ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

Design and Implementation of the 'Dialect' Learning System based on Speech Recognition

Jiale Li, Mingming Hu

Ningbo ladder Education Technology Co., Ltd, 315000, China

Abstract

Speech recognition technology can make pronunciation learning software with pronunciation feedback function to help learners correct the wrong pronunciation in time, so that learners can avoid forming wrong pronunciation habits due to multiple repetitions. This paper introduces a dialect pronunciation training system with functions of pronunciation following, pronunciation evaluation and pronunciation correction developed by using artificial intelligence and speech recognition technology. It aims to realize the intelligence, popularization and portability of pronunciation learning of local dialects and improve the pronunciation level of dialect learners to a certain extent.

Keywords

Dialect learning; Speech recognition; Artificial intelligence.

1. Introduction

As Chinese people travel around the country, dialect learning is conducive to strengthening human interaction in different places. But dialect pronunciation has become a difficult problem in dialect learning for everyone, and the reasons for that are mainly in the following 3 aspects.

(1) There are great differences in pronunciation methods between dialects. If you grow up in a non-native environment, you will make many pronunciation mistakes that are hard to notice when you learn the dialect. And if you don't correct them in time at the beginning of learning, you will often acquire a very unstandard dialect.

(2) There is a lack of qualified dialect teachers. Even in primary and secondary schools in large and medium-sized cities, there is a shortage of dialect teachers who are able to provide accurate instruction in spoken pronunciation. In general, multimedia teaching can only be taught unilaterally, and teachers cannot teach interactive dialects to meet the specific needs of students effectively.

(3) There is a lack of time and environment to practice spoken dialects. Language is a way of communication, and the most important thing is to pronounce more and practice more. However, in traditional dialect learning, people often spend a lot of time on reading and writing dialects, and they do not have enough time and practice opportunities in spoken language pronunciation. Most of the dialect learning software in the current market focus on the improvement of dialect reading and writing skills. Some of the only spoken pronunciation learning software has a single function, and can only perform simple operations such as pronunciation following. They lack effective feedback on learners' pronunciation and the training effect is not satisfactory.

2. Project Implementation Content

The main function of the dialect learning system based on speech recognition technology is to learn and train the pronunciation of dialects in the form of animation, sound, pictures and text. It can achieve effective feedback on learners' pronunciation, guide and improve learners'

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

continuous training and improve their dialect pronunciation level. It also provides a friendly, intuitive and brief operating interface. Based on the requirement analysis, the functions of the system are determined as follows.

(1) Pronunciation demonstration. Pronunciation demonstration means that when learning pronunciation, the standard pronunciation animation video or the standard pronunciation sound is played first, together with the pronunciation structure diagram and the introduction text, etc. It makes the learners have a correct understanding of the pronunciation, the main points of pronunciation, the mouth and the tongue movement characteristics, etc.

(2) Pronunciation following. The system first plays the correct pronunciation animation or pronunciation sound, and then prompts the learner to follow the pronunciation. The learner follows the prompt to pronounce. And at the same time, the system records the learner's pronunciation to the phone memory for subsequent processing.

(3) Pronunciation comparison. The system firstly plays the standard pronunciation oral animation video or sound, and then plays the recorded learner's pronunciation. The pronunciation comparison function is mainly to compare the standard reference pronunciation with the learner's pronunciation, so that the user can have a direct understanding of the difference between their pronunciations and the standard pronunciation.

(4) Pronunciation scoring. Pronunciation scoring is one of the core functions of the system, which mainly uses speech recognition technology and related pronunciation scoring algorithm to have a quantitative evaluation of learners' pronunciation results. Accurate and reliable pronunciation scores enable learners to have an accurate understanding of their own pronunciation performance, which in turn allows them to continuously improve their pronunciation and enhance their pronunciation level.

(5) Image display of pronunciation results. The image display of pronunciation results is mainly in the form of image feedback to compare the learner's pronunciation with the standard pronunciation. The system uses pronunciation resonance peak comparison diagram to reflect the difference between standard pronunciation and learner pronunciation resonance peaks. According to the relationship between resonance peaks and pronunciation mouth and tongue position, the reference diagram also reflects the pronunciation mouth and tongue position movement characteristics of learner pronunciation and standard pronunciation to some extent.

By analyzing the functional requirements of the system, it is finally determined that the core of the system should include the following modules: voice recording module, voice and video playback module, AP-based pronunciation scoring module, and image display module of pronunciation resonance peak.

