Research and Application of Computer Artificial Intelligence Technology in Machine Pronunciation

Yisha Zhang*
Jiangxi University of Engineering

*Corresponding author: yishazhang@jxue.edu.cn

Abstract. The traditional oral English training system uses a layered training strategy, which cannot meet the system performance design requirements. For this reason, the thesis designs a computer artificial intelligence system for oral English training based on human-computer interaction. The system hardware structure is composed of C8051F021 single chip microcomputer, LM317 chip, LM337 chip, A/D chip and IN5822 diode, which is used to reliably control the sampling frequency of the chip. At the same time, the paper introduces Sugeno points to realize the correct evaluation of spoken pronunciation. First, the randomness of natural language pronunciation and the instability of the speech processing system are modeled by fuzzy measure and credibility, and then they are integrated into the Sugeno integration framework to evaluate the language score instead of the specific score. The experimental results prove the credibility and stability of the evaluation model in closed and open tests.

1. Introduction
The development of the oral English training system is based on the needs of students. At the same time, it explains English knowledge points with reference to the teaching plan proposed by the Ministry of Education. The functions provided include after-school tests, oral exercises, etc. In the process of actual use, students must register and log in to the spoken English training system, use various functions to arrange a study plan, and make reasonable arrangements for learning through tests. The current oral English training system can meet the needs of students, but there are still many problems that need to be solved uniformly. Therefore, a cloud computing-based oral English learning system is designed, and users can use mobile terminals to learn oral English knowledge, simulate oral English testing, and build exclusive dictionaries [1]. Administrators can perform learning resource management, user information management, and vocabulary management through the cloud server. The system can realize functions such as cloud computing-based mobile learning and cloud sharing of learning resources, enhance the user's learning experience, and effectively promote the user's oral learning. This article first gives an evaluation model of continuous reading in spoken English pronunciation. The continuous reading assessment of oral English is an important problem to be solved in computer-assisted oral learning. Although there are many factors that affect the spoken language evaluation algorithm, the randomness of spoken language pronunciation and the instability of existing speech processing systems have become two main problems to be solved.
2. System Design

2.1. Voice signal preprocessing

2.1.1. Speech production mechanism. In the process of human pronunciation, due to the contraction of the lungs, the compressed air flow is generated, and the bronchus passes through the glottis and airways to cause audio oscillations to produce sound. There are three different types of excitation in the process of human pronunciation, so three different types of sounds can be produced, namely, unvoiced, voiced, and popping. Although human beings can make an infinite variety of sounds, languages use limited phonemes as the basic pronunciation units. Generally, there are only dozens of phonemes in a language [2]. A phoneme is equivalent to a collection of coded symbols in a communication system, which includes a limited number of symbols. According to the different states of pronunciation actions, phonemes can be divided into two types: closed type and open type. The closed phonemes are the consonants in English pronunciation, and the open phonemes are vowels. Although some phonemes are basically unobstructed in their vocal channels, the vocal tracts of certain phonemes are relatively narrow and produce slight friction sounds. This kind of phonemes are called semivowels.

2.1.2. Digitalization of analog signals. The voice signal is a one-dimensional analog signal whose amplitude changes continuously with time. Only the digitally processed signal can be analyzed by the computer. That is, the digitization of voice signals is a prerequisite in digital processing. The process of speech signal digitization includes two processes: sampling and quantization. After these two processes, a discrete digital signal in time and amplitude can be obtained.

