





Optimization of Computer Aided English Pronunciation Teaching System Based on Speech Signal Processing Technology

Lina Ma¹  and Yanjie Lei² 

¹School of Foreign Languages, Zhengzhou Sias College, Zhengzhou, Henan, 451100, China, malina15290833069@163.com

²Henan Provincial Senior Technology School of Industry & Information, Zhengzhou, Henan, 451150, China, leiyanjie136@163.com

Corresponding author: Lina Ma, malina15290833069@163.com

Abstract. After the development of speech signal processing technology has matured, various language learning tools have begun to emerge. The speech signal processing technology has many functions, such as standard tape reading, making audio aids, synthesizing speech, and performing speech evaluation. Therefore, the adoption of speech signal processing technology in English pronunciation teaching can meet different teaching needs. Voice signal processing technology can present teaching information in different forms, and promote multi-form communication between teachers and students, and between students and students. This will help stimulate students' interest in learning English and improve the overall teaching level of English pronunciation. This research first investigates and studies the current level of English pronunciation mastery. After combining the relevant principles of speech signal processing technology, it puts forward the areas that need to be optimized in the design of the English pronunciation teaching system. Through the demand analysis and function analysis of the system, this research uses speech signal processing technology to extract the characteristics of the speech signal---Mel Frequency Cepstrum Coefficient (MFCC), The system's speech signal preprocessing, speech signal feature extraction and dynamic time warping (DTW) recognition algorithms are optimized. At the same time, this research combines multimedia teaching resources such as text, pronunciation video and excellent courses to study the realization process of each function of the system.

Keywords: Voice signal processing technology; English pronunciation teaching system; System Optimization; Computer-assisted English teaching;

DOI: <https://doi.org/10.14733/cadaps.2021.S3.129-140>

1 INTRODUCTION

Language is the main tool for communication between people. Through promptly, accurate and objective feedback information, it can help students detect the gap between their pronunciation and standard pronunciation, and correct pronunciation errors. Golonka, E-M believed that pronunciation is the basis of language. As long as the pronunciation is accurate and fluent, even if the vocabulary and grammar points are limited, it will sound quite authentic [1]. On the contrary, if the pronunciation is below a certain level, effective communication cannot be established even if the grammar and vocabulary are completely correct. Although speech is so important, it is difficult for learners to correct inaccurate pronunciation entirely by themselves [2]. How to improve pronunciation teaching in the process of English teaching and minimize inaccurate pronunciation has become an urgent problem in English pronunciation teaching. The main contents of speech signal processing technology is the basic knowledge of speech signal processing and various analysis and processing techniques of speech signal. Including time domain and frequency domain processing, homomorphic processing, linear predictive analysis, vector quantization, hidden Markov model technology, speech detection analysis, etc. Gilakjani, A. P pointed out that the development of voice signal processing technology had made it possible for humans to interact with computers. The application of speech signal processing technology to the teaching of English pronunciation is helpful to the formation of students' sense of language and can also integrate course resources [3]. So, with the voice signal processing technology, the classroom atmosphere can be enlivened, and teaching methods can be enriched, thereby reducing the difficulty of teaching for teachers. Therefore, this research aimed at the shortcomings of the current English pronunciation teaching, combined with the relevant principles of speech signal processing technology, proposed the design of the English pronunciation teaching system, and described the optimization process of the system function in detail. This research used speech signal processing technology to extract the characteristics of the speech signal---Mel Frequency Cepstrum Coefficient (MFCC), and through matching with the standard speech pronunciation sample library, the gap between the pronunciation of the test speech and the pronunciation of the standard speech was obtained. At the same time, this research combined multimedia teaching resources such as text, pronunciation videos and quality courses to test the function of the English pronunciation teaching system, and conduct research in the practical application of teachers and students, and get feedback. Of course, the increasing maturity of voice signal processing technology also provides the possibility to realize this kind of system. This system also has positive significance in the promotion of English pronunciation teaching.

2 DESIGN AND OPTIMIZATION OF ENGLISH PRONUNCIATION TEACHING SYSTEM BASED ON SPEECH SIGNAL PROCESSING TECHNOLOGY

The increasingly mature speech signal processing technology has injected vitality into the pronunciation teaching system, making the function of the English pronunciation teaching system more effective. To some extent, mastering English pronunciation plays a decisive role in accumulating vocabulary. If the learner is proficient in the pronunciation of English phonetic symbols and words, then in the process of memorizing words, the learners can greatly reduce the heavy learning burden and pressure, easily master English words and gradually develop a strong interest in English. The integration of information technology and English courses can create a real English learning environment for learners, and the application of multimedia technology can also effectively stimulate learners' interest in learning, thereby forming a motivation for lasting learning [4].

