Intelligent Correction System of Students' English Pronunciation Errors Based on Speech Recognition Technology

MEILI DAI
School of Foreign Languages, Zhengzhou Sias University, Xinzheng 451150, CHINA

Abstract: With the increasingly frequent international exchanges, English has become a common language for communication between countries. Under this research background, in order to correct students' wrong English pronunciation, an intelligent correction system for students' English pronunciation errors based on speech recognition technology is designed. In order to provide a relatively stable hardware correction platform for voice data information, the sensor equipment is optimized and combined with the processor and intelligent correction circuit. On this basis, the MLP (Multilayer Perceptron) error correction function is defined, with the help of the known recognition confusion calculation results, the actual input speech error is processed by gain mismatch, and the software execution environment of the system is built. Combined with the related hardware structure, the intelligent correction system of students' English pronunciation error based on speech recognition technology is successfully applied, and the comparative experiment is designed the practical application value of the system is highlighted.

Keywords: Speech recognition; English pronunciation; Sensors; processor; Calibration circuit; MLP function; Identify confusion degree; Gain mismatch

Received: June 2, 2021. Revised: October 13, 2021. Accepted: October 29, 2021. Published: November 29, 2021.

1. Introduction

According to different recognition objects, speech recognition tasks can be divided into three categories, namely isolated word recognition, keyword recognition and continuous speech recognition. Among them, the task of isolated word recognition is to recognize isolated words known in advance; the task of continuous speech recognition is to recognize arbitrary continuous speech. Keyword detection in continuous speech stream aims at continuous speech, but it does not recognize all words, but only detects where some known keywords appear. According to the speaker, speech recognition technology can be divided into speaker specific speech recognition and speaker independent speech recognition. The former can only recognize one or several people's speech, while the latter can be used by anyone. Obviously, speaker independent speech recognition system is more in line with the actual needs, but it is much more difficult than the recognition for specific person. Speech recognition method is mainly pattern matching method [1, 2]. In the training phase, the user speaks each word in the vocabulary one by one, and stores its feature vector as a template in the template library. In the recognition stage, the feature vectors of the input speech are compared with each template in the template library in order of similarity, and the one with the highest similarity is taken as the output of the recognition result.

Voice is one of the most convenient, accurate and natural ways to communicate and interact between people. The speaker's brain produces language information, which is transmitted to the vocal organ to produce speech. The listener's ear, as a frequency analyzer, converts the sound waveform into an intermediate representation and sends it to the listener's brain to decode the desired language information. The process of communication can be regarded as a communication problem consisting of encoder and modulator at the speaker end and demodulator and decoder at the listener end. Without eyes and hands, they can easily communicate with each other. With the emergence of artificial intelligence machines, compared with the use of keyboard and mouse for

traditional human-computer interaction, people are naturally full of expectations for machines with highly intelligent voice communication ability [3, 4]. Intelligent machine which can understand people's speech, think and understand people's intention, and ultimately respond to people's voice or action has always been one of the ultimate goals of artificial intelligence. Intelligent voice interaction technology has naturally become one of the first research hotspots.

2. Hardware Design of Intelligent Correction System for English Pronunciation Errors

2.1 Optimization design of speech recognition sensor

Speech recognition sensor uses diode to complete speech recognition and convert English pronunciation into digital signal. At present, the most widely used speech recognition sensors are CMOS sensor and CCD sensor. CMOS Each diode in the sensor will be connected with a recognizer and conversion circuit, which will output the recognized English pronunciation in a way similar to the memory circuit. CCD sensor optimizes the internal structure of CMOS sensor, with only one recognizer, and the English pronunciation data of each diode will be transmitted to the next unit at one time, which will be output after the integration of the bottom part of the sensor, and the most accurate result will be obtained After that, it passes through the recognizer at the end of the sensor and outputs after successful recognition.