2.1.3. Time-dependent processing. As a typical non-stationary signal, the production of speech signal is closely related to the movement of the articulator. This physical movement is much slower than the speed of sound propagation [3]. Therefore, we usually assume that the speech signal is short-term. Stationary, that is, the spectral characteristics within 10 to 20 milliseconds can be approximately regarded as constant. Like this time-dependent processing method, usually a finite-length window sequence \{w(m)\} is used to intercept a segment of speech signal for analysis, and let this window East China analyze the signal near a certain moment. Its general formula is:

\[ Q_n = \sum_{m=-\infty}^{\infty} T[x(m)] w(n-m) \]  

(1)

Among them, \( T[] \) represents some kind of operation, and \{x(m)\} is the input signal sequence. Equation (1) is in the form of convolution, so \( T[] \) can be understood as the output of a discrete signal \( T[x(m)] \) through two FIR (Finite Impulse Response) low-pass filters whose unit stimulus is \{w(m)\}, as shown in Figure 1. Since the window function is generally a smooth function with a large middle and small end of \( x(n) \), the filter corresponding to such a response has a low-pass characteristic. The response of frequency and bandwidth depends on the choice of window function.

![Figure 1. Schematic diagram of speech recognition](Image)
Through comparative analysis, the side lobes of the rectangular window are too high, and serious leakage will occur, so it is rarely used. The Hamming window has the lowest side lobes, which can effectively prevent leakage and has a more stable low-pass characteristic. Therefore, the Hamming window is the most widely used. In addition, the longer the window, the more obvious the averaging effect on the signal, and the higher the frequency resolution of the signal, but the lower its time resolution. Therefore, to quickly reflect time-varying information, the selected window length should be relatively short.

2.1.4. Voice endpoint detection. Correct endpoint detection is the key to all automatic speech recognition systems (ASR). Research results show that more than half of the recognition errors in speech recognition systems are caused by unreliable endpoint detection. Endpoint detection, as the name suggests, is to detect the start and end of the voice. There are two commonly used methods: multi-threshold front-end endpoint detection method and double-threshold front-end endpoint detection method. In order to facilitate real-time feature extraction, a dual-threshold front-end endpoint detection algorithm is usually used. Although the former can reduce the front-end error, there is a long time delay, which is not conducive to real-time control. If the time-domain parameters of the speech signal, the short-term parameters E (short-term energy) and Z (short-term zero-crossing rate), are used for endpoint detection, the problem in multi-threshold front-end detection can be solved.

Short-term energy: The energy in the speech signal will change significantly over time. Generally, the energy of the voiced part is much larger than the energy of the unvoiced sound. Therefore, it has a more obvious effect in the application of distinguishing unvoiced and voiced sounds. For signal \( \{x(n)\} \), the short-term energy is defined as follows:

\[
E_s = \sum_{n=-\infty}^{\infty} \left[ X(n) * w(n-m) \right]^2 = \sum_{n=-\infty}^{\infty} x_s^2(n)
\]

(2)

\( S_v(n) = x(n) * w(n-m) \) is a voice signal after windowing. The paper sets \( h(n) = w(n) \), then \( E_s = x(n) * h(n-m) \), where \( h(n) \) is the unit impulse response to the filter. Because short-term energy refers to the squaring of the signal, which widens the gap between the high and low signals, this method is not suitable for some specific scenarios. For this reason, we can use the short-term average amplitude instead of the short-term energy to represent the change in energy. The formula is:

\[
M_s = \sum_{n=-\infty}^{\infty} \left| x(n) \right| w(n-m) = \sum_{n=-\infty}^{\infty} \left| x_s(n) \right|
\]

(3)

The meaning of the short-term average zero-crossing rate refers to the number of times the signal in each frame passes the zero value. For discrete speech signals, the short-term average zero-crossing rate is essentially the number of times the sampling point symbol changes. The zero-crossing rate of signal \( \{x(n)\} \) is defined as:

\[
Z_a = \sum_{m=-\infty}^{\infty} \left[ \text{sgn}[x(m)] - \text{sgn}[x(m-1)] \right] w(n-m)
\]

(4)

2.2. System flow

Similar to the recognition process of a general speech recognition system, the recognition process of this system consists of two steps: the first step is to obtain the user’s pronunciation, and then the characteristic value of the user’s pronunciation is extracted. Different systems can select suitable characteristic values; the second step is to extract the feature value of is matched, aligned and scored in the reference model (HMM), which is the decoding process. Hidden Markov model is applied in the second step of decoding process. It includes an acoustic model, a language model and a word 5178 database. The acoustic model is trained through the Baum-Welch algorithm through a large number of
standard pronunciation samples [4]. The language model is a probability model of words or multi-phrases obtained by counting the words in the sentences in the corpus. The dictionary provides the range of recognizable words and the phoneme mapping relationship corresponding to the words. These two models are the basis of the probability data of the entire HMM, and their quality directly affects the accuracy of recognition.

