Design and Implementation of Oral Training System Based on Automatic Speech Evaluation

Zhengzhong Pan(✉)

Department of Foreign Languages, Jiangxi Teachers College, Yingtan 335000, Jiangxi, China
juanpan520.good@163.com

Abstract. With the development of human-computer interaction technology, autonomous learning using computer-aided language learning system (call) has become an important learning method for students' English learning. Oral learning only helps you on the road. As for the latter, it depends on yourself, And to this extent, adults are fully capable of summing up their own learning experience and methods, and counseling appears in the form of mutual communication. Aiming at the application of speech recognition technology in oral English training system, according to the characteristics of voiced pronunciation, each syllable must have voiced part and voiced pronunciation energy is large, an improvement of HMM Speech automatic segmentation algorithm is proposed; It is proposed to evaluate the quality of oral pronunciation from the three dimensions of correctness, rhythm and fluency, which better ensures the authority of the evaluation results, and makes the results of machine evaluation and expert evaluation have high likelihood. The research on speech signal and speech recognition is not only an important research content.

Keywords: Automatic language · Speech recognition · Oral training · System design

1 Introduction

As the most widely used language in the world, English has become an indispen.

The function of a few oral pronunciation learning software is relatively single, which can only carry out simple repeated operations such as pronunciation and reading, lack of effective evaluation and feedback, and the training effect is not ideal.

In the past two decades, the rapid development of computer technology, network technology and electronic information technology has provided a good opportunity for English teaching reform. It has become an inevitable development trend for colleges and universities to carry out English teaching reform with the help of computer-assisted language learning (call) system. The mature application of human-computer interaction technology and voice processing technology in call system makes the ways of language learning more diversified, and practitioners can learn in a virtual environment similar to the real environment. In social practice, the use of language includes listening, speaking,

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2022, LNICST 465, pp. 73–84, 2023.
https://doi.org/10.1007/978-3-031-23950-2_9

reading and writing, in which listening and speaking is the foundation of language. Call system has become the best way of oral learning. In the oral learning of call system, the evaluation of pronunciation quality is the most key technology. It uses the computer to evaluate the learners' oral pronunciation by using the voice processing technology to replace the evaluation of experts. Therefore, the evaluation result of computer is required to be accurate and can reach the level of expert evaluation. At present, many speech scoring mechanisms of call system are based on extracting the acoustic features of speech signal. This evaluation mechanism ignores some information hidden by the features of speech signal in other aspects. Therefore, the study of a better voice scoring mechanism to achieve a more accurate evaluation of the learners' oral pronunciation can better help the learners find the shortcomings of their oral pronunciation, better improve the efficiency of the learners' oral English learning and improve the effect of oral English learning.

Oral English is an important tool to realize communication and exchange between people, and oral English level is an important standard to measure conversational ability. With the implementation of the new curriculum reform, China's current English teaching pays more and more attention to the cultivation of communicative ability such as English listening and speaking, and more and more English learners pay more and more attention to their oral ability, Oral English teaching has become one of the important research fields of applied language.

However, as a non-native second language, oral English teaching is weak in English teaching. There are many factors that lead to the poor effect of students' oral English learning, which leads to a large number of primary and secondary school students and even college graduates in China who are still "dumb" who can only write but not speak after more than ten years of English learning. There are many factors, including teachers, There are also student factors. Under the traditional teaching conditions, the only place for communication training is the classroom or English corner. Oral English teaching can only be seen, read and mechanically imitated. Teachers usually only use blackboards and books to provide students with a real language communication environment. Most of the time, teachers are talking and students are listening. Teachers pay attention to the teaching of knowledge and ignore the cultivation of oral communication ability. There are few opportunities for students to speak. These boring "cramming" teaching methods, which are completely divorced from the context, let students speak for the sake of speaking, and do not mobilize students' enthusiasm for active participation in the real sense [1]. Therefore, it is impossible to effectively improve students' verbal communication ability. In addition, a series of tests such as the high school entrance examination and the college entrance examination make students blindly pursue their grades, while the proportion of English oral test scores is much less than that of written test, which makes students mistakenly think that reading and writing are more important than speaking. However, this violates the gradual law of learning language from listening, speaking, to reading and writing. Therefore, "dumb English" has become the primary problem to be solved urgently in China's current English teaching.