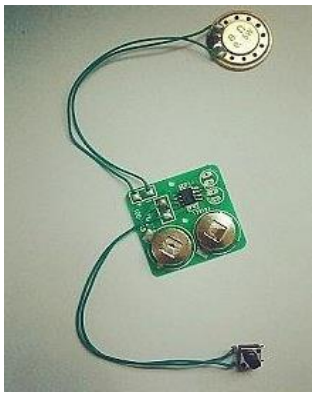

Figure 1 Structure diagram of speech recognition sensor

After acoustic feature extraction and acoustic model and language model training, speech recognition refers to the combination of acoustic model and language model. The training data includes speech data and its corresponding text data. The features extracted from the original speech and the corresponding annotation of sentences are used for supervised acoustic model training. The training text is used for language model modeling. After training acoustic model and language model, combined with decoder search algorithm, the recognition results of input speech features can be obtained. Good acoustic characteristics should consider the following three factors [5, 6]. Firstly, it should have excellent distinguishing characteristics, so that different modeling units of acoustic model can model conveniently and accurately. Secondly, feature extraction can also be considered as the process of speech information compression and coding. It not only needs to eliminate the channel and speaker factors and retain the content related information, but also needs to use the parameter dimension as low as possible without losing too much useful information, so as to train the model efficiently and accurately.

2.2 Improved design of speech recognition processor

According to speech recognition technology to complete the analysis and recognition of English pronunciation, it emphasizes the real-time of English speech processing. Therefore, in addition to improving the accuracy of speech processing and instruction processing of the common processor, it also improves the accuracy of speech recognition [7, 8]. The choice of DSP chip is also very important for the improvement of speech recognition processor. It is not only related to the processing speed of English pronunciation, but also related to the improvement difficulty and process of the processor.

The improvement of speech recognition processor needs to consider the processing speed and recognition accuracy of the chip. For an automatic English pronunciation system, the processing speed of speech recognition is the most important, and the processor must

be required to complete the corresponding processing tasks within a limited time, otherwise it is difficult to ensure the real-time performance of English pronunciation processing. When designing a speech recognition processor, the recognition speed of the processor chip is determined according to the DTW algorithm and the processing time requirements [9, 10]. Generally speaking, the accuracy of English pronunciation recognition of floating-point DSP chip is higher than that of fixed-point DSP chip, and it has high accuracy in automatic correction of English pronunciation errors. The internal structure of the processor chip is shown in Figure 2.

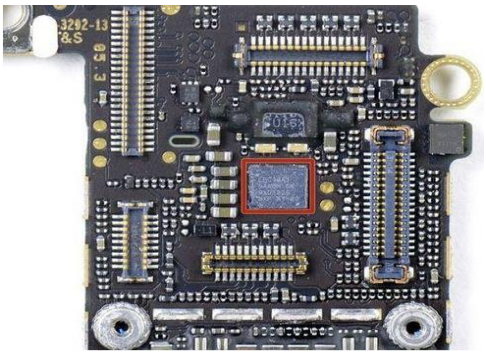

Figure 2 Internal structure of processor chip

The processor chip based on automatic correction of English pronunciation errors mainly aims at the characteristics of fast data processing speed and wide hardware resources, and chooses the processor chip to meet the requirements of English pronunciation data processing. A large number of operations in the automatic correction system of English pronunciation errors are the data processing of English pronunciation errors. Therefore, considering the processing speed factor, we decided to choose TDSP - TF887 as the choice of processor chip. Speech recognition processor architecture is a hybrid 16 / 32 bit instruction level architecture based on optimal code density. This architecture is suitable for processing complex English pronunciation error data, and can improve the accuracy of automatic correction of English pronunciation error.

2.3 Intelligent correction circuit

Intelligent correction circuit can provide current supply guarantee for the realization of speech recognition technology. When the sensor receives the sound from the students, the relevant circuit elements will spontaneously generate continuous sensing current, which can be transmitted to other system application elements under the action of related resistance elements. Generally speaking, the speech recognition processor can be directly connected with the intelligent correction circuit. As a sound wave collection structure, the former can collect and process the received speech signals, and then spread these speech signals to the lower processing host with the help of transmission wires [11, 12]. Due to the existence of the collector elements, the intelligent correction circuit can directly interfere with the system host's perception of students' English pronunciation errors. In general, the larger the transmission current is, the stronger the intelligent correction circuit's perception of students' English pronunciation errors is. Under the action of R, s and other resistance elements, the voice signal can be directly fed back from the input structure to the output structure, and in this process, the fluctuation degree of the voice will not change. This is also the main reason why the new system can better correct students' English pronunciation errors.

