width learning network. Experiments verify that the proposed method is effective. The designed method can provide feedback and correction of the wrong pronunciation of Chinese learners, solve the problem of slow improvement of learners' pronunciation, and provide a good learning environment for Chinese learners. At the same time, speech recognition is a competitive high-tech industry, and its development needs the support of related industries. As a basic application, speech error classification has indispensable practical significance. Since more experiments are conducted on standard Mandarin, we can conduct in-depth research on dialects in the follow-up.
References

[1] Neumeyer, L. (1996) Automatic text-independent pronunciation scoring of foreign language student speech [J]. Proc. of ICSLP-96.

[2] Yang, C., Yang, L., Jing, Y. (2010) Automatic pronunciation assessment for mandarin proficiency test based on hmm. Journal of Computers, 5(7), 1062-1069.

[3] Liu, Z., Luo, Y. (2005) Research on voice scoring method based on feature comparison [J]. Journal of Computer Applications, (12): 212-214.

[4] Jin, Q., Chen, S., Li, X., et al. (2015) Speech Emotion Recognition Based on Acoustic Features [J]. Computer Science, 42(9):24-28.

[5] You, L., Guo, W., Dai, L., Du, J. (2019). Deep neural network embeddings with gating mechanisms for text-independent speaker verification[J]. arXiv.

[6] Zhang, Y., Du, J., Wang, Z., Zhang, J., Tu, Y. (2018). Attention based fully convolutional network for speech emotion recognition. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

[7] Chen, C., Liu, Z. (2018). Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. IEEE Transactions on Neural Networks & Learning Systems, 29(99), 10-24.

[8] Chen, C., Liu, Z. (2017). Broad learning system: A new learning paradigm and system without going deep. 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE.

[9] Ravanelli, M. & Bengio, Y. (2018). Interpretable convolutional filters with sincnet[J]. arXiv preprint arXiv:1811.09725.

[10] Li, T., Zhao, Z., Sun, C., Cheng, L., Gao, R. X. (2019). Waveletkernelnet: an interpretable deep neural network for industrial intelligent diagnosis [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021: 1-11.