Construction and Analysis of English Learning Model Based on Classroom Network Environment

Qi Ding

Rizhao Polytechinic, Rizhao, Shandong, 276826, China

Keywords: Network environment, English learning, oral English.

Abstract: In classroom teaching, learners' acceptance of English knowledge is mainly achieved through listening and speaking skills, so, in terms of listening comprehension and spoken English, this paper studied the algorithm of intelligent English learning model. This paper introduced the process of speech recognition, such as preprocessing, feature extraction, pattern matching and so on, the algorithm of pattern matching using MFCC feature extraction and DTW dynamic time warping has been improved and selected, and the specific parameters have been given. Finally, the test was carried out. After testing, the pronunciation accuracy of the algorithm was higher, for vowel and word pronunciation, the similarity with expert scores was above 90%; the effective rate of pronunciation correction was 80%, which could improve the pronunciation level of learners to a certain extent.

1. Introduction

At present, with the continuous development of science and technology, English classroom teaching is gradually to the direction of information and intelligence. In classroom teaching, learners' acceptance of English knowledge is mainly achieved through listening and speaking ability, however, in recent years, with more and more Chinese learning English, the corresponding English learning software has become more and more, However, most softwares do not have good pronunciation evaluation and feedback correction for oral English pronunciation (Choi J, Yi Y. 2016)[1]. However, in English pronunciation learning, especially for not English speaking learners, effective feedback motivation is very important. This has become a bottleneck restricting the intelligent English learning software (Paramudia P, Habil H. 2015)[2].

With the maturity of speech recognition technology, people begin to use pronunciation recognition technology to assist pronunciation learning at home and abroad, but most of them stay in the theoretical stage(Sun Y, Franklin T,et al)[3]. With the rapid development of mobile Internet technology, the popularity of smart phones is becoming higher and higher. As an excellent smartphone operating system, Android system has developed rapidly in recent years, and its share in smart phones has reached more than 80%(Pingxiao W. 2016)[4]. Android smart phone not only has powerful data processing ability and gorgeous graphical user interface, but also because of the system is completely open, rich development components, it can be quick and convenient for application development (Huang L H, Chen C Y.2016)[5]. Compared with the traditional computer software, the software based on the Android smartphone platform has great convenience and practicability (Hegedus S J, Dalton S, et al.2015)[6]. Therefore, it has a great practical and social significance to develop a portable and popular English pronunciation aided learning system with pronunciation feedback function based on Android platform(Hashmi N A.2016)[7].

2. State of the art

With the deepening of the research, the expansion of the application field and the relaxation of the constraints of speech recognition, speech recognition technology has encountered some difficult problems to solve(Rui Z, Yang H Y. 2017)[8]. For example, filtering and building template data become difficult as the vocabulary library expands; the acoustic characteristics of different people on

  DOI: 10.25236/icepms.2018.221

the same pronunciation are quite different, which makes the identification of non - specific people unreliable(Sampson D G, Sergis S, et al.2015)[9]. At the end of the 1980s, there was a breakthrough in the study of speech recognition: Researchers have succeeded in solving 3 major problems of large vocabulary, continuous speech and non-specific people in the laboratory, and a recognition system with corresponding characteristics is also implemented. The successful representative of the research results is the Sphinx system of Carnegie Mellon University(Carnegie Mellon University), this is the first reliable recognition system for large vocabulary, non-specific and continuous speech recognition(Webster G D, Gesselman A N, et al.2016)[10]. From the beginning of this period, the research of speech recognition is more thorough. Speech recognition is also changed from initial recognition based on specific person, isolated word to non-specific person, continuous sentence recognition, and the recognition method is gradually transformed from template matching method to statistical model method. The most important research results of this period are the application of hidden Markov model (HMM) and artificial neural network (ANNs) in speech recognition. Related research uses statistical methods to transform the field of vision from micro to macro, and to build the best speech recognition system from the perspective of the whole rather than the details, so as to achieve a better recognition effect.