2 Related Work

2.1 Basic Principles of Speech Recognition

The general recognition process is divided into: speech signal preprocessing, feature parameter extraction, reference model training, pattern matching (recognition), rule judgment, and output recognition results (as shown in Fig. 1).

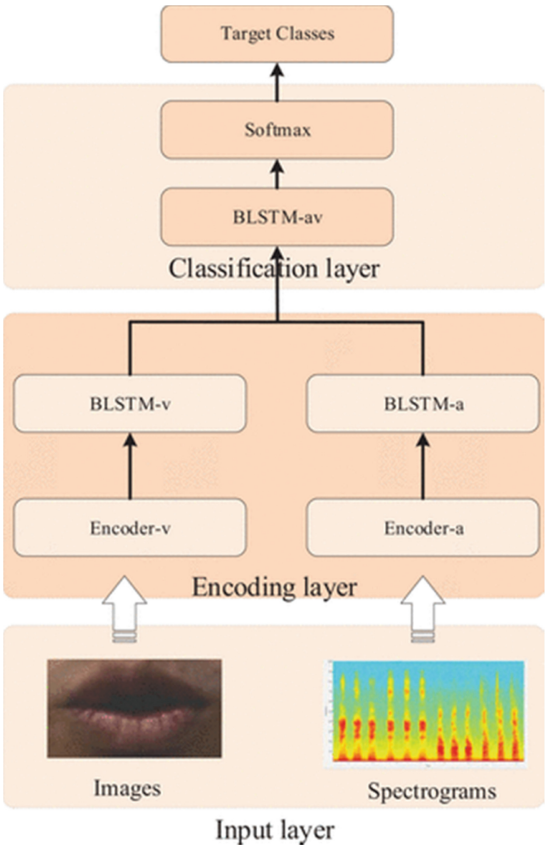

Fig. 1. Schematic diagram of speech recognition

Speech signal processing mainly includes the sampling of speech signal, anti aliasing filtering, the removal of the influence of individual pronunciation differences, and the influence of noise caused by environment and hardware equipment. At the same time, speech signal processing also involves the selection of speech recognition primitives and endpoint detection.

Speech signal feature extraction is mainly used to extract the acoustic parameters that reflect the essential features of speech, such as average energy, average zero crossing rate, formant, etc., for acoustic model processing. The extracted special parameters must meet the following requirements:

The language model includes a grammar network composed of speech recognition commands or a language model composed of statistical methods. For the input signal, language processing finds the word string that can output the signal with the maximum probability according to the acoustics, language model and expert knowledge (such as word formation rules, grammatical rules, semantic rules, etc.)[2].

According to the different applications in practice, the speech recognition system can be divided into: the recognition of specific people and non-specific people, the recognition of isolated words and continuous words, the recognition of small vocabulary and large vocabulary, and the recognition of infinite vocabulary. However, no matter what kind of recognition system, its basic principle and processing method are generally similar, as shown in Fig. 2 below.