2.1 Overviews of Speech Signal Processing Technology

The basic process of speech recognition is shown in Figure 1. It mainly contains several functional packages such as speech signal preprocessing, feature extraction, feature modeling (building a reference pattern library), similarity measurement and post-processing. A speech recognition system mainly contains two phases of training and recognition. It is necessary to preprocess the input original speech and perform feature extraction. The preprocessing module is to process the input original voice signal, filter out unimportant information and background noise, etc., and perform endpoint detection of the voice signal. The characteristic parameters commonly used in speech recognition systems are amplitude, energy, zero-crossing rate, linear prediction coefficient (LPC), LPC cepstrum coefficient (LPCC), line spectrum pair parameter (LSP), short-term frequency spectrum, formant frequency, reflecting the human ear Mel frequency cepstrum coefficient (MFCC) of auditory characteristics, etc. Therefore, the selection and extraction of speech characteristics are the key to system construction.

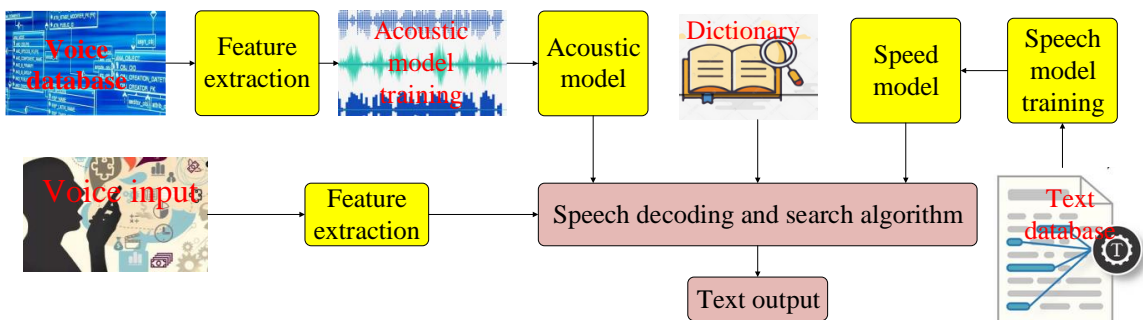


Figure 1: Framework diagram of the principle of the speech recognition system.

2.2 System Requirement Analysis and Function Analysis

The system is designed mainly based on the Android phone system to meet the needs of English pronunciation, pronunciation follow-up, pronunciation evaluation and pronunciation feedback in English pronunciation training. The system should fully consider the characteristics of the Android smart phone system when designing the system, and make reasonable calls to the functions of the Android system. When designing the system, it is necessary to combine multimedia technology to express the internal structure of the pronunciation mouth in the form of animation video. In this way, the system is first started, then the driver is initialized and OSAL is initialized and started, and finally it enters the polling phase of the event.

2.2.1 Login function analysis

According to the demand analysis, the system introduces the characteristics of age and gender into the voice recognition system because the gender and the voice characteristics of each age group are different. This can improve the accuracy of pronunciation recognition, and the system will store the learner's information in the database. The login subsystem provides user login with two permissions. One is to log in as a teacher, using the built-in teacher account, which can perform new user registration, modification of registration information and database update, and save the registration information for account login. Another is to log in as a student, you can only choose an existing account to enter and use, and you cannot modify user information. The use case diagram is shown in Figure 2(a).

2.2.2 Analysis of pronunciation course navigation function

Figure 2(b) is a use case diagram of the course navigation function. Through the analysis of the user's needs, when the user returns from the pronunciation practice history statistics of the phonetic transcription, he can enter the course navigation function. According to the needs of users, you can also browse the historical practice scores to see the previous practice scores. Use the main window interface to return to the login subsystem, view the system help, and exit the system.

2.2.3 Analysis of pronunciation practice function

Obtained from the demand analysis, after the user learns the pronunciation, if he wants to practice his own understanding of the English pronunciation, the system provides the user with the corresponding switch label. The use case diagram is shown in Figure 2c. This function mainly provides users with pronunciation exercises. First select the content of the exercises, and then the user starts to pronounce, and at the same time records their pronunciation. Then play back the recording and hear your own pronunciation.