3. Methodology

3.1 Preprocessing of Speech Recognition Signal

In the field of signal processing, speech recognition technology is one of the key research contents. In essence, speech recognition technology belongs to a pattern recognition process, the purpose is to make the machine can effectively identify the human voice. The speech signal is generated by the sound path excited by the airflow. The voiceless, voiced plosives and the three types of speech, the difference in spending is the pronunciation incentives. The sound source is excited by the air turbulence in the contraction area of the sound channel, and the voiced sound originates from the quasi periodic pulse signal of the glottal position. The burst sound source is generated from the sudden release of the accumulated air pressure at the closed point of the sound channel. Because the speech signal changes with the movement of the vocal organs, the vibration of the vocal cords is much faster than the movement of the vocal organs. In relatively short time frames (10 to 30 milliseconds), some characteristic parameters, such as speech spectrum, can’t be changed, and the signal is treated as a short-term stationary signal. After that, the processing of speech signal is based on this theory.

Preprocessing is the first part of speech signal processing, and it is also a very important link. Before the feature extraction of signal, preprocessing must be carried out. Generally speaking, speech signal preprocessing covers signal digitization, pre emphasis, endpoint detection and so on.

In the digital part of the speech signal, we use the Android mobile phone comes with a headset as an input device of the speech signal. The digital processing of the language signal is realized by the combination of software and hardware. Android mobile phone comes with audio processing chip, which can sample and quantify the speech signal, but the related sampling quantization parameters need to be configured by software programming. There are many ways to control sound acquisition in Android SDK, in which the Audio Record class can control the specific way of the underlying data acquisition, which is very suitable for the system. In this way, the sampling rate and quantization bits of the signal can be flexibly configured by using the related methods of the Audio Record class, so that the audio chip can be controlled digitally according to the specified parameters. According to the Nyquist theorem, the sampling rate is greater than or equal to 2 times of bandwidth during the sampling operation of the unfolded signal. The frequency of the general speech signal is 300~3400Hz. This paper adopts the sampling rate of 8000Hz, which is enough for general speech signal, moreover, the lower sampling rate can reduce the workload of FFT transform and save computing time, which is more important for real-time speech recognition system. Quantization bits are the data range that each sampling point can represent. The commonly used 8 bit, 16 bit, 24 bit and other.8 bit word length

quantization quality is lower, the 16 bit quantization quality is higher, and convenient for computer processing. In this paper, 16 bit quantization bits are selected:

$$H(z) = 1 - a * z^{-1} \tag{1}$$

As far as the speech signal pre emphasis is concerned, the speech signal will attenuate in the high frequency region under the excitation of glottis and the radiation of the nose and mouth. In order to make the signal frequency become more platform, and to provide more convenience for the analysis of the related characteristic parameters such as spectrum, it is necessary to strengthen the signal high frequency region, that is, pre emphasis. The energy loss usually follows the following laws: When the signal frequency is increased by 2 times, the power spectrum amplitude will be reduced by about 6Db. Therefore, the signal can be weighted (or upgraded) according to the proportion of 6D B/oct. The speech signal is usually expressed by the first order high pass filter in the form of time domain, and the pre emphasis signal S2 (n) is used as the signal:

$$s_2(n) = s(n) - a * s(n-1) \tag{2}$$

A is the pre weighting coefficient, and the general value is close to 1. The value is 0.97.

In the frame windowing processing part of speech signal, the principle of short-time analysis is to intercept the speech signal into small intervals, each segment is called a frame. The system uses 8000Hz sampling rate, that is, 8000 sampling points per second, 256 sampling points occupy the length of about 32ms. 256 points are conducive to the calculation of FFT transform. Based on the overall performance of the system, this paper selects 256 sampling points as a frame, and uses the method of continuous framing to frame the speech signal. In order to reduce the influence of incoherent signal on two adjacent frames, a window function is usually used before each frame of speech is processed. The window function is mainly used to increase the continuity of the adjacent two frame signals, and better preserve the overall characteristics of the signal. The commonly used window functions have rectangular window functions and Hamming window functions, and their expressions are as follows (where N is the length of a frame):

$$w(n) = \begin{cases} 1, 0 \le n \le (N-1) \\ 0, n = others \end{cases} \tag{3}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], 0 \le n \le (N-1) \\ 0, n = others \end{cases} \tag{4}$$

The short-time characteristic of speech signal is closely related to the selection of window function, and the choice of window function w (n) can affect the result of short-time analysis to a great extent. Finally, the endpoint detection of speech signal, for the characteristics of the Android platform and system functional requirements, in this project, an endpoint detection method based on short-time energy and short-time zero crossing rate is adopted. The endpoint detection technology based on short time energy and short-time zero crossing rate belongs to time domain analysis. The method is simple, small amount of calculation, but also can guarantee a certain degree of reliability.

3.2 MFCC Feature Extraction

In this paper, MFCC coefficients, which are suitable for Android platform English pronunciation training system, are selected as characteristic parameters. The Mel frequency represents the sensitivity of the human ear to the frequency. The research shows that the sound level of human ear is not consistent with the actual frequency of sound. In fact, the human auditory system has a special nonlinear system that has different auditory sensitivity for different frequencies of signals. In the low frequency part, the human ear is sensitive; in the high frequency part, the human ear feels rougher and rougher. Below 1000Hz, the sensitivity and frequency are approximately linear, and the logarithm of the sensitivity increases approximately 1000Hz. Type 3 is the relationship between the two, and

figure 1 is the diagram of the two.

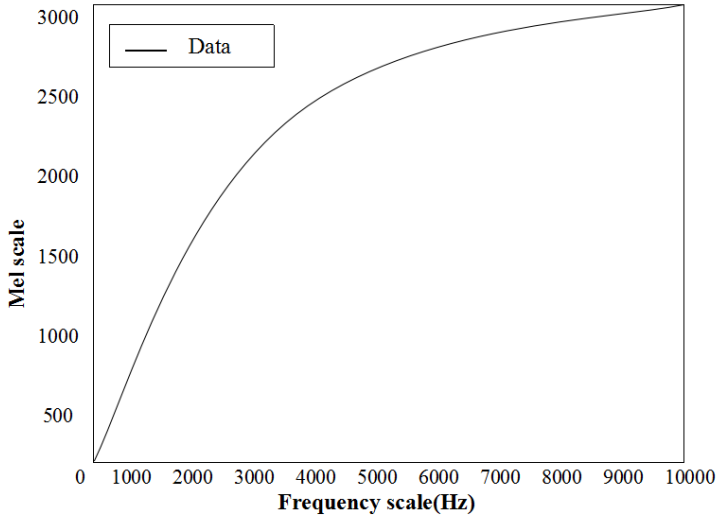

Figure 1 Mel scale and frequency diagram

The relationship between the Mel scale and the frequency has the following relation. (F is the real frequency of the signal):

$$f_{mel} = 2595 * \log_{10}(1 + f / 700) \tag{5}$$

The calculation process of MFCC is shown in figure, and the concrete steps of extracting MFCC characteristic parameters are as follows:

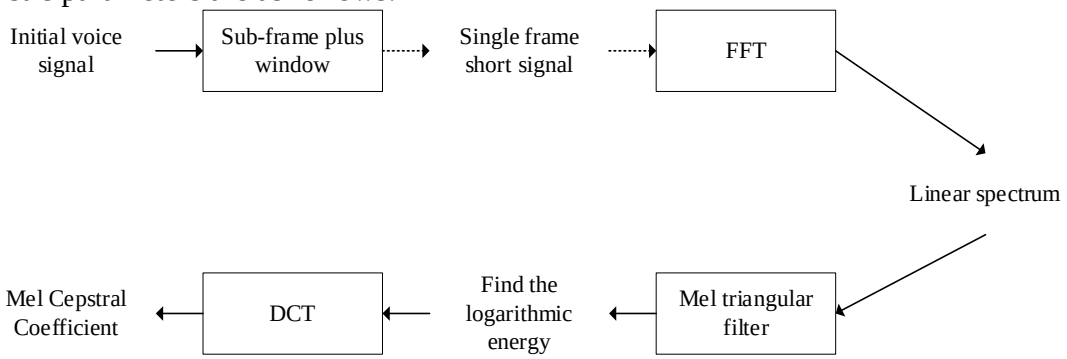

Figure 2 MFCC parameter extraction process

3.3 Speech Signal Pattern Matching

Dynamic time warping (DTW) is a nonlinear regularization method that combines time warping and distance disparity computation. The characteristic vector sequence of standard reference template is R={R (1), R (2),... R (m),... R (M)}, where M is the total number of frames containing speech standard template, m sequence label standard template frame, R (m) is the feature vector of the M Frame of speech.

The feature vector sequence of the input speech test template is T={T (1), T (2),... T (n),... T (N)}, where N is the number of frames in the voice test template, n as template in the speech sequence label, T (n) as the feature vector of the N frame of speech. The matching similarity between the test template and the reference template feature vector can be expressed by the distance between the vectors, and the larger the distance, the smaller the matching similarity. The distance between characteristic vectors T (n) and R (m) is usually represented by Euclidean distance:

$$d[T(n), R(m)] = \sum_{i=1}^{p} (t_i - r_i)^2 \tag{6}$$

In the formula, Ti and RI represent the dimensional feature vectors of T (n) and R (m) respectively,

and P is the order of characteristic vectors. Dynamic time warping algorithm is to find the time warping function, and the time axis n of the test template is mapped nonlinearly to the time axis m of the reference template, so that the minimum matching distance between the test template and the reference template is D, that is,

$$D = \min_{w(n)} \sum_{n=1}^{N} d[T(n), R(w(n))] \tag{7}$$

DTW algorithm is an optimization problem. It uses the time warping function w (n) to describe the time correspondence between the test template and the reference template, so as to solve the matching distance between the two templates in this case.

4. Result analysis and discussion

The part of speech recognition is the key component of intelligent spoken English pronunciation training. The basic theory and algorithm of speech recognition are introduced in this paper, as well as the process of speech recognition, such as preprocessing, feature extraction and pattern matching. Based on the analysis of the commonly used speech recognition technology, a recognition scheme suitable for this system is selected. According to the limited data processing ability of Android platform and the requirements of real-time and reliability of the system, the algorithm of feature extraction and DTW dynamic time warping for pattern matching is improved and selected, and the specific parameter details are given in MFCC. Based on this, we test the above algorithms based on the set of test environment. The test mainly includes three aspects, they are voice input test, scoring accuracy test and feedback correction test.

In the beginning, we did voice input tests, and in this test, the specific tests included: English pronunciation test when the user read, whether the system can correctly identify and record the corresponding pronunciation. Figure 3 is a speech signal of this test. We determine the starting point and termination point of the speech segment by short-time energy and short-time zero crossing rate, as shown in figure 4, through pictures, we can clearly determine the endpoint of the speech segment, which is prepared for further processing of the signal.

The test case data includes 20 vowels, 24 consonants, and 12 words, and we take the first pronunciation as the test. We define test results as: success rate = number of successful use cases / total number of test cases. Specific test shows in Table 1, we can see that the system for vowel and word can all correctly input, but because some consonants pronunciation time is short, pronunciation energy is small, the system can’t correctly identify input.

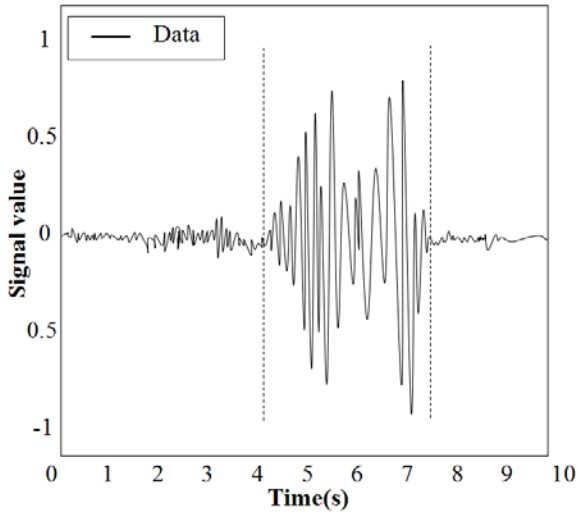

Figure 3 Original waveform

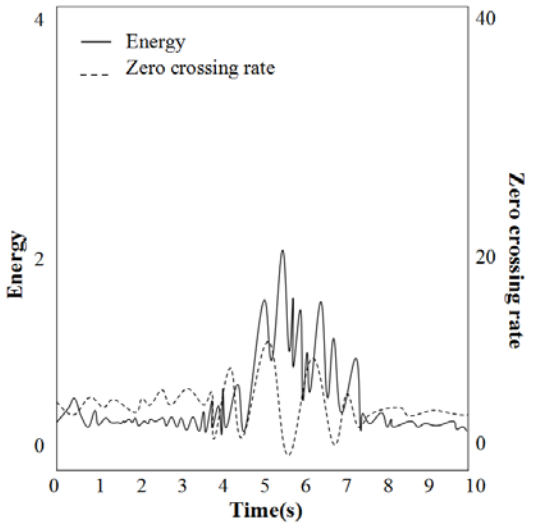

Figure 4 Endpoint test results of speech signals

Table 1 Test cases and results

Pronunciation type		vowel	consonant	word
Voice input test	Quantity	20	24	12
	Success rate	100%	91.6%	100%
Scoring accuracy test	Quantity	20	24	12
	Accuracy	95.52%	75.26%	91.68%
Feedback corrective test	Quantity	20	/	12
	Efficient	85%	/	75%

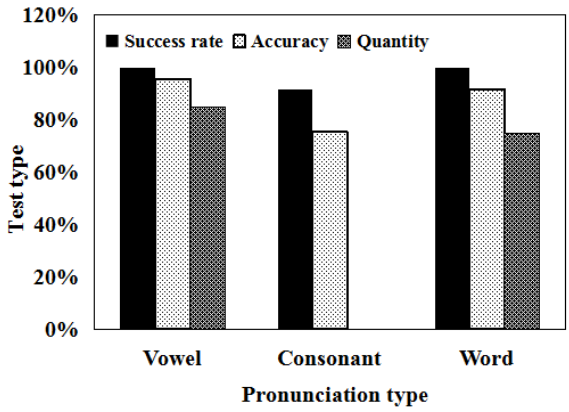

Figure 5 Test cases and results

Then, we test the scoring accuracy, and the specific test is to test the similarity between the automatic scoring system and the expert artificial score to verify the reliability of the system. The test case data include 20 vowel pronunciation scores, 24 consonant pronunciation score and 12 pronunciation score. Here we define test results: Suppose that i pronunciation system is automatically rated as S_1, Experts rated the pronunciation as S_2, The score similarity of the pronunciation is defined as $R_i = 1 - |S_1 - S_2| / S_2$, The pronunciation score accuracy of n samples was defined as $\overline{R} = (R_1 + R_2 + \cdots + R_n)/n$, By Figure 5 we can see that the system for vowel and word pronunciation scoring accuracy is higher, and expert experience score similarity of more than 90%, which has high reliability; For pronounce consonants, the accuracy of consonant pronunciation is limited by the length of pronunciation and energy, and the accuracy is 75.26%. It can reflect the quality of pronunciation to some extent, but it is not ideal enough.

Then we do feedback correction test, and the specific test content is based on the sound formant image contrast given by the system, to detect whether the effective feedback correction can be achieved. The test case data includes 20 vowels and 12 pronunciations (consonants without formant). The test results are defined as: for each pronunciation, if the pronunciation of the formant image is improved, the pronunciation score can be improved, otherwise the effect is small or invalid. The effective rate = the effective number / total pronunciation number. By Figure 5 we can see that the system through pronunciation formant image contrast method, pronunciation correction efficiency is about 80%, to a certain extent, guide the learners to pronounce correction. From the above analysis, we can see that the English learning model designed in this paper has good effect on speech input, but the accuracy of scoring accuracy test is not high, but it can still reflect the quality of pronunciation to some extent, This is in the future research we need further improvement, finally in the feedback correction test, we only use the vowel pronunciation and word pronunciation test, the test results of the correction efficiency can reach about 80%, and the learners' pronunciation can play a certain role in the correct, although the English learning model has achieved some results, but there is still room for improvement, in the future research, we will further improve and enhance the various evaluation indicators.

5. Conclusion

At present, with the continuous development of science and technology, English teaching is gradually moving toward an information-based and intelligent direction, and in the classroom teaching, learning to accept the knowledge of English is mainly achieved through listening and speaking ability, therefore, this point in listening and spoken English, English learning model of intelligent algorithm. This paper mainly introduces the basic theory, algorithm of speech recognition, the pretreatment, feature extraction and pattern matching and other speech recognition processing flow. Based on the analysis of the commonly used speech recognition technology, a recognition scheme suitable for this system is selected. According to the limited data processing ability of Android platform and the requirements of real-time and reliability of the system, the algorithm of pattern matching using MFCC feature extraction and DTW dynamic time warping is improved and selected, and the specific parameter details are given. Finally, we tested the algorithm, and tested the accuracy of the pronunciation scoring algorithm. For vowels and word pronunciation, the similarity score with expert scores was above 90%; The effective rate of pronunciation correction is 80%, which can improve the pronunciation level of learners to a certain extent.

References

[1] Choi J, Yi Y. Teachers' Integration of Multimodality Into Classroom Practices for English Language Learners, *Tesol Journal*, 2016, 7(2):304-327.

[2] Paramudia P, Habil H. Initial Oral English Communication Needs of Learners in the Business English Classroom, *Journal of Education & Learning*, 2015, 9(1):35.

[3] Sun Y, Franklin T, Gao F. Learning outside of classroom: Exploring the active part of an informal online English learning community in China, *British Journal of Educational Technology*, 2017, 48(1),47-53.

[4] Pingxiao W. Research on the English teaching and autonomous learning based on multimedia platform and smart classroom system, *International Journal of Smart Home*, 2016, 10(9):373-384.

[5] Huang L H, Chen C Y. GM(0,N)Model -Based Analysis of the Influence Factors of Network English Learning Platform, *Journal of Grey System*, 2016, 19(1):31-40.

[6] Hegedus S J, Dalton S, Tapper J R. The impact of technology-enhanced curriculum on learning advanced algebra in US high school classrooms, *Educational Technology Research & Development*, 2015, 63(2):203-228.

[7] Hashmi N A. Computer-Assisted Language Learning (CALL) in the EFL Classroom and its Impact on Effective Teaching-learning Process in Saudi Arabia, 2016, 5(2):202-206.

[8] Rui Z, Yang H Y. Construction and analysis of the foreign language learning model based on the classroom network environment, *Agro Food Industry Hi Tech*, 2017, 28(1):1205-1208.

[9] Sampson D G, Sergis S, Vlachopoulos P. Flipped Classroom Teaching Model Templates for STEM Education, *Politiques Et Management Public*, 2015, 24(11):41-68.

[10] Webster G D, Gesselman A N, Crosier B S. Avoidant adult attachment negatively relates to classroom popularity: Social network analysis support for the Parent–Partner–Peer Attachment Transfer model, *Personality & Individual Differences*, 2016, 96:248-254.