Eng Int Syst (2019) 3: 103–109

Engineering
Intelligent Systems

Emotional Data Mining and DTW Algorithms in English Speech Teaching Recognition

Ning Wang*

Xingtai Medical College, Xingtai, Hebei 054000, China

The main research focus of this paper is the non-specific English spoken speech recognition method under the PC system. At the same time, this paper improves the detection method of speech endpoints, improves the recognition algorithm of speech recognition under PC, and uses the DTW algorithm to match the template, as it is easy to implement. In addition, the endpoint detection method proposed in this paper improves the efficiency of speech recognition. Many English consonants are clear consonants but, when disturbed by noise, they are easily drowned. In the specific person recognition system, the recognition rate is higher. In the non-specific person recognition system, the recognition rate is lower. This paper explores the teaching of English phonetics recognition based on emotional data mining and dynamic time integration algorithm. In the actual application of speech recognition systems, there is a strong demand for real-time software functions, which requires improving the operational efficiency and running time of the system. Such a system can help to improve oral recognition and teaching efficiency in the English classroom.

Keywords: Emotional data mining; Dynamic time integration algorithms; Speech recognition; Operational efficiency

1. INTRODUCTION

Phonetics is the acoustic expression of language. It is the most basic, natural, effective and convenient means by which humans communicate. It is also a kind of support for human thinking (Beuscart et al., 2016). Although the study of human speech has a long history, the original research was limited to the basic theory of acoustics (Ghaderi et al., 2015; Centen et al., 2015). With the popularity of computers and the advent of the information age, speech communication systems which involve digital signal processing, speech synthesis and recognition, are a timely technological development. Automatic Speech Recognition (ASR) refers to the recognition of the sounds of a person's speech and converting them into digital signals for input into a computer for processing (Lin et al., 2014). As a leader in the field of intelligent computer research and the key technology of human-machine voice communication, speech recognition technology has attracted much attention from the scientific communities of various countries (Abdel et al., 2015).

Academic emotion is a non-intellectual factor closely related to the teaching and learning process, which plays an important role in students' learning process. Analyzing students' academic emotions in feedback texts is an important way to discover students' level of learning. Therefore, how to quickly and effectively analyse the emotional aspect of students' feedback texts is a current problem that needs to be resolved (Ullah & Zeb, 2015; Shi et al., 2017). The emotion analysis method in artificial intelligence technology can quickly and effectively analyze emotions in text, and has been applied in many fields (Yuan, 2017; Marcos et al., 2019). However, emotional analysis has a unique application in the field, and the corpus and dictionaries needed for analysis have strong domain relevance (Chen et al., 2018). The visualization of academic emotions can enable teachers and curriculum managers to more intuitively find the emotional changes in students'learning process, and assist teachers to provide intervention and personalized teaching. Identifying

*Email: wangninglunwen123@163.com

Table 1 Performance comparison of 27-dimensional and 39-dimensional eigenvalues.

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

students' emotional state in online learning is essential to predicting students' learning behavior and using it as decision support (Cattivelli et al., 2011; Nishida et al., 2018). In the process of learning spoken English, real-time feedback on the pronunciation of English learning is also very necessary. Immediate feedback on pronunciation can greatly improve the learning efficiency of English learners and will encourage learners to learn and use spoken English.

ISLE (Interactive Spoken Language Education) was introduced in 2010, mainly funded by several universities in Europe, and its main purpose was to provide real-time feedback on voice quality. In order to extract the prosodic features of the music after reading aloud, Audhkhasi N. et al. used the formants to extract the variation of audio frequency. (Yu et al., 2018; Choetkiertikul et al., 2018). However, their research detected only the pause in sound, but did not carry out related research to determine the quality of reading aloud. In order to help learners to use the interactive dialogue system to learn spoken English, the University of Cambridge and MIT have jointly launched the SCILL (Spoken Conversation Interaction in Language Yoga Loamlng) spoken communication language learning project. Another research institute that conducts pronunciation-quality scoring is SRI (Stanford International Research Institute), and their research results have been quite successfully applied in the EduSpeak and WebGrader series of software (Chutani et al., 2012; PATEL et al., 2015). The ISLE (Interactive Spoken Language Education) system is a spoken English pronunciation practice system specially designed for German learners of English. The system has a pronunciation evaluation function that can detect the position of the learner's pronunciation error. However, this system is not ideal for error detection and final feedback (Sun et al., 2015). There is also a spoken English reading practice system for children, called Technology Based Assessment of Language and Literacy. Its main role is to perform speech recognition and automatic evaluation of words. Ordinate in the United States has developed Versant, which is a system used for automatic oral evaluation. It can locate words in a sentence, extract many evaluation features of word speech, and finally fit the pronunciation evaluation score by statistical model (BI et al., 2016).