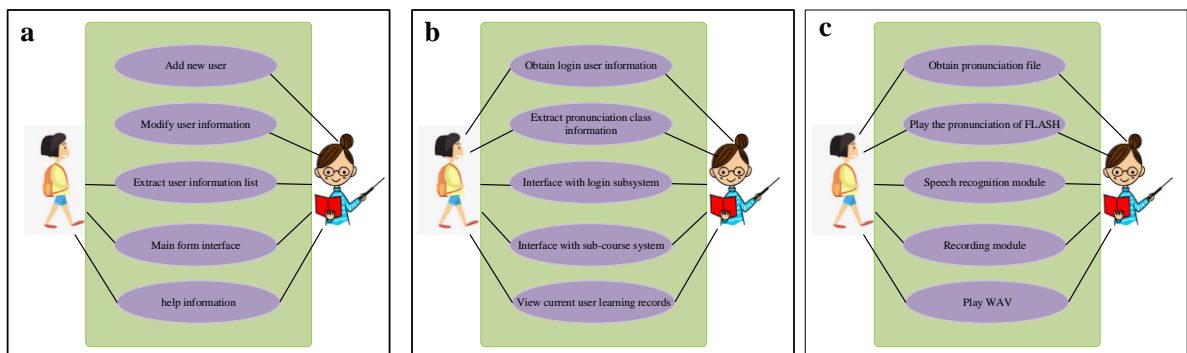


Figure 2: An example diagram of the various functions of the system in this study (a: login function; b: pronunciation course navigation function; c: pronunciation practice function).

2.3 The Overall Layout of English Pronunciation Teaching System According to Speech Signal Processing Technology

Liu, S.-C and his group's research showed that the learners' various senses should be fully mobilized to maximize the acquisition of knowledge and information during language learning. Among them, especially the mobilization of visual and auditory organs should be considered [5]. Gilakjani, P considered that a new model of teaching environment can be created by efficaciously integrating information technology into the teaching process of many disciplines [6]. Therefore, in English pronunciation learning, the design of an English phonetic transcription teaching system based on speech signal processing technology can create a good English phonetic learning environment for learners. It can better help learners master the pronunciation of words and sentences, and improve the efficiency of learners' autonomous learning. Based on the premise of effective stimulation to the learners' senses, the presentation of learning content should be organized and arranged as much as possible. This requires us to consider the effective integration of the curriculum content of the English British Standard with information technology. So the overall design purpose of this article is mainly on the App side, in order to improve the students' English pronunciation, and then improve the students' oral English level, and finally achieve the purpose of promoting the students' overall English level. The core framework of this research system can be seen for detail in Figure 3:

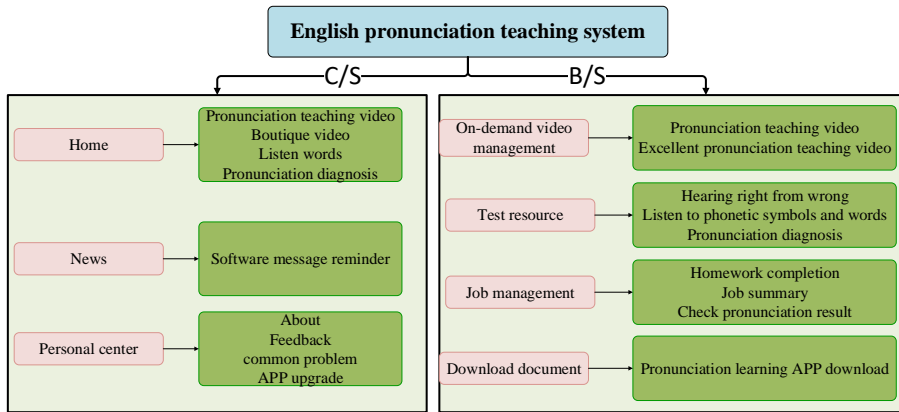


Figure 3: The core framework of the English pronunciation teaching system in this research.