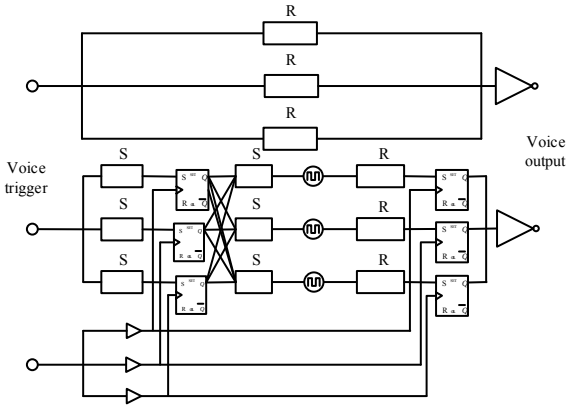

Figure 3 Schematic diagram of intelligent correction circuit

The ultimate goal of English pronunciation correction is that English speech recognizer or machine can effectively deal with the impact of different speech styles, channel differences, environmental changes and so on. Because speech signal is affected by these variables, it is not an easy task to recognize it. We naturally want to use a powerful model to model speech signal.

3. Software Design of Intelligent Correction System for English Pronunciation Errors

3.1 MLP error correction function

The purpose of speech recognition technology is to learn enough knowledge from the known English pronunciation data, and then extend it to the new data in the future to make effective decisions. Do well in data that doesn't appear in training. But with the progress of the learning process, the learning curve in the training set is always getting better. When do we stop the learning process and get good generalization? An effective method to improve the generalization is the early stop strategy combined with cross validation. Generally, the training data is divided into a separate data set, which does not participate in the network training and is called verification set. The learning process of the network alternates between the training stage and the verification stage [13, 14]. Different from the fact that the learning curve on the training set always gets better, the error function on the verification set generally decreases first and then increases, so when the error function on the verification set is the smallest, the iteration can stop.

One of the most successful areas of MLP is in English pronunciation error correction and recognition, especially in the acoustic modeling of students' speech. In these works, MLP is usually used as a pattern classifier. The input training features of MLP are usually standard acoustic features, such as MFCC or PLP. The output layer node of MLP corresponds to the Subword unit of speech, such as phoneme, syllable, etc. the nonlinear activation function of output layer is expressed in the form of formula (1). In this way, the trained MLP will estimate the posterior probabilities of each phoneme class in the output layer under the condition of the current input data. These posterior probabilities can be used by the following automatic speech recognition systems [15, 16]. In the experiment, it is found that it is very helpful to select about 90 ms feature expansion in the input layer of MLP. Assuming that e represents the minimum recognition coefficient of English pronunciation error and k represents the maximum recognition coefficient of English pronunciation error, the MLP error correction function can be defined as:

$$B = \sum_{e=1}^{k} w_e \frac{1}{\sqrt{(2\mu)^2 |\overline{y}|}} \tag{1}$$

In the formula, w_e —the minimum amount of English pronunciation input, μ — the correction coefficient term, and $\overline{y}$ — the recognition average amount of the input speech.

3.2 Calculation of recognition confusion

In the intelligent correction system of students' English pronunciation errors, the lower the recognition confusion is, the higher the certainty of using the language model to predict the possible words based on the given historical word sequence is. Therefore, the language model is trained to minimize the confusion of sentences in the training set. In the process of training, first of all, we need to count the frequency of each word and the combination of words in the training corpus, and then calculate the parameters of the language model. Because the number of word combinations increases exponentially with the size of the word list, in the situation of large word list and insufficient training data, we need to use speech recognition technology for some phrases with very small probability of word combination, even for phrases that never appear in the training set.

In the aspect of the characteristic domain of the students' English pronunciation error data, an improved method is proposed based on the sensor element. Firstly, the speech segments which compete with the correct mark are selected, and then the competitive neural network is trained by using these competitive information. In the aspect of model correction, a new method for modeling multi flow acoustic characteristics is proposed based on the hybrid modeling framework of speech recognition technology and MLP function [17-19]. The multi stream original features are automatically learned by the unsupervised features layer by layer, and the high-level feature representation is fused to continue the following feature learning. The whole process of

information fusion is combined with the training of deep neural network. This method of feature fusion in the middle layer of deep neural network has achieved more performance than the traditional multi flow information modeling method. Assuming that $r_{\max}$ represents the largest characteristic value of students' English pronunciation errors, the simultaneous formula (1) can express the recognition confusion degree of the correction system as follows:

$$P = \frac{\left[B(r_{\max})^{-\frac{1}{\kappa}} \right]}{\lambda \times f^2} \tag{1}$$

In the formula, κ represents the power correction coefficient, λ represents the behavior processing period of speech correction, and f represents the correction partial derivative coefficient based on speech recognition technology.