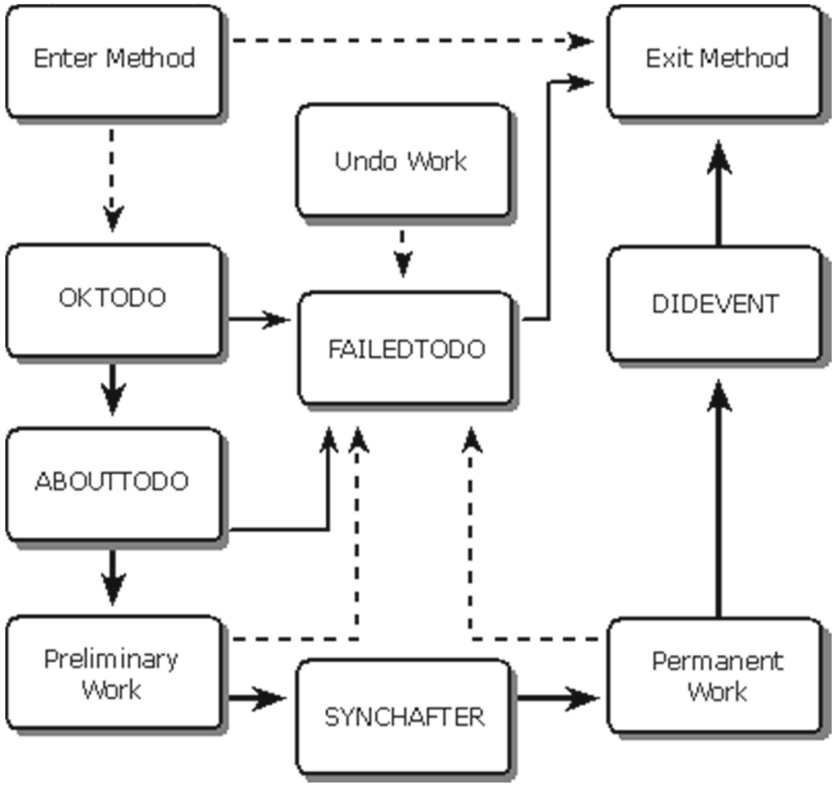

Fig. 2. Forward algorithm probability transition diagram

Pattern matching: match the unknown patterns with the existing model base one by one through the original set standards, so as to obtain the best matching.

Training of acoustic model: after processing the characteristic parameters according to certain standards, the model parameters that can represent the characteristics of these modes are extracted from a large number of known modes to form a template library.

In speech recognition, the main technologies applied to pattern matching and acoustic model training are: dynamic time warping (DTW), hidden Markov model (HMM), artificial neural network (ANN).

HMM is a parameter representation of the time-varying characteristics of language signals. The statistical characteristics of signals are described by two interrelated random processes [3]. The model parameters include HMM topology, state transition probability and a set of random functions describing the statistical characteristics of observation symbols.

The application of ANN in speech recognition system is a research hotspot at present. The network is essentially an adaptive nonlinear dynamic system, which simulates the basic principle of human brain neuron activity, and has the abilities of learning, memory, judgment, association, comparison, reasoning, generalization and so on.

Compared with HMM and ANN, DTW is an earlier pattern matching and model training technology. It successfully solves the problem of unequal time length in the comparison of speech signal characteristic parameter sequence by using dynamic programming method, and obtains good performance in isolated word speech recognition. However, because it is not suitable for continuous speech large vocabulary speech recognition system, it has been replaced by HMM and ANN.

2.2 Analog Signal Digitization

Speech recognition based on Hidden Markov model, on the one hand, uses the hidden state to correspond to the relatively stable pronunciation units in the acoustic layer, and describes the changes of pronunciation through state transition and state residence; On the other hand, the probability calculation model is introduced. The probability density function is used to calculate the output probability of speech feature parameters to HMM model. The recognition result is determined by finding the best state sequence of the function and taking the maximum a posteriori probability as the criterion.

The technical difficulty of speech recognition is how to carry out acoustic analysis, which includes preprocessing and feature extraction. Finally, the feature vector of speech signal in frame is obtained to prepare for HMM model modeling and training. Speech signal is a one-dimensional analog signal, and its amplitude changes continuously with time. Only after digital processing can the signal be analyzed by computer. That is, the digitization of speech signal is the premise of digital processing. The processing process of speech signal digitization includes sampling and quantization. After these two processes, the digital signal with discrete time and amplitude can be obtained.

Automatic speech segmentation algorithms have different classification methods: Based on ASR (automatic speech recognition) algorithm and boundary detection technology.

1) Based on boundary detection algorithm
   This algorithm mainly adopts similar edge detection technology, uses some feature information with high resolution in time domain or frequency domain, and determines the boundary points according to a certain detection mechanism according to

its variation characteristics in time domain. The commonly used characteristic information includes short-time energy, zero crossing rate, etc. The algorithm is simple and fast, but the accuracy is poor[4].