In both China and abroad, the research on speech evaluation has been carried out for a long time. Although speech evaluation research began later in China than in other countries, it has achieved very promising results. However, much of the software resulting from the research has shortcomings. For example, from the scope of application, they are limited to 121 set sentences which cannot possibly meet the learner's goals and needs. In terms of application, many software programs are used only in business or for personal leisure time learning, although a few English speech evaluation systems are used in middle school classrooms. One objective of this paper is to address the shortcomings of the above-mentioned oral English learning software. studies it.

2. RESEARCH METHOD

2.1 Sound Feature Value Extraction

Generally, when the system is used in a real environment, the recognition performance of the sound feature values is significantly degraded. There are several factors that contribute to this decline in recognition rate:

(1) Additive noise. Additive noise is the sum of the true speech signal and the background noise. Speech signals are often subject to background noise in real-world environments, and background noise is usually additive.

(2) Channel distortion. Speech signals are also affected by phenomena such as speech generation processes, recording processes, and channel distortions that occur during transmission.

(3) Other factors. In addition to the effects of additive noise and channel distortion, the extraction of characteristic parameters is also affected by some other artificial or transient noise.

It can be seen that, given the different environments, the system needs to consider the impact of noise in the English learning environment. Since the system relies only on the traditional MFCC eigenvalues, its immunity to noise is inadequate. Therefore, this paper adopted a variety of parameters to combine anti-noise, introduced parameters that can suppress noise, and introduced more speech through parameter expansion.

We used the first and second order differences of MFCC to suppress the stationary noise and improve the recognition rate. For the differential calculation, the following formula was used:

$$d(k) = \frac{1}{\sqrt{\sum_{i=-n}^{n} i^2}} \sum_{i=-n}^{n} i^* c(k+i) \tag{1}$$

In the above formula, c is a frame speech parameter, and k is a constant, which is usually taken as 2. Typically, based on the identification system on the PC platform, the 39-dimensional MFCC feature is taken. However, considering the limitation of the calculation amount of the platform, the experiment finally selects the 12-dimensional MFCC, the 12-dimensional first-order differential MFCC, the 1-dimensional normalized energy, the 1-dimensional first-order differential energy, and the second-order differential energy. Moreover, a total of 27-dimensional feature values is selected. Table 1 is a comparison of the parameters of the performance of different dimensional eigenvalues.

According to the output probability distribution (B parameter), HMM can be divided into:

Table 2 Comparison of the calculated amount of continuous HMM and semi-continuous HMM.

	The number of Gaussian models required to observe the signal per frame
Continuous HMM	HMM number × number of each HMM state $\times 8 = 31010 \times 3 \times 8 = 744240$

Table 3 Main calculation time ratio.

	PC	Embedded Systems
Gaussian calculation	26.98%	23.87%
Gaussian mixture model calculation	7.76%	11.68%
HMM calculation	25.61%	23.06%

Table 4 Comparison of different feedforward steps

Base	Feed forward step	Speedup ratio	The complexity
4	2	2	2
2	1	1	1

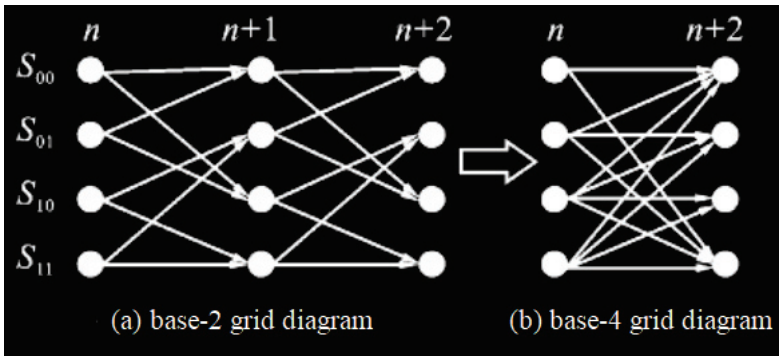

Figure 1 Viterbi grid diagram.

(1) Discrete HMM: The feature parameters are quantized, and the discrete probability distribution is used to approximate the output probability distribution.