3.3 Gain mismatch processing of speech errors

Different from offset mismatch, the spurious component caused by gain mismatch is related to the input speech signal. Because the noise introduced by the offset mismatch is direct flow, it is simply superimposed with the input signal. But the noise caused by gain mismatch multiplies the input signal, which is similar to amplitude modulation, so the noise energy caused by gain mismatch is proportional to the amplitude of input signal. Speech error correction technology is usually divided into foreground correction and background correction. The characteristic of foreground correction is that it needs to use various reference signals as the input of time interleaved ADC system. The mismatch parameters are extracted by sampling and converting the input signal of actual ADC, and the feedback adjustment is carried out. For the foreground correction technology, the biggest problem is that the correction process interrupts the normal operation of ADC, because the foreground correction needs to be carried out after the power on reset or during the normal conversion interval of ADC [20, 21]. If the calibration is performed only when the power is on and reset, the acoustic wave, volume and other environmental factors may change with time in the process of ADC operation, and the mismatch and error characteristics of the system will also change. Because the system does not make further real-time calibration, it can not adapt to the changing mismatch parameters, and the performance of the ADC system will fluctuate accordingly. If real-time correction is needed, the ADC will be interrupted if correction operation is needed in the normal conversion process. Obviously, this requirement greatly limits the application scope of ADC system and is difficult to be accepted by system users.

Supposing that β represents the foreground speech recognition coefficient of the intelligent correction system and ϖ represents the background speech recognition coefficient of the intelligent correction system, and the simultaneous formula (2) can express the gain mismatch processing result of the speech error as follows:

$$N = \sqrt{(\frac{P}{k-1})\frac{3}{\beta}\frac{1}{2^{\varpi}}} \tag{2}$$

(3)

In the formula, k represents the gain adjustment coefficient of English pronunciation error.

4. System Practicability Test

Two hundred freshmen in a university were randomly divided into two groups with 100 experimental researchers in each group, one group as the experimental group and the other as the control group. Two experimental hosts with the same configuration were used as the original control equipment. The experimental host was equipped with an intelligent correction system for students' English pronunciation errors based on speech recognition technology, and the control host was equipped with a neural network correction system [22].

The γ-wave length value can reflect the accuracy of the system host for correcting students' wrong English pronunciation. Generally, the larger the γ-wave length value is, the stronger the accuracy of the system host for correcting students' wrong English pronunciation is, and vice versa. The specific changes of γ-wave length in experimental group and control group are shown in Table 1.

Table 1 Comparison of γ-wave length values

Number of students /(person)	γ -wave length value /(mm)	
	Test group	Control group
10	0.8	0.3
20	0.8	0.2
30	0.8	0.3
40	0.8	0.2
50	0.8	0.3
60	0.9	0.3
70	0.9	0.3
80	0.9	0.2
90	1.0	0.3
100	1.0	0.2

According to the analysis of Table 1, with the increase of the number of students participating in the experiment, the γ-wave length of the experimental group always keeps a slow rising trend. When the number of students reaches 60, the rising trend has a small increase. The length of γ- wave in the control group kept a stable trend. From the perspective of limit value, the maximum value of 1.0 mm in the experimental group increased by 0.7 mm compared with 0.3 mm in the control group.

The ψ wave length value can reflect the correction stability of the system for the wrong speech signal. Without considering other interference conditions, the larger the ψ wave length value is, the more stable the correction behavior of the system for the wrong speech signal is. The specific changes of ψ wave length in the experimental group and the control group were recorded in Table 2.

Table 2 Comparison of ψ-wave length values

Number of students /(person)	Ψ wave length value /(mm)	
	Test group	Control group
10	0.7	0.5
20	0.8	0.6
30	0.9	0.6
40	1.0	0.5
50	1.1	0.4
60	1.2	0.4
70	1.3	0.3
80	1.3	0.3
90	1.3	0.2
100	1.3	0.2

By analyzing the recorded values in Table 2, we can see that the value of ψ wave length in the experimental group gradually tends to be relatively stable after a period of rising, and the maximum value is 1.3 mm, which can maintain a stable state for a long time. In the control group, the value of ψ wave length began to decrease steadily after a slight increase, and the maximum recorded value was only 0.6 mm, which decreased by 0.7 mm compared with the maximum value in the experimental group

According to the above experimental results, the application of intelligent correction system for students' English pronunciation errors based on speech recognition technology can improve the numerical results of γ detection wavelength and ψ detection wavelength at the same time, and can achieve accurate and stable correction for students' English pronunciation errors, which meets the practical application requirements.