2) ASR algorithm
At present, this kind of algorithm is the most mature and has the best application effect. The package mainly includes dynamic time warping (DTW) algorithm and hidden Markov model (HMM) algorithm. Such time-dependent processing means usually use a finite length window sequence {w (m)} to intercept a speech signal for analysis, and let the window to analyze the signal near a certain time. Its general formula is:

$$Q_n = \sum_{m=-\infty}^{x} T[x(m)]^* w(n-m) \tag{1}$$

where $T[]$ represents a certain operation, and $[x(m)]^*$ is the input signal sequence, as shown in Fig. 3 below.

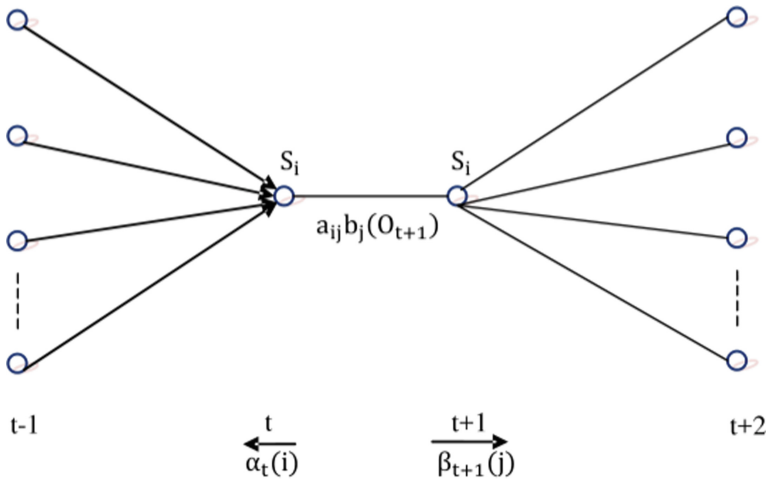

Fig. 3. Algorithm calculation training plan

2.3 Voice Endpoint Detection

HMM training method is one of the important factors affecting the effect of automatic segmentation. This is because based on the statistical method, how to select the appropriate training data and methods is the key to ensure the statistical processing method, which can truly reflect the characteristics of the training data and maximize the characteristics of the data to be processed. Common methods to improve performance include[5]:

1) HMM is trained by speech data after segment segmentation, which is also called restricted HMM.

Training and test results show that this method can greatly improve the segmentation accuracy; It separates the two by using the property that the convoluted components of the signal can be separated after proper homomorphic filtering (as shown in Fig. 4).

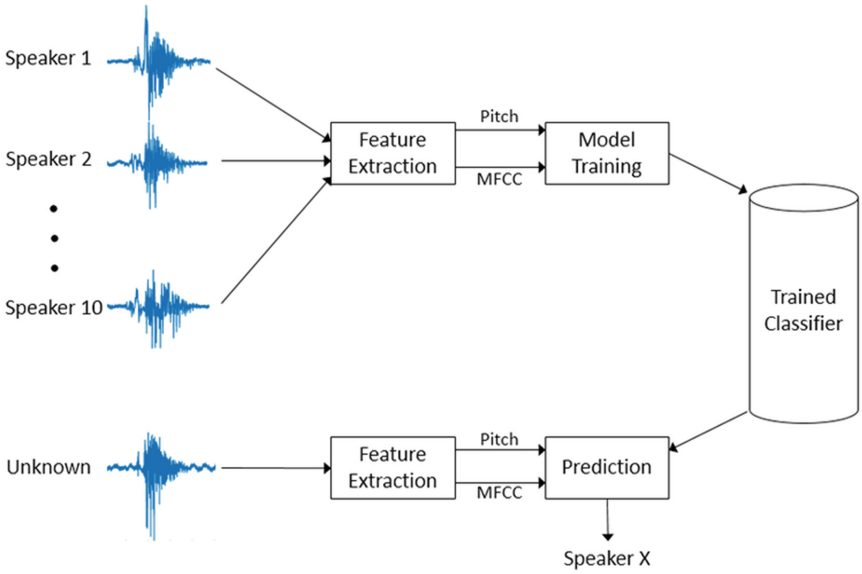

Fig. 4. Speech signal cepstrum processing process

2) In the actual segmentation of speaker speech, using speaker adaptation technology, such as maximum likelihood regression or maximum a posteriori probability method, using a small amount of adaptive data to establish speaker adaptation can also improve the segmentation accuracy.