(2) Continuous HMM: Multiple (generally 8) state-dependent Gaussian distributions are used to fit the output probability distribution.

(3) Semi-continuous HMM: Multiple (typically 256) state-independent Gaussian distributions are used to fit the output probability distribution.

(4) Comparison of the calculated amount of continuous and semi-continuous models.

According to the experimental results, the recognition calculation amount is mainly concentrated in the calculation of GMM (Gaussian mixture model), and the GMM calculation amount mainly lies in calculating the probability of each Gaussian distribution.

As can be seen from the table above, in the platform, it is more appropriate to use semi-continuous HMM.

2.2 Improved Viterbi Alignment Algorithm

Due to the large amount of computation required by HMM, and given limited computing power, this paper optimized the processing to reduce the computational complexity of the recognition process, while at the same time guaranteeing the recognition rate. The key operations used in the experiment include: Gaussian calculation, HMM calculation (Viterbi path finding) and Gaussian mixture model calculation. The ratio of these three parts to the total running time is shown in Table 3.

Therefore, the key to reducing speech recognition time is to reduce acoustic HMM calculations and Gaussian calculations. This can be achieved by reducing the number of iterations of the path iteration.

The Viterbi decoding algorithm can be depicted with a grid diagram; Figure 1(a) is a 4-state base-2 grid diagram. In order to reduce the number of iteration calculations, the feedforward technique is used to combine the two branch metric grids into a branch metric grid to form the base-4 structure, as shown in Figure 1(b). By merging the two-step branch metric of base-2 into one step, the number of iterations is reduced by half, thereby significantly reducing the decoding delay.

Table 3 lists the decoder speedup, branching complexity, and efficiency for the different feedforward steps in the experiment. The results show that the use of the base-4 structure can achieve better performance.

When selecting the base-4 structure, in order to not reduce the accuracy, it is necessary to study a clipping algorithm of a suitable decoding path to delete the path with the lower probability score. The general judgment method is that the difference between the path and the optimal path is greater than a certain threshold. Here, threshold selection is important. If the threshold is chosen properly, the amount of calculation required can be greatly reduced, and the performance will not drop significantly. The threshold F_c is:

Table 5 Sound processing functions.

Function name	Functional description
wavread	Read wav file
wavplay	Play voice
wavrecord	Recording
wavwrite	Write wav file
sound	Play voice
soundsc	Normalized play voice

$$F_c = C \times P(t) \tag{2}$$

Here, c is a constant and $P(t)$ is the probability score of the optimal path at time t. Since the observation probability of the state in the acoustic model follows the Gaussian distribution and it is evenly distributed after taking the logarithm, the influence on the path score is relatively large. However, the jump probability between states is more concentrated, and the impact on path scores is relatively small. Therefore, when the data of a certain frame is concentrated, there will be a certain path probability value that is very prominent, but it is likely that it is not the global optimal path. In this case, using a fixed threshold will exclude the optimal path, which will seriously affect the accuracy of the recognition. The solution is to obtain the highest and lowest scores in all paths of the current frame and use the split point method to determine the dynamic threshold. The specific method is:

$$P_{\min}(S_t) = \min_{1 \le i \le N} \left(P'(j)\right) \tag{3}$$

$$P_{\max}(S_t) = \max_{1 \le i \le N} \left(P'(j)\right) \tag{4}$$

$$c = 0.618 \left(P_{\max}(S_t) - P_{\min}(S_t)\right) \tag{5}$$

Among them, $P'(j)$ is the probability of path j, and $P_{\min}(S_t) P_{\max}(S_t)$ are the minimum path probability and the maximum path probability of state S at time t, respectively. c is the golden point of the difference between the two, and it is substituted into equation (2) to find the threshold.

2.3 Identification Process Clips

The entire identification process was reviewed:

$$W = \arg\max P(W) \cdot P(W, U) \cdot \prod_{u_i \in U} P(O|u_i) \tag{6}$$

Among them, W is the word string and the language model $P(W)$ represents the probability of the word string W appearing under the model. U is the acoustic model string, the pronunciation dictionary defines the mapping from W to U, O is the observation, and $P(O|u_i)$ is the probability of obtaining the observation under the acoustic model u_i.

For the corpus of the system, the amount of calculation in the identification process is as follows:

Input semaphore: 1s of speech has 200 input data to process (10ms for one frame, 50% overlap).

Language model: 1053 unary models, 3326 binary models, 4240 ternary models.