Shown in Figure 2 is the process of the speech recognition system, where the training process is that a large number of standard pronunciation feature values are repeatedly calculated by the Baum-Welch algorithm, and the result is the reference model. The recognition process is the pronunciation feature value input by the user, which is decoded by the Viterbi algorithm.

![Figure 2. System flow chart](image)

2.3. **System module design**

For the voice processing schematic diagram described in Figure 2, the system is divided into three parts, namely: user interface GUI, input and output I/O, and scorer the. The relationship between the system modules is shown in Figure 3:

![Figure 3. System module relationship diagram](image)
2.3.1. Interface Design. VC++6.0 provides a set of methods for processing resources. Through the navigation of the "Resource View" of the work space, you can create resources or activate the corresponding resource editors of various resources to perform visual editing. Right-click the resource item in the "Resource View" in the "work space" on the left, vc++6.0 displays a shortcut menu, select the "insertdialog" option to add a new dialog box. Various controls in the toolbar on the right can be used to design the function of the target dialog box. Add button, edit, text and other related controls in the newly created dialog window and make appropriate adjustments. Modify the button control name to "select", "play", "record", "stop", "save" And other attributes; modify the attributes of the text control in turn, so that it can meet the normal display of sentence content, waveform, score feedback and error correction suggestions. Don't make sure that the external program interface can coordinate the control of the underlying I/O module and the Scorer module Call [5]. Enable the scorer module to operate on the data acquired by the latest I/O module.

2.3.2. I/O Design. There are two ways to process sound files in current computer systems: One is to use ready-made software, such as Microsoft's voice recorder, Sound Forge, Cool Edit, etc., which can record, edit, and play sound signals, but their functions are limited. Yes, in order to process sound data more flexibly and to a greater extent, you have to use another method, which is to use the multimedia services provided by Microsoft and write your own programs in the Windows 7 environment to perform sound processing to achieve some specific functions. The following begins to introduce the format of the sound file and the method of using the Visual C++ 6.0 development tool to process the sound file in the Windows7 environment. All the program codes in this article are compiled and passed under the Windows7 and Visual C++ 6.0 environment and run normally.

This system is designed to operate sound files with I/O stream, that is, save the language input by the user as a wav format file with the attribute of 16KHz/Mono/16bit/PCM, and then open the WAVE file during playback to obtain the sound data in it. Perform corresponding mathematical operations according to the required sound data processing algorithm, and then re-store the result in the WAVE format file. You can use the CFILE class to implement the read operation, or you can use another method, which is to use the multimedia processing functions provided by Windows (these functions are all started with mmio). The schematic diagram of WAVE file operation is shown in Figure 4:
3. Spoken speech feature extraction

Generally, when the system is used in an actual environment, the recognition performance of the sound feature value will be significantly reduced. The reasons are as follows: (1) Additive noise refers to the sum of the real speech signal and background noise. The speech signal is often interfered by background noise in the actual environment, and the background noise is usually additive. (2) Channel distortion the voice signal is also affected by some channel distortions such as the voice generation process, the recording process, and the channel distortion generated during the transmission process. (3) Other factors such as the extraction of feature parameters are also affected by some other man-made or instantaneous noises. It can be seen that the diversity of the mobile phone use environment determines that the system needs to consider the impact of noise in the environment, and the noise resistance of the traditional MFCC feature value alone is not the best. This article considers the use of a variety of parameter combinations to resist noise, introduce parameters that can suppress noise, and at the same time introduce more speech through parameter expansion. Assuming that the noise is stable, the precondition that the noise is relatively stable in each speech frame is obtained [6]. Therefore, we consider using the first-order and second-order difference of MFCC to suppress the stationary noise and improve the recognition rate. The difference calculation uses the following formula:

\[
2 \sum_{i=0}^{n} \sum_{j=0}^{k} c(k+i) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{k} i c(k+i)}{\sqrt{\sum_{i=0}^{n} \sum_{j=0}^{k} i^2}}
\]