2.4 Optimization of Speech Signal Processing

The voice signal processing module generally includes five steps: voice signal digitization, endpoint detection, framing, windowing, and pre-emphasis. The voice signal of this system is recorded through the microphone of the Android phone. In the background, the *android.media.Audio Record* API provided by Android is used to configure the sampling rate and quantization bits in the digitization process to obtain the desired digital audio signal. The aim of pre-emphasis is to strengthen the high-frequency signal of the voice, remove the low-frequency signal in the voice signal, and make the signal spectrum flatter [7]. For making the voice signal have the same segment on the time axis, short-term analysis can be used to divide the voice signal into frames. This system uses a sampling rate of 55.2Hz, that is, 5520 sampling points per second, and 256 sampling points occupy about 27ms. For the purpose of reducing the affect aroused by the distance relationship between the two frames, a certain window function $Q(n)$ is used to multiply the speech signal $s(n)$ when processing the frame signal. Window functions generally use rectangular windows or Hamming windows, and their expressions are as follows:

Rectangular window

$$Q_r = \begin{cases} 1 & (0 \leq n < N-1) \\ 0 & (Other) \end{cases} \quad (1)$$

Hamming window

$$Q_{HM} = \begin{cases} 0.5 - 0.46 \cos(2\pi n / (M-1)) & (0 \leq n < M-1) \\ 0 & (Other) \end{cases} \quad (2)$$

Hanning window

$$Q_{HN} = \begin{cases} 0.5 - 0.5 \cos(2\pi n / (M-1)) & (0 \leq n < M-1) \\ 0 & (Other) \end{cases} \quad (3)$$

In this study, considering the different application ranges of the rectangular window and the Hamming window, the system uses the rectangular window function when detecting the end point of the signal, and uses the Hamming window when transforming the frequency domain. The influence of background noise in the real environment increases the difficulty of accurate endpoint detection for voice signals. For the analysis of the characteristics of the Android platform and the functional requirements of the system, this system uses a short-term zero-crossing rate analysis

method to perform endpoint detection [8]. We define the short-term zero-crossing rate of the speech signal:

$$K_m = 0.5 \sum_{j=0}^{M-1} |\text{sgn}[x_m(m)] - \text{sgn}[x_m(j-1)]| \quad (4)$$

Where A is a symbolic function, namely $\text{sgn}[x] = \begin{cases} 1 & (x \geq 0) \\ -1 & (x < 0) \end{cases}$. Voiced sounds with low energy frequency have a lower zero-crossing rate, while unvoiced sounds with high energy frequency have a higher zero-crossing rate.

2.5 Feature Extraction and Optimization of Speech Signal

So far, after processing the speech signal, several often selected as characteristic parameters are: linear prediction coefficient (LPC), linear prediction cepstral coefficient (LPCC) and Mel cepstrum coefficient (MFCC). These parameters can describe the characteristics of the speech signal. Since the MFCC characteristic parameters are better in noise immunity and robustness, it has good recognition performance and noise immunity. Moreover, the Mel cepstrum coefficient, the MFCC characteristic parameter, can fully characterize the human hearing characteristics. Below 1KHz, the human ear's sensitivity to sounds of different frequencies has a roughly linear relationship with the frequency, and above 1KHz, there is an approximate logarithmic increase. Figure 4 is a description of this relationship.

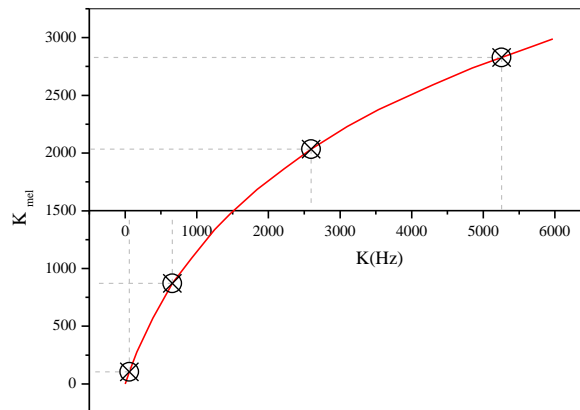


Figure 4: The relationship between human ear hearing range and Mel frequency.

The equation below may express the relationship between Mel scale and frequency (equation k is the true rate of the signal), and the unit of actual frequency k is HZ.

$$k_{mel} = 2595 \log_{10}(1 + f / 700) \quad (5)$$

This study uses the relevant functions provided by the MATLAB voice analysis tool VoiceBox to extract the MFCC parameters of the voice. The input voice sequence is a wav format audio with 64-bit quantization and a sampling rate of 55.2KHZ. First, use the Hamming window function to

frame the speech sequence, and require that each frame has a length of 512 points and a frame shift of 256; Then perform FFT transformation on each frame of speech signal to obtain the corresponding frequency spectrum and amplitude spectrum; Finally, the FFT parameters are de-logarithmized through the Mel filter group and then the cepstrum is calculated to obtain the 12-dimensional MFCC parameters. Therefore, the partial extraction process of MFCC in this study is as follows:

① Define mfcc() function

```
function cc=mfcc (x,fs,p,frameSize,inc)
bank=melbankm (p,frameSize,fs,0,0.5,'m');
% Normalized Mel filter bank coefficients
bank=full(bank); bank=bank/max(bank(:));
p2=p/2;
% DCT coefficient, p2*p
for k=1:p2; n=0:p-1;
dctcoe f(k,:)=cos((2*n+1)*k*pi/(2*p));
end
```

② Give an input to the mfcc() function

```
Fname=sprintf f('%s\\%d.wav',model File Name,i);
x=fname; [x,fs]=audio read(x)
```

③ Call the function mfcc() to get the output result.

2.6 Optimization of Dynamic Time Warping (DTW) Recognition Algorithm

Golonka, E. M pointed out that in the process of pronunciation recognition, the frequently used discrimination algorithms include dynamic time warping algorithm (DTW) and algorithms based on hidden Markov model (HMM) [9]. In view of the fact that the system is doing the recognition of English pronunciation, the remaining functions of the system will also consume a lot of resources, and in line with the requirements of simplicity and easy calculation, this study uses the DTW algorithm to calculate the similarity between the test pronunciation and the standard pronunciation. At the same time, due to the traditional DTW algorithm, the whole word is used as the basic recognition unit for training and recognition, and the problem of word segmentation is not considered. After experiments in this study, it is found that the optimized DTW search path will not search the entire data matrix area in the graph. This area is a parallelogram named "Search domain" in the figure. A parallelogram formed by the origin and the end point (upper right corner) of the rectangular grid line and the two sides $2/3$ and $3/2$. Finally, the following two points X_a and X_b are calculated, which are the best choice for search speed and similarity in such an area. The improvement of the traditional DTW algorithm is mainly to improve the efficiency of voice comparison. Figure 5 shows the search path diagram of the improved DTW algorithm.

y_{\min} and y_{\max} are calculated as follows:

$$y_{\min} \begin{cases} 0.5x & 0 \leq x \leq x_b \\ 2x + (G - 2V), & x_b \leq x \leq V \end{cases} \quad y_{\min} \begin{cases} 2x & 0 \leq x \leq x_b \\ 0.5x + (G - 0.5V), & x_a \leq x \leq V \end{cases} \quad (6)$$

In this research, *Matlab* is used to implement the algorithm. The *dist()* function is defined in *Matlab* to represent the improved DWT algorithm. By verifying the comparison efficiency of the traditional DTW algorithm and the improved DTW algorithm, it is found that the comparison rate of the improved DTW algorithm is greatly increased contrasted with the unmodified DTW algorithm.

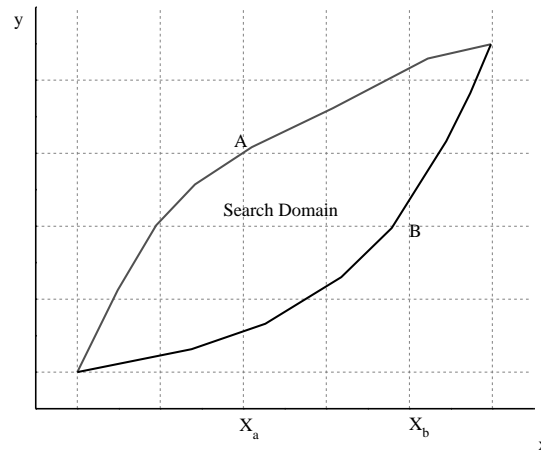


Figure 5: The search path diagram of the improved DTW algorithm.

3 THE FUNCTION REALIZATION OF THE COMPUTER-ASSISTED ENGLISH PRONUNCIATION TEACHING SYSTEM OPTIMIZED BY SPEECH SIGNAL PROCESSING TECHNOLOGY

Rao, K believed that the basis of speech recognition is the ability to have a standard speech sample library for recognition and comparison [10]. After the speech of the sample library is processed and analyzed, it can represent and reflect the characteristics of the speech. These models will become a template for reference comparison, directly related to the objectivity and accuracy of platform feedback. The key to the voice diagnosis module in this system is to upload the collected user's phonetic transcription and pronunciation to the server for comparison with the template in the standard voice sample library [11]. According to the comparison result, choose whether to provide pronunciation guidance to the user. The entire standard speech library of this system has a total of 259 segments, of which 6-7 segments of audio are given for the pronunciation of each English phonetic symbol, including wrong speech and correct speeches, which is convenient for subsequent comparison. Now take the pronunciation module and pronunciation diagnosis module of listening phonetic symbols as an example for description.

3.1 Realization of Phonetic Transcription Pronunciation Module

The listening phonetic transcription module mainly examines 44 English phonetic symbols. For each phonetic symbol audio, four phonetic symbols are given. Choose the one corresponding to the given phonetic symbol audio [See Figure 6(a)]. In this module, students' ability to recognize phonetic symbols can be strengthened, and at the same time, students' listening skills can be improved. The interface is shown in Figure 6(b) below. When the user clicks the "Check pronunciation" button after the selection is completed, the system will automatically determine the correctness. At the same time, the system will pop out the prompt box to remind the user, and urge the user to answer the question carefully and master the pronunciation knowledge of the phonetic symbol.

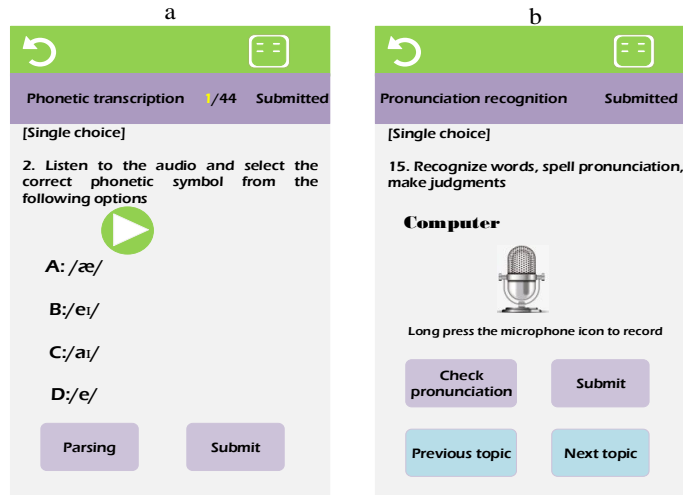


Figure 6: Schematic diagram of the interface displayed on the client side of the pronunciation module (a) and pronunciation diagnosis module (b) of this system.

3.2 Realization of Pronunciation Diagnosis Module

The main principle of the voice diagnosis module in this system is to compile the Matlab code into a dynamic link library, and then deploy it to the server in the form of Webservice. The Android side uses a built-in recording method to record the user's audio, and then automatically upload it to the server, where the user's voice is extracted and recognized. Finally, the recognition result is returned to the Android terminal, displayed and presented to the user. The interface display is shown in Figure 6(b). Click on pronunciation diagnosis, enter the pronunciation teaching video module, you can recognize the pronunciation of phonetic symbols. According to the phonetic transcription picture given at the top of the page, press and hold the microphone icon to start recording audio. After the pronunciation is over, let go of the icon, and the system will automatically upload the recorded audio to the server for judgment. If the learner's pronunciation is correct, the system will display "Phonetic symbol detection is correct" and you can enter the next question to answer the study; If the learner's pronunciation is wrong, the system displays the warning "Phonetic notation detection error, please click the correct pronunciation button!" At the same time the page will display the "Correct pronunciation" button.

4 TEST OF COMPUTER-ASSISTED ENGLISH PRONUNCIATION TEACHING SYSTEM OPTIMIZED BY SPEECH SIGNAL PROCESSING TECHNOLOGY

Testing is very important for developing a system. Faced with intricate problems, errors may occur at every stage of the software life cycle. If the errors in the software are not found and corrected before the software is put into productive operation, these errors will be exposed in the production process sooner or later, which will often cause very bad consequences. Therefore, we must attach great importance to software testing. When designing and developing the system, the system test process we used can be illustrated in Figure 7. In the equipment management system, the unit test is mainly to clean up the junk code in all source programs. The focus of integration testing is the connection between modules and the transfer of parameters. In the development process of this system, the main content of integration testing includes interface testing between units, performance testing, boundary and performance testing under artificial conditions. The focus of system testing is the operation of the entire system and its compatibility with other software. In

order to find defects and measure product quality, tests are carried out in accordance with the function and performance requirements of the system. According to the function of each interface, the system test has interface control function and data transfer relationship, including data transfer between form and form and data transfer between form and database. This part of the content takes the login function test, pronunciation practice function test, and system recognition rate test as examples.

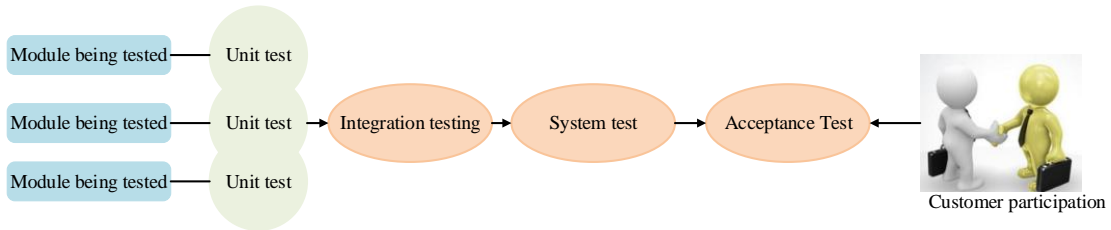


Figure 7: Schematic diagram of the testing process of the system in this study.

4.1 Login Function Test

By traversing the user name information in the database, using SQL statements to query and format specific information. If the last record is searched and the same record is not found, it proves that the username is available. Enter the following examples (see Table 1). After testing, enter the content that meets the requirements to complete the new user login.

<i>Test items</i>	<i>Test input</i>	<i>Test Results</i>
<i>New user name</i>	<i>1 word; 2-4 letters; contains letters and numbers</i>	<i>Verify that the authenticated new username can pass</i>
<i>Gender</i>	<i>Male, female, other</i>	<i>Verify that the authenticated user's gender input is normal</i>
<i>Registration date</i>	<i>2020.06.20</i>	<i>Date that verified users can register</i>

Table 1: Examples of the system login function test.

4.2 Pronunciation Practice Function Test

In the pronunciation practice function module, the system automatically generates an analytical report on the user's learning situation for the user's practice for a period of time, and gives suggestions such as grades (Excellent/Very, Good/Good/Try Again). The test process of pronunciation practice function is: the practice content for a period of time, the results of database opening, closing, updating, modification, etc., can be recorded in two TXT files in text form in real time. At the same time, another test tool is used. The main function of the test tool is to read out the practice content stored in the database for a period of time from the data table, and import it into another TXT file according to the date and test content.

<i>Test contents</i>	<i>Pitch score</i>	<i>Syllable score</i>	<i>Result</i>
<i>Phonetic symbol</i>	<i>Excellent</i>	<i>Good</i>	<i>Operating normally</i>
<i>Bus</i>	<i>Good</i>	<i>Excellent</i>	<i>Operating normally</i>
<i>/ɔɪ/</i>	<i>Try again</i>	<i>Good</i>	<i>Operating normally</i>
<i>/ə/</i>	<i>Excellent</i>	<i>Excellent</i>	<i>Operating normally</i>

Table 2: Examples of functional tests for system pronunciation practice.

After testing, the pitch and syllable scores of the user's learning records (see Table 2) are obtained to meet the user's needs for scoring in pronunciation practice.

4.3 System Recognition Rate Test

In this system, only the voice diagnosis module uses the voice recognition function, so only the function test of this module is performed here to verify whether the recognition performance of the previously designed English phonetic transcription auxiliary learning platform is achieved. In order to achieve this recognition rate test, the experiment uses a standard voice sample library that has been recorded. That is to record 2 correct and 2 wrong voices respectively. In order to minimize the impact of environmental noise, a relatively quiet laboratory is the ideal test environment for this test. The recognition rate test of this platform can be demonstrated in Table 3:

Phonetic	Results	number 1	number 2	number 3	number 4	Correct rate
		√	√	×	×	
/ɔ:/		1	1	1	0	0.85
/ɜ:/		0	1	0	1	0.91
/æ/		0	1	0	1	0.61
/aɪ/		1	0	1	1	0.83
/dʒ/		1	1	0	1	1
/ts/		1	1	1	1	0.96

Table 3. The recognition rate test results of this system.

In Table 3, 1 means the recognition is correct, and 0 means the recognition is wrong. Judging from the recognition accuracy rate in Table 3, the speech diagnosis module can basically make relevant judgments on the correctness of speech, and then cooperate with other functional areas in the platform to learn English phonetic symbols. Therefore, from the test results, in the recognition rate test of the voice diagnosis module, the system can basically achieve the corresponding learning effect proposed at the beginning of this research.

5 CONCLUSION

At present, the importance of speech signal processing technology applied to English pronunciation teaching system has been reached a consensus. This research combined with the relevant principles of speech signal processing technology, and put forward the design of pronunciation teaching system in the field of English teaching. Through the demand analysis and function analysis of the system, this research optimized the system's speech signal preprocessing, speech signal feature extraction and dynamic time warping (DTW) recognition algorithm. This study used speech signal processing technology to extract the characteristics of the speech signal---Mel Frequency Cepstrum Coefficient (MFCC). By matching with the sample library of standard speech pronunciation, the gap between the pronunciation of the test speech and the pronunciation of the standard speech was obtained. Also, this research combined multimedia teaching resources such as text, pronunciation video and excellent courses to study the realization of each function of the system. The system used voice matching technology to evaluate the pronunciation of English phonetics, so that learners could get feedback in time. The results of the test showed that the English pronunciation teaching system optimized under the speech signal processing technology in this research was very helpful for teachers' English pronunciation teaching and students' English pronunciation learning. In summary, the application of speech signal processing technology in English pronunciation teaching can improve the effectiveness of classroom teaching.

Lina Ma, <https://orcid.org/0000-0002-1079-978X>
 Yanjie Lei, <https://orcid.org/0000-0002-5420-7918>

REFERENCES

- [1] Golonka, E.-M.: Technologies for Foreign Language Learning: A Review of Technology Types and Their Effectiveness, *Computer Assisted Language Learning*, 27(1), 2014, 70–105. <https://doi.org/10.1080/09588221.2012.700315>
- [2] Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M. A.; Schuller, B.; Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, 5200–5204. <https://doi.org/10.1109/ICASSP.2016.7472669>
- [3] Gilakjani, A.-P.; Sabouri, N.-B.: Advantages of Using Computer in Teaching English Pronunciation, *International Journal of Research*, 2(3), 2017, 78–85. <https://doi.org/10.18869/ACADPUB.IJREE.2.3.78>
- [4] Levis, J.-M.: COMPUTER TECHNOLOGY IN TEACHING AND RESEARCHING PRONUNCIATION, *ACM Sigapl Apl Quote Quad*, 27(1), 2007, 184–202. <https://doi.org/10.1017/S0267190508070098>
- [5] Liu, S.-C.; Hung, P.-Y.: Teaching Pronunciation with Computer Assisted Pronunciation Instruction in a Technological University, *Universal Journal of Educational Research*, 4(9), 2016, 1939–1943. <https://doi.org/10.13189/UJER.2016.040902>
- [6] Gilakjani, P.; Abbas.: Teaching Pronunciation of English with Computer Technology: A Qualitative Study, *International Journal of Research*, 3(2), 2018, 94–114. <https://doi.org/10.29252/IJREE.3.2.94>
- [7] Wang, D.; Chen, J.: Supervised Speech Separation Based on Deep Learning: An Overview, *IEEE Transactions on Audio, Speech, and Language Processing*, 26(10), 2018, 1702–1726. <https://doi.org/10.1109/TASLP.2018.2842159>
- [8] Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, 4960–4964. <https://doi.org/10.1109/ICASSP.2016.7472621>
- [9] Golonka, E.-M.; Bowles, A.-R.; Frank, V.-M.; Richardson, D.-L.; Freynik, S.: Technologies for foreign language learning: a review of technology types and their effectiveness, *Computer Assisted Language Learning*, 27(1), 2014, 70–105. <https://doi.org/10.1080/09588221.2012.700315>
- [10] Rao, K.; Sak, H.; Prabhavalkar, R.: Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer, *IEEE Automatic Speech Recognition and Understanding Workshop*, 2017, 193–199. <https://doi.org/10.1109/ASRU.2017.8268935>
- [11] Henderson, A.; Frost, D.; Tergujeff, E.; Kautzsch, A.; Murphy, D.; Kirkova-Naskova, A.; Curnick, L.: The English pronunciation teaching in Europe survey: selected results, *Research in Language*, 10(1), 2012, 5–27. <https://doi.org/10.2478/V10015-011-0047-4>