2.1. Scoring Module Design

The system scoring module adopts the AP-based pronunciation scoring technology, which includes the scoring parameter generation part and the pronunciation scoring part. It is responsible for the adaptive generation of scoring parameters and the correct scoring of learners' pronunciation, and the workflow diagram of both is shown in the figure.

First of all, the tested pronunciation and the standard reference pronunciation are pre-processed separately. The preprocessing includes pre-emphasis of pronunciation, framing and windowing, endpoint detection and other processes. After the tested pronunciation and standard pronunciation are preprocessed, then the feature extraction and pattern matching are calculated. The MFCC characteristic parameters and the DTW dynamic time regularization method are applied to the system. After the above processing, the frame-average matching distance between the tested pronunciation and the standard reference pronunciation can be obtained.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

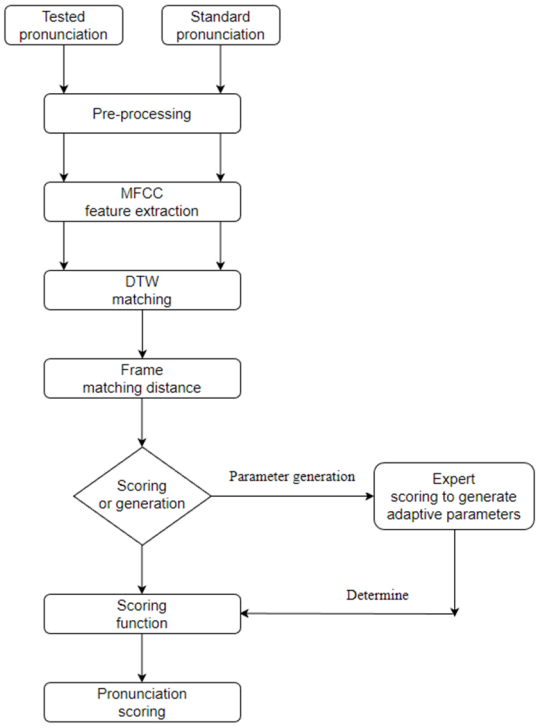

Figure 1. System pronunciation scoring flow

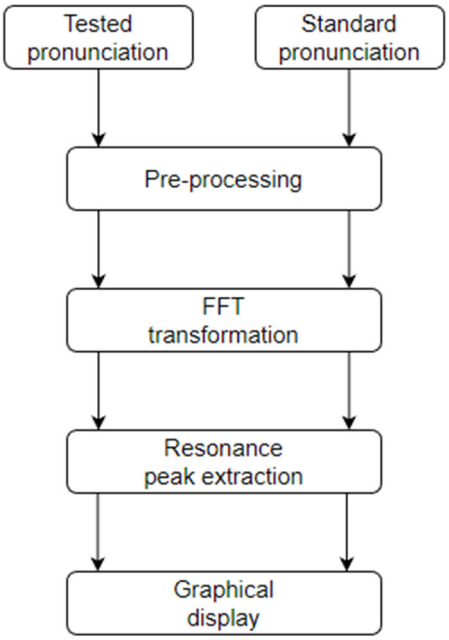

Figure 2. Resonance peak comparison flow chart

2.2. Feedback Module Design

The resonance peak image display module is mainly responsible for graphically depicting the changes of resonance peaks between the standard reference pronunciation and the learner's pronunciation over time, thus qualitatively reflecting the differences in pronunciation patterns between the two. After the process of preprocessing, FFT transformation, and resonance peak extraction, the system obtains the resonance peak information of learner pronunciation and standard reference pronunciation. In order to display this resonance peak information

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

graphically on the mobile terminal, the system utilizes AchartEngine, an open source chart generation library developed for Android applications, which supports line, bar, and pie charts, etc. Using this library, the system is able to display comparative charts of pronunciation resonance peaks.

3. Key Technologies

3.1. Speech Recognition Method

Based on the principle of pattern matching technology, the project first deposits the feature vectors of known speech signals as templates in the template library. After feature extraction, the feature vectors of the input speech to be measured are compared with the reference templates in the template library for similarity, and the recognition results are finally derived. The main processes of speech recognition include pre-processing, feature extraction, and pattern matching, etc. The figure is the block diagram of automatic speech recognition system based on pattern matching principle.