Correct endpoint detection is the key of all automatic speech recognition systems.

The energy of voiced speech varies significantly with the time. Therefore, the energy of voiced speech is more obvious than that of voiced speech. For signal {x (n), short-time energy is defined as follows:

$$E_n = \sum_{m=-\infty}^{\infty} [X(n) \times w(n-m)]^2 \tag{2}$$

The meaning of short-time average zero crossing rate refers to the number of times that the signal of each frame passes through the zero value. For discrete speech signals,

the short-time average zero crossing rate is essentially the number of symbol changes at the sampling point. As shown in Fig. 5 below, the framework of language signal processing system is shown.

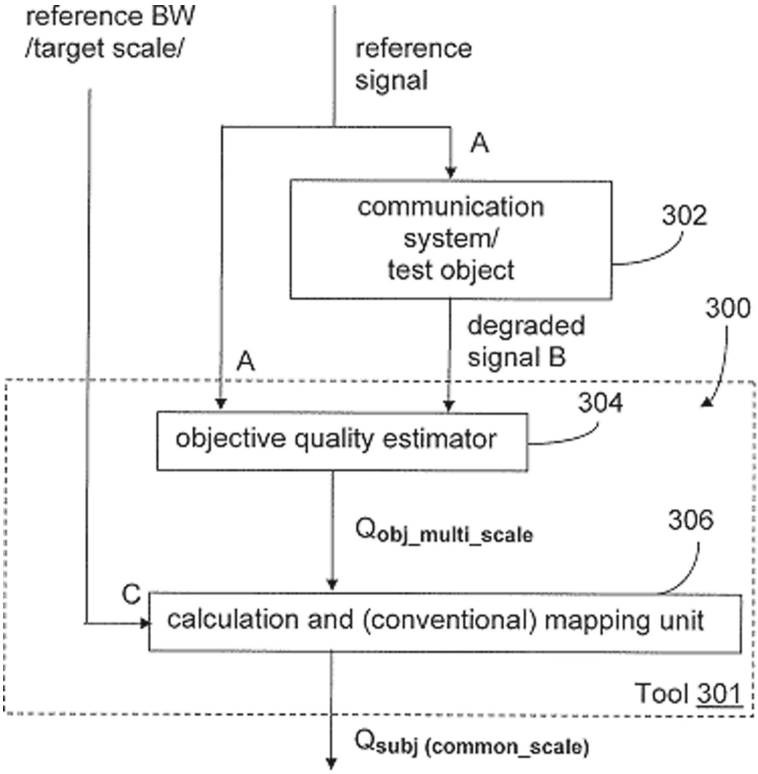

Fig. 5. Framework of speech signal processing system

In order to overcome this, it is found that the signal is the result of convolution between the excitation source and the channel frequency, and the "cepstrum feature" is established.

3 Design and Implementation of Oral training System Based on Automatic Speech Evaluation

Automatic speech evaluation, also known as automatic evaluation of pronunciation quality, refers to the technology that automatically evaluates the pronunciation level, detects errors and gives correction guidance through the machine [6]. It uses the computer to automatically evaluate the standard degree of the voice sent by the user, and its application in computer assisted language learning (call) system, It can help practitioners understand their deficiencies in pronunciation and promote practitioners to improve their

pronunciation [7]. At present, automatic speech evaluation is mainly based on speech recognition, which uses the acoustic gap mapping between the practitioner's pronunciation and the standard pronunciation to obtain the practitioner's pronunciation. The complete speech evaluation research includes the establishment of standard pronunciation model, segmental prosody analysis of pronunciation and manual score mapping training.

In the aspect of oral evaluation based on oral learning, the following problems should be solved:

(1) How to better apply oral evaluation technology to oral learning and play a greater role. For example, analyze the special pronunciation of the practitioner, accurately detect the pronunciation problems of the practitioner, and put forward effective suggestions for improvement.

(2) How to ensure the accuracy of the evaluation results and make the machine evaluation results consistent with the expert evaluation results. This requires that we must conduct more in-depth research and development on all aspects of evaluation technology, including technical details, and effectively improve the evaluation algorithm.

(3) How to eliminate various factors (such as environmental noise, local voice, equipment abnormalities, etc.) that affect the oral evaluation technology, and ensure the accuracy of the evaluation by properly correcting and compensating the evaluation results..The speech recognition module is shown in Fig. 6:

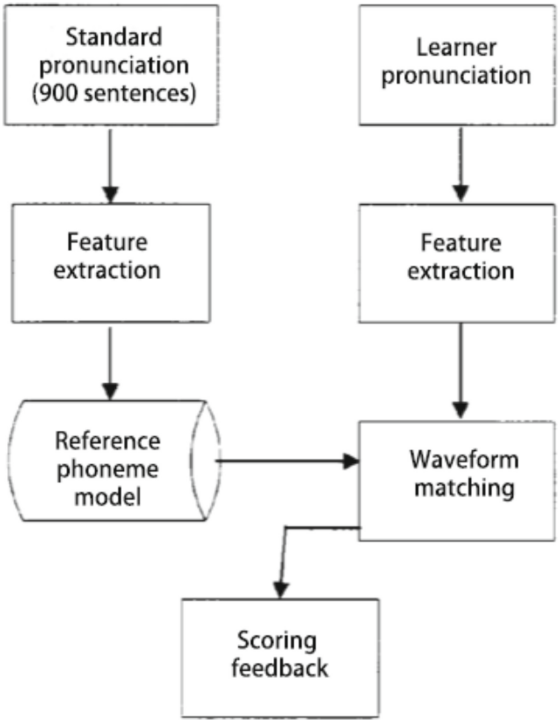

Fig. 6. Schematic diagram of speech recognition module

Basic process of automatic voice evaluation:

1) Pretreatment
   Acoustic features for speech recognition are extracted from speech signals. And analyzing the text to obtain the speech segmentation position, speech acoustic model, recognition network and other information required for speech recognition;
2) Speech recognition
   Pattern matching is carried out according to the language acoustic model to complete speech recognition, and the recognized phonemes and phoneme boundaries are output;
3) Extract scoring features:
   According to the recognition results, imitating the evaluation of experts, combined with the characteristics of text and acoustics, the scoring features that can be quantified to describe the integrity of content, fluency and standard of speech are extracted;
4) Calculated score:
   Use the trained speech model to segment the speech features (usually using vierbi algorithm), then calculate various pronunciation quality test degrees on the segments, and finally have a trained score mapping model to convert the pronunciation quality measure into an intuitive score. Figure 7 below shows speech model vs. speech features.

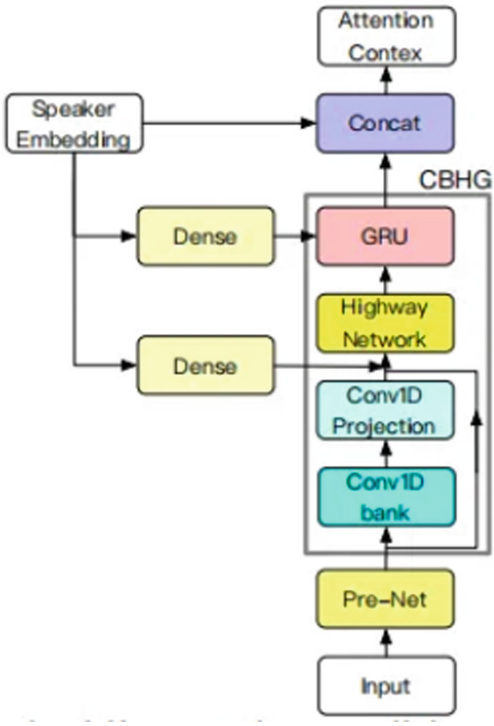

Fig. 7. Speech model vs. speech features

External function: realize visual window. Example sentence Library View: play according to the specified example sentences or learning strategies. Information view: display sentences and phonetics; Display monosyllabic score and correction; Display sentence prosody score and correction, and play correction. User management interface: display user records, study files, analyze fallible phonemes and common errors [8]. The system can recognize English learners with strong Chinese accent.