5. Conclusion

Under the function of the optimized sensor and processor, the new intelligent correction system redefines the MLP error correction function. With the help of the intelligent correction circuit, it calculates the recognition confusion value and performs the gain mismatch processing on the speech error data. The experimental results show that γ detection wavelength and ψ detection wavelength have a trend of synchronous increase, which not only completes the accurate correction of students' English pronunciation errors, but also better maintains the application stability of the experimental environment, and has strong feasibility ability.

References

[1] Kumar A, Aggarwal R K. Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling. Journal of Intelligent Systems, 2020, 30(1): 165-179.

[2] Alotaibi Y A, Selouani S A, Yakoub M S, et al. A

canonicalization of distinctive phonetic features to improve arabic speech recognition. Acta Acustica united with Acustica, 2019, 105(6): 1269-1277.

[3] Tiwari V, Hashmi M F, Keskar A, et al. Speaker identification using multi-modal i-vector approach for varying length speech in voice interactive systems. Cognitive Systems Research, 2019, 57: 66-77.

[4] Bobkov S A, Kurushin D S, Perevalov A M, et al. Using linguistic anticipation to improve the quality of speech recognition in robotic systems. Russian Electrical Engineering, 2020, 91(11): 669-672.

[5] Deuerlein C, Langer M, Sener J, et al. Human-robot-interaction using cloud-based speech recognition systems. Procedia CIRP, 2021, 97(2): 130-135.

[6] Veisi H, Mani A H. Persian speech recognition using deep learning. International Journal of Speech Technology, 2020, 23(4): 893-905.

[7] Song Z. English speech recognition based on deep learning with multiple features. Computing, 2020, 102(99): 1-20.

[8] Revathi A, Sasikaladevi N. Hearing impaired speech recognition: Stockwell features and models. International Journal of Speech Technology, 2019, 22(4): 979-991.

[9] Yazdani R, Arnau J M, Gonzalez A. A Low-Power, High-Performance speech recognition accelerator. IEEE Transactions on Computers, 2019, 68(12): 1817-1831.

[10] Viswanathan N, Kokkinakis K. Listening benefits in speech-in-speech recognition are altered under reverberant conditions. The Journal of the Acoustical Society of America, 2019, 145(5): EL348-EL353.

[11] Toledo T D, Lee H D, Spolaor N, et al. Web System Prototype based on speech recognition to construct medical reports in Brazilian Portuguese. International Journal of Medical Informatics, 2019, 121: 39-52.

[12] Kim G, Lee H, Kim B K, et al. Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition. IEEE Signal Processing Letters, 2019, 26(1): 159-163.

[13] Ri H C. A usage of the syllable unit based on morphological statistics in Korean large vocabulary continuous speech recognition system. International Journal of Speech Technology, 2019, 22(4): 971-977.

[14] Hu S, Shang X, Qin Z, et al. Adversarial examples for automatic speech recognition: attacks and countermeasures. IEEE Communications Magazine, 2019, 57(99): 120-126.

[15] Cui X, Zhang W, Finkler U, et al. Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies. IEEE Signal Processing Magazine, 2020, 37(3): 39-49.

[16] Rumagit R Y, Alexander G, Saputra I F. Model comparison in speech emotion recognition for indonesian language. Procedia Computer Science, 2021, 179(1): 789-797.

[17] Hammami N, Lawal I A, Bedda M, et al. Recognition of Arabic speech sound error in children. International Journal of Speech Technology, 2020, 23(3): 1-7.

[18] Jermsittiparsert K, Abdurrahman A, Siriattakul P, et al. Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 2020, 23(4): 1-8.

[19] Hai Y. Computer-aided teaching mode of oral English intelligent learning based on speech recognition and network assistance. Journal of Intelligent and Fuzzy Systems, 2020, 39(4), 5749-5760.

[20] Poorna S S, Nair G J. Multistage classification scheme to enhance speech emotion recognition. International Journal of Speech Technology, 2019, 22(2): 327-340.

[21] Krobba A, Debyeche M, Selouani S A. Maximum entropy PLDA for robust speaker recognition under speech coding distortion. International Journal of Speech Technology, 2019, 22(4): 1115-1122.

[22] Zhao L., Liu Y., Chen L., et al. English oral evaluation algorithm based on fuzzy measure and speech recognition. Journal of Intelligent and Fuzzy Systems, 2019, 37(1), 241-248.