In the formula, \(c\) is a frame of speech parameters, \(k\) is a constant and usually takes 2. Generally, the recognition system based on the PC platform takes the 39-dimensional MFCC features, but considering the limitation of the calculation of the mobile phone platform, after experiments, the final choice is the 12-dimensional MFCC, the 12-dimensional first-order difference MFCC, the 1-dimensional normalized energy, and the 1-dimensional the first-order difference energy and the second-order difference energy have a total of 27-dimensional eigenvalues. Table 1 compares the performance parameters of different dimensional eigenvalues.

| 27-Dimensional | 39 Dimensions |
|----------------|--------------|
| Recognition rate (%) | 92.01 | 97.88 |
| Recognition speed (ms/word) | 5 | 48 |

4. Sugeno integral continuous reading spoken English pronunciation assessment model

Solving the problem of continuous reading assessment requires us to consider the instability of automatic speech recognition systems and the randomness of natural language pronunciation. Obviously, neither of these two factors can be described in an accurate way. As mentioned above, Sugeno integral is an excellent solution for fusing different fuzzy measures [7]. In order to introduce this technology, it is necessary to map the problem domain of continuous reading evaluation to the Sugeno integration space \(\text{sug} \{X, f, \mu\}\). The difficulty of applying Sugeno integral is to find a fuzzy measure according to the power set and satisfy the conditions in the definition.

We can assess the practitioner’s oral and continuous reading skills for each continuous reading group. Based on this simplification, the continuous reading group can be directly mapped to the attribute set \(X\) of the Sugeno integral. This paper defines the attribute set \(X = \{x_1, x_2, \ldots, x_L\}\) according to the length \(L\) of the longest continuous reading group in the entire corpus (6 in our training set). Each element \(x_i \in X\) is just a placeholder, it represents a specific instance of continuous reading classification, and is determined by the corresponding continuous reading group [8].
For example, the attribute sets corresponding to continuous reading groups \( G_i \) and \( G_j \) are 
\[ A = \{ x, y \} \]
and 
\[ B = \{ x, z, y, \} \]
respectively. Although there are elements in the two attribute sets, they correspond to different continuous reading fragments that should and good enough, or different continuous reading types CC and CV. The benefit of using placeholders instead of specific continuations or continuation types can be reflected in the definition of fuzzy measures.

Even native speakers may miss some continuous reading candidates when they pronounce naturally. Based on this, it is unreasonable to determine the value of the score based on the number of continuous pronunciations of the practitioner. The next question is, if the practitioner wants to get a good or excellent score, how many can he/she miss out without pronunciation? Obviously, there is no precise answer [9]. When the length of the continuous reading group is greater than 2, the most common situation is that only 2 continuous readings are pronounced. Based on this phenomenon, the continuous reading score should depend on the actual number of continuous reading pronunciations of the practitioner. The fuzzy measure of "belongs to or better than good" (or goodness):
\[
\mu(A) = \begin{cases} 
0.0, & |A| = 0 \\
\left[ 1 + \left( \frac{|A| - \alpha}{\beta} \right)^2 \right], & 1 \leq |A| \leq L 
\end{cases}
\]  

In the formula, L represents the maximum length of the continuous reading cluster in the training corpus; A represents the sub-attribute set composed of continuous reading placeholders; \(|A|\) represents the potential of the sub-attribute set A. The fuzzy measure of "the degree of excellence":
\[
\mu(A) = \begin{cases} 
e^{-\frac{|A|}{c}}, & 1 \leq |A| \leq c \\
1.0, & c \leq |A| \leq L \text{ or } L < |A| 
\end{cases}
\]  

5. System Test
In the process of performance testing, compare the traditional training strategy and the training strategy designed by the human-computer interaction-based English training system proposed this time, and then test the average deviation of the same batch of sample data and the distribution of training topic exposure [10]. The test model designed in the experiment has a total number of 400 sets in the training question bank, 3,000 participants in the test, and 50 training items. The test results are shown in Table 2 and Table 3.

| Table 2. Performance test results of traditional spoken English training system |
|---------------------------------------------------------------|
| **Comparative adoption** | **Training strategy** |
| Mean deviation | -0.0029 |
| Exposure rate distribution | 7.242 |
| The number of training items whose exposure rate is less than 2% | 2 |
| Number of training items greater than 30% | 14 |
| Training item repetition rate | 0.243 |
Table 3. Performance test results of spoken English training system based on human-computer interaction

| Comparative adoption | Training strategy |
|----------------------|------------------|
| Mean deviation       | -0.0011          |
| Exposure rate distribution | 5.325       |
| The number of training items whose exposure rate is less than 2% | 1 |
| Number of training items greater than 30% | 0 |
| Training item repetition rate | 0.132    |

From the experimental results in Table 2 and Table 3, it can be seen that the indicators of the training strategy proposed by the human-computer interaction-based oral English training system are significantly higher than those of the traditional oral English training system, and have achieved improved oral English. Training system strategy performance requirements.

6. Conclusion

Aiming at the problems existing in the traditional oral English training system, a design of an artificial intelligence-based oral English training system is proposed. The main tasks of the design are: Aiming at the problem of the recording ability of the oral English training system, using artificial intelligence technology to upload each training situation to the database, based on artificial intelligence technology to analyze the students' oral English training situation; for the system power problem, choose different chips to adjust; Through experiments, test the performance of the traditional system and the built system.

References

[1] Tao, Y. Design, development and evaluation of academic oral English curriculum reform. Theory and Practice in Language Studies, vol. 7, pp.427-435, June 2017.
[2] Bing, W. A. N. G. The Teaching Reform in Oral English for Non-English Majors Based on FIF. Sino-US English Teaching, vol. 14, pp.205-210, April 2017.
[3] Xiaojing, F. A. N. Application of Interactive Teaching Approach to Oral English Teaching in Junior Middle Schools. Studies in Literature and Language, vol. 19, pp.94-100, February 2019.
[4] Yao, J., Zhang, F., & Li, H. Research on the Optimization of Oral English Teaching in Junior High School in the Environment of Man-machine Dialogue. Journal of Language Teaching and Research, vol. 10, pp.1047-1054, May 2019.
[5] Chen, S. Design of internet of things online oral English teaching platform based on long-term and short-term memory network. International Journal of Continuing Engineering Education and Life Long Learning, vol. 31, pp. 104-118, January 2021.
[6] Hummel, R., Bruers, J., van der Galiën, O., van der Sanden, W., & van der Heijden, G Outcome measures for oral health based on clinical assessments and claims data: feasibility evaluation in practice. BMC Oral Health, vol. 17, pp.1-18, January 2017.
[7] Yang, Y., & Xiao, L. The construction of college oral English teaching model based on multimodal perspective. Creative Education, vol. 10, pp.423-428, February 2019.
[8] Wang, S., & Liu, C. Task-based flipped classroom in Chinese college EFL teaching: An empirical study in oral English course. International Journal of Contemporary Education, vol. 1, pp. 12-18, January 2018.
[9] Berenguer, J., Gil-Martin, Á., Jarrin, I., Moreno, A., Dominguez, L., Montes, M., ... & González-García, J. J. All-oral direct-acting antiviral therapy against hepatitis C virus (HCV) in human immunodeficiency virus/HCV–coinfected subjects in real-world practice: Madrid coinfection registry findings. Hepatology, vol. 68, pp.32-47, January 2018.
[10] Walqui, A., & Heritage, M. Meaningful Classroom Talk: Supporting English Learners' Oral Language Development. American Educator, vol. 42, pp.18-29, March 2018.