4 System Process

Speech evaluation is basically similar to speech recognition, but the ultimate purpose is different. Speech recognition is to search and compare the segmented units with the dictionary, and select the one with the greatest probability as the recognition result [9]; The evaluation does not need to give specific recognition results, but only calculate the similarity with the maximum recognition results internally.

Speech evaluation needs multi-dimensional comparison. The main comparison dimensions of the paper are: correctness, prosody and fluency: 1) Correctness: correctness has been introduced in detail in the speech recognition part. After segmentation, only.

The distance can be calculated according to the corresponding characteristic information, which is commonly used as Euclidean distance.

2) Fluency: in this paper, fluency defines the effective pronunciation duration. This definition is for simplicity.

The definition can be defined according to the accuracy and the duration of effective pronunciation unit in the follow-up study.

3) Prosody: prosody is an important measurement dimension. If the voice of the statement is gentle, the interrogative sentence is generally ascending. According to these characteristics, prosody can be characterized by pitch period.

For example: can I take your order?

Because it is difficult to accurately calculate the pitch period, and the calculated period position may not be corresponding, it is not easy to directly convert the distance into fraction [10].

5 Conclusion

The design and implementation of oral training system based on automatic speech evaluation studied in this paper. To evaluate the quality of oral English, the purpose of the system should be to provide the user with the most accurate information about his / her oral performance in order to improve this performance. In addition, it is important that such a system should be easy for users to use and understand. Automatic speech recognition (ASR) technology uses a computer to convert audio files into text. The ASR system uses an algorithm to recognize words or phrases from recorded sound samples. These systems can be used in voice mail transcription, dictation, document retrieval, call center automation and other applications.At the same time, by embedding automatic voice evaluation technology in the system, the system can timely and accurately feed back the results of practice after training, so that students can quickly find their weaknesses, and then supplement learning, so as to improve their oral English level.

Acknowledgements. "Speech Recognition Technology in College Oral English Teaching System" (NO. GJ213205), scientific and technological research project of Jiangxi Provincial Education Department.

References

1. Tripodi, D., Cosi, A., Fulco, D.: The Impact of Sport Training on Oral Health in Athletes. Dentistry J. **9**(5) (2021)
2. Rakhee, P., Claire, R., Gallagher, J.E.: Collaborating for oral health in support of vulnerable older people: co-production of oral health training in care homes. J. Public Health (2021)
3. Stengel, S., Merle, U., Peters-Klimm, F., et al.: Intersectoral online training and exchange during the COVID-19 pandemic in a district. ZFA. Zeitschrift fur Allgemeinmedizin **97**(6), 252–256 (2021)
4. Hsu, W.C., Hsieh, Y.P., Lan, S.J.: Home care aides' attitudes to training on oral health care. PLoS ONE **16**(4), e0249021 (2021)
5. Qiu, X., Chen, Y., Chen, H., et al.: Learning syntactic dense embedding with correlation graph for automatic readability assessment. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)
6. Tripathi, A., Bhosale, S., Kopparapu, S.K.: Automatic Speaker Independent Dysarthric Speech Intelligibility Assessment System. Comput. Speech Lang. (2021)
7. Ahmed, B., Konje, J.C.: Fetal lung maturity assessment: A historic perspective and Non – invasive assessment using an automatic quantitative ultrasound analysis (a potentially useful clinical tool). Europ. J. Obstetrics Gynecol. Reproductive Biol. (2021)
8. Liguori, P., Al-Hossami, E., Orbinato, V., et al.: EVIL: Exploiting Software via Natural Language (2021)
9. Pappagari, R., Cho, J., Joshi, S., et al.: Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios (2021)
10. Phipps, A.: Decolonising the languages curriculum: linguistic justice for linguistic ecologies (2021)