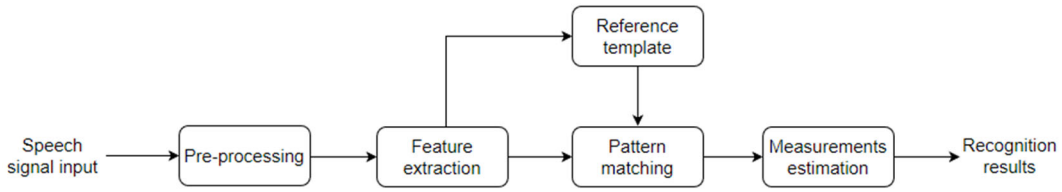

Figure 3. Flowchart of recognition based on pattern matching

3.2. Speech Informative Feature Extraction

After pre-processing the speech signal, feature extraction of the signal is also required. The feature extraction part is responsible for calculating and extracting the key parameters reflecting the signal characteristics, and effectively describing the speech signal by a small number of parameters for subsequent processing. The feature extraction of the signal not only highlights the data features of pattern matching and improves the recognition rate of the system, but also compresses the information and reduces the storage and operation of the system.

3.3. Scoring Mechanism based on Adaptive Parameters

In the single-reference template-based scoring method, the scoring parameters a and b need to be determined in conjunction with the empirical expert scoring results when performing scoring operations. The existing pronunciation scoring systems are constantly debugged and modified for a particular computer or hardware platform through methods such as system simulation and testing to determine the values of a and b. Once the system is completed, the values of a and b cannot be changed. This approach has the limitation that the scoring parameters determined are often only appropriate for the platform system being tested. Once the platform or speech capture peripherals used are changed, the scoring parameters will no longer be accurate and the scoring results will no longer be reliable. Considering the relatively large hardware differences of Android phones, this method is very unfavorable to the application and popularity of the system.

In order to solve the limitation of fixed scoring parameters in the above scoring methods, this paper proposes an adaptive parameter (AP) based scoring method. The aim is to achieve platform adaption of the system, enhance the compatibility of the system, and improve the reliability and accuracy of the scoring mechanism. The AP-based scoring method is an improvement to the single-reference template-based scoring method, and the AP-based scoring algorithm is defined here as follows.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

$$score = \frac{100}{1 + x(d)^y}$$

Where x and y are adaptive parameters. The parameters x and y for scoring operations are not deterministic, but can vary adaptively according to the computer or hardware device. The adaptive parameters x and y are generated by the adaptive training of the system, and the specific computational flow is shown in the figure.

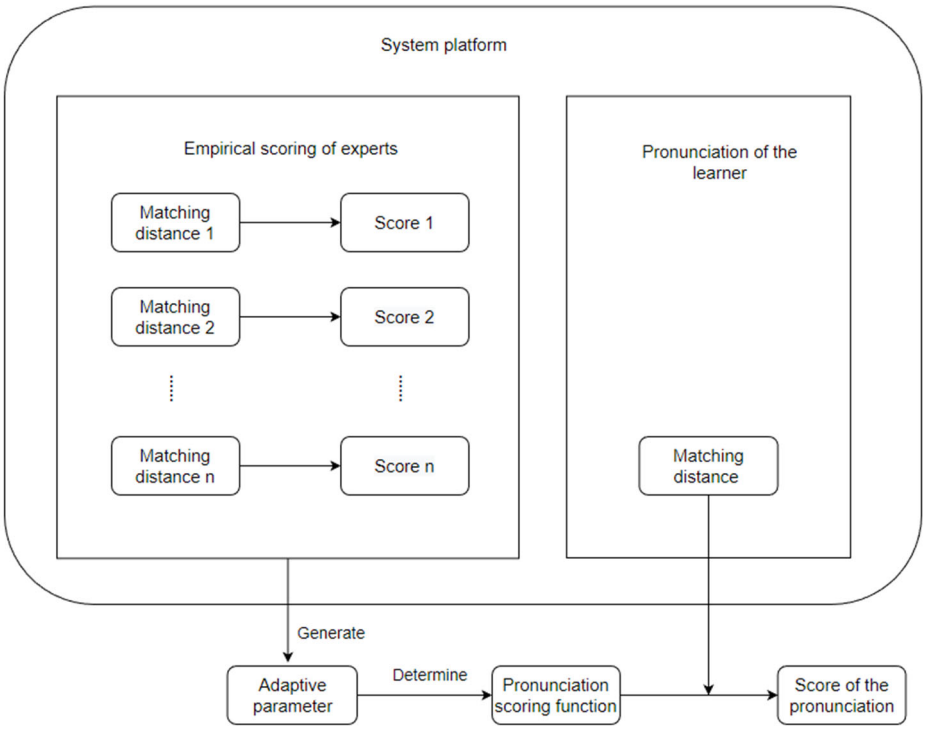

Figure 4. Schematic diagram of AP-based scoring calculation

Before the system performs pronunciation scoring, there is a separate scoring parameter generation module to generate adaptive parameters. In the scoring parameter generation module, the learner pronounces the pronunciation for several speech sounds, while the expert empirically scores the learner's pronunciation. Thus, for each pronunciation, the MFCC frame matching distances and the corresponding expert scores will correspond to each other. Let the set of MFCC frame average matching distances for all training pronunciations be A={d1,d2,... .di,... .dn}, and the corresponding set of expert scores is B={s1,s2,... si,... .sn}. As a result, n pairs of frame matching distances and expert scores are obtained, and the relationships are as follows.

$$\begin{cases} s_1 = \frac{100}{1 + x(d_1)^y} \\ s_2 = \frac{100}{1 + x(d_2)^y} \\ \cdots\cdots \\ s_n = \frac{100}{1 + x(d_n)^y} \end{cases}$$

In order to find the parameters x and y, the best values of x and y can be obtained by using the least squares curve fitting method. And theoretically, the larger the sample space, the more accurate the obtained fitting function is. However, since the scoring parameter generation module is implemented on the Android cell phone platform, the system requires a high level of real-time computation and a relatively low level of accuracy for the parameters. In order to

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202210_5(10).0123

simplify the process of scoring parameter generation as much as possible and reduce the computational effort, the system selects 5 speech samples for training, and selects the most suitable 3 samples from them for calculation. As a result, the estimated values of parameters x and y can be calculated quickly for scoring operation.

Since the scoring parameter generation module and pronunciation scoring module are located on the same mobile device, the operational parameters of pronunciation scoring are generated based on expert scoring training before performing pronunciation learning. The generated scoring parameters reflect the characteristics of the current system hardware platform, and the scoring scores have a high similarity with the empirical scoring of experts. Therefore, the AP-based method has strong adaptiveness, high accuracy and reliability, while greatly improving the compatibility of the system.

4. Conclusion

This system realizes a set of dialect pronunciation training system with multi-functional functions such as pronunciation following, pronunciation evaluation and pronunciation correction based on mobile terminal and by using relevant artificial intelligence and speech recognition technology. It aims to realize the intelligence, popularization and portability of pronunciation learning of local dialects. The system has been tested to have a high accuracy in pronunciation scoring and the efficiency of pronunciation correction reaches 80%, which can improve the pronunciation level of dialect learners to a certain extent.

References

[1] Huang Wei, Shi Jiaying. Research on speech recognition based on deep neural network[J]. Modern Computer,2016,(7).20-25.

[2] Xing Anhao, Zhang Pengyuan, Pan Jilin, et al. DNN cropping method and retraining based on SVD[J]. Journal of Tsinghua University (Natural Science Edition), 2016, (7). 772-776. doi:10.16511/j.cnki.qhdxxb.2016.21.043.

[3] Mo Yuanyuan, Guo Jianyi, Yu Zhengtao, et al. A Chinese - Vietnamese bilingual word alignment method based on deep neural network (DNN)[J]. Journal of Shandong University (Science Edition), 2016,(1).77-83. doi:10.6040/j.issn.1671-9352.3.2014.289.

[4] Zhang Chi. Application of deep neural networks in environment detecting systems for mobile terminal [D]. University of Electronic Science and Technology of China, 2017.1-102.

[5] Wang Zhenyu. Research on speech recognition technology under embedded platform [D]. Guizhou University,2017.1-76.

[6] Gong Yanting. Audio recognition based on acoustic spectrogram saliency detection[D]. Hefei University of Technology,2015.1-54.

[7] Guo Shengqiang. Research on cross-domain speech recognition based on deep learning [D]. Chongqing University of Posts and Telecommunications,2017.

[8] Zhao Tiankun. Music information retrieval based on deep neural network[D]. Beijing University of Posts and Telecommunications,2015.1-70.