Acoustic model: 40 monophonic models (uniphone), 10675 triphonic models (trihpone).

If all cases are considered, the amount of calculation in Equation 4–6 will be very large. ($200 \times (40 + 10675) \times$ All string possibilities). Therefore, to speed up the identification process, the clips should be taken for identification.

In the learning of English language pronunciation, speech text has specific characteristics. Therefore, it can be used as a priori knowledge to delete any unnecessary recognition. That is, $\prod_{u_i \in U} P(O|u_i)$ in equation (6) uses $\prod_{u_i \in j} P(O|u_i)$. u_j is the acoustic model string corresponding to the j-th sentence read by the current user.

3. SYSTEM DESIGN

3.1 Function Setting

The system can executethe following functions:

(1) The system can record, extract data and display the waveform of the speech to be recognized, so that the program can be processed in real time.

(2) The system performs feature analysis of the speech to be recognized: Feature analysis is very important for speech recognition and can further study speech recognition algorithms. It is divided into two types: time domain feature analysis and frequency domain feature analysis. Of the two, time domain feature analysis includes: short-term energy analysis, short-term average amplitude analysis, short-term zero-crossing rate analysis, endpoint detection; the frequency domain feature analysis includes a series of MFCC-based feature analyses.

(3) The system can perform dynamic matching and use the DTW algorithm to calculate the Euclidean distance between the characteristic coefficient of the reference speech signal and the actual signal, and obtain the speech matching result, obtaining the score for the learner's pronunciation.

3.2 Data Collection and Function Realization

MATLAB itself provides a certain number of audio processing capabilities. MATLAB provides wav file reading and writing functions and sound card recording and playback functions. These functions can be used for some simple speech signal processing. It mainly includes the six functions shown in Table 5.

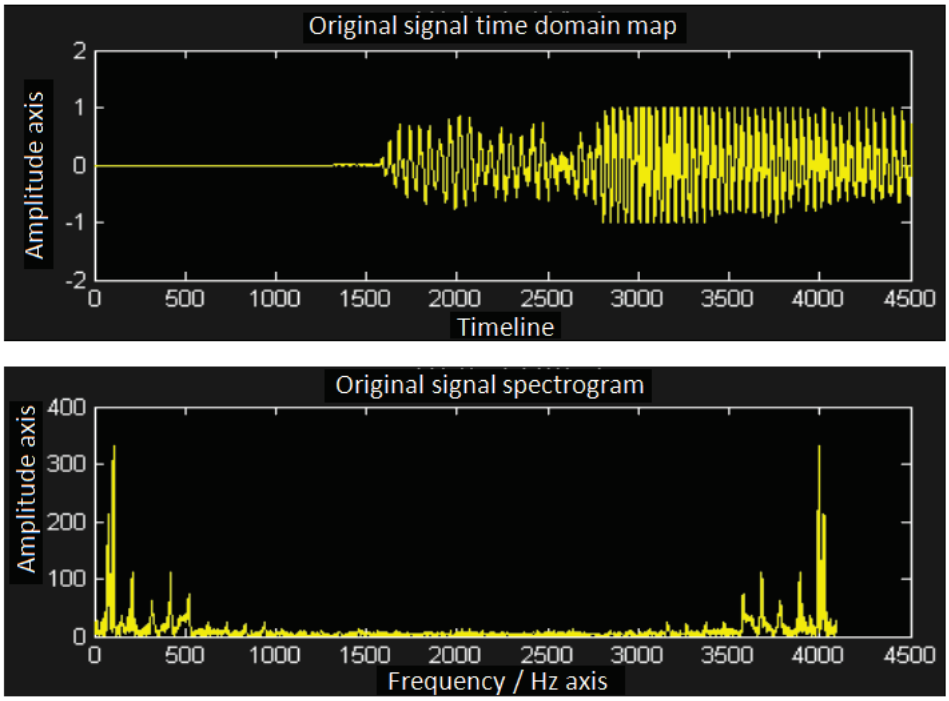

Figure 2 The original signal of the word "za".

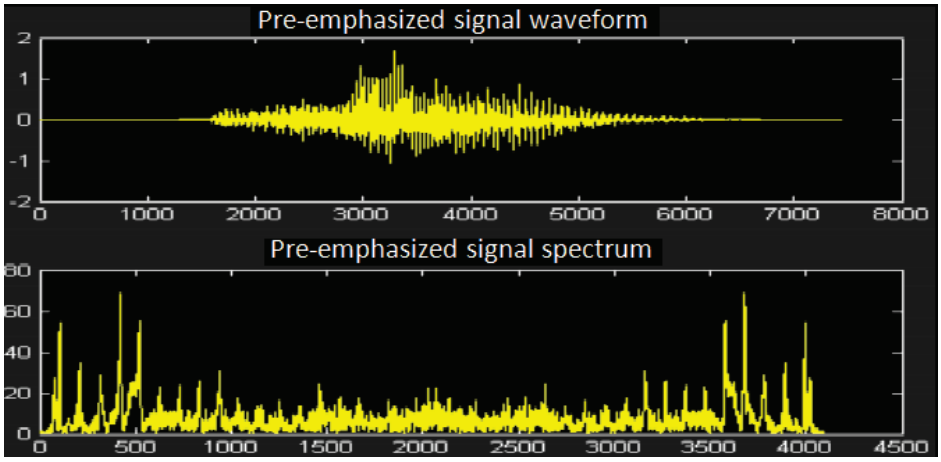

Figure 3 Pre-emphasis of “za.wav”.

When using MATLAB software to process speech signals, the recorded speech can be processed and played back using the sound processing functions shown in Table 5 to quickly verify the data and reliability of the calculations. Voice signal acquisition can be recorded using the wavrecord (n, fs, ch, dtypc) function, or it can be recorded as a wav file using the Windows “recorder” program, and then read using the wavrecord (file) function. The template voice of this system is recorded using the “recorder”. The basic functions of recording are: start recording, stop recording, read the recorded voice data, and read the calculated short-term energy and zero-crossing rate. First, we should adjust the volume setting in the recording feature of the audio device, monitor it with the recorder program provided in the accessory, and adjust the volume to fit. Figure 2 depicts the original speech signal of za.wav.

The Wavrecord function must set the time of the voice when recording the voice. If the time is too short or the user does not speak within the specified time, some or all of the voice data will be missed, which cannot be easily managed. In addition, in a speech recognition system, the detection of a voice command issued by a user should be performed at any time. Moreover, the program must automatically determine whether the current is muted, or the user is speaking, and should save the voice command issued by the user and delete the silent part of the head and tail.

Windowing is used to maintain the short-term stability of the speech signal. The selection of the window is very important, and different windows will make the average result of the energy different. The type used in this system is the Hamming window. Figure 3 below shows the pre-emphasis of the word "za", the waveform after the framed window, and the spectrum.

At runtime, the program will stop and wait for the user’s voice input; the interval between each query is 0.1 seconds. When the endpoint detection is successful, the

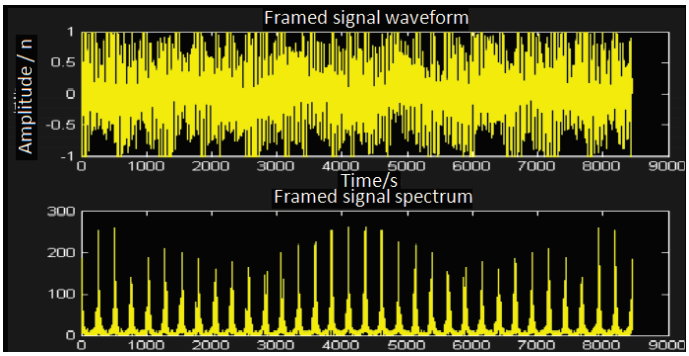

Figure 4 Framing of "za.wav".

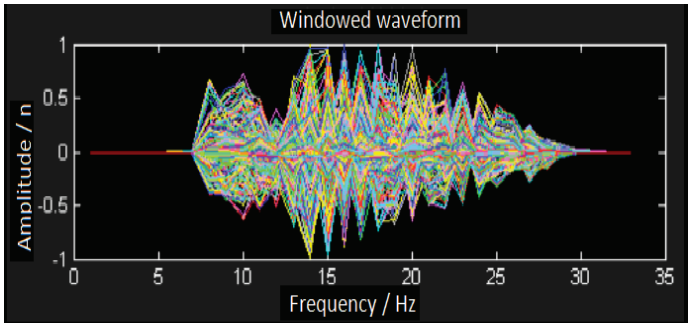

Figure 5 Windowing of "za.wav".

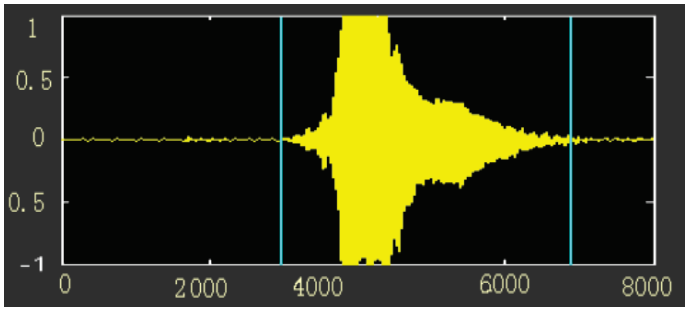

Figure 6 Endpoint detection graph of the word "za.wav".

speech waveform, short-time energy and zero-crossing rate are plotted, as shown in Figure 6.

4. ANALYSIS AND DISCUSSION

With the advancement of science and technology, the rapid development of the economy, and the tremendous improvement of global intelligent terminals, major manufacturers have increased the software development of intelligent platforms, and various products have appeared. However, the general aim of speech recognition is to provide users with the function of interpersonal interaction control. In the field of English language learning, there is still a lack of research and development of applications.

The existing PC-based intelligent English learning software has been able to provide computer-assisted learning technology, so that learners can receive the intelligent function of pronunciation quality score in time. If we want to export this software to the platform, it will be limited by factors such as the speed of operation, the amount of storage space and the bandwidth of the bus. In order to address the hardware and software limitations of embedded systems, this paper examined a set of English learning systems based on continuous speech recognition technology running smoothly on the platform. The system proposed in this paper uses speech recognition technology to effectively evaluate the learner's pronunciation quality and give feedback on the user's pronunciation. The system was developed based on Carnegie Mellon University's SPHINX as the core recognition engine of the whole system, which has the advantages of a large vocabulary and continuous pronunciation recognition.

Due to the large amount of computation of HMM, under the condition of limited computing power, this paper optimized the processing to reduce the computational complexity of the recognition process, but at the same time guaranteeing the recognition rate. The experiments found that key operations include: Gaussian calculation, HMM calculation (Viterbi path finding) and Gaussian mixture model calculation.

The disadvantage of HMM is that the establishment of a statistical model relies on a large speech library. In practice,

this requires a large amount of work, and the amount of storage required for the model and the calculation of the matching calculation (including the output probability calculation of the feature vector) are relatively large. Moreover, the completion of the algorithm usually requires a DSP with a certain SRAM capacity. In view of the limited RAM resources of the device, the language model in the HMM is set to read-only and stored directly in the ROM so that it can be accessed through the I/O operation function of the memory-mapped file.

The binary data within the acoustic model is rearranged and the mapping of the triples is represented in a more efficient manner. Traditional Sphinx uses two large text files to represent the mapping of triples and access them with a hash table. We compress it into a model definition file and access it through a tree structure, which allows us to greatly reduce memory consumption and obtain a faster system start-up speed.

Embedded systems are very fast in integer and Boolean calculations, but weak in floating-point operations (e.g. ARM does not have a processor that handles floating-point arithmetic). In the code of the recognition process, the amount of floating-point operations should be reduced as much as possible by integer operations. At the same time, since ARM is a 32-bit access architecture supported by 16 general-purpose registers, it is the fastest to read data at 32 bits each time. Therefore, we also try to avoid unaligned access, which is mainly achieved by manually expanding some loop code in the code.

5. CONCLUSION

At present, in more and more environments, oral communication is being conducted in English. The use of intelligent portable terminals offering intelligent English learning systems that are independent of time, location and teacher resources will provide users with better and faster e-learning tools. The main research focus of this paper is the non-specific spoken English speech recognition method under the PC system. This paper improves the detection method of speech endpoints, and the recognition algorithm of speech recognition under PC, and used the DTW algorithm to match the template. In addition, the endpoint detection method proposed in this paper improves speech recognition efficiency. The speech recognition system has high requirements for endpoint detection. Moreover, the end detection method used in this paper can accurately detect the starting point of speech and improve the speech recognition rate. Through the verification study, we demonstrate that the speech recognition algorithm of this paper achieves a high recognition rate.

REFERENCES

1. Abdel A., Xu Y., Josang A. (2015). A normal-distribution based rating aggregation method for generating product reputations. Web Intelligence, 13, pp. 43–51.
2. Beuscart J., Mellet K., Trespeuch M. (2016). Reactivity without legitimacy? Online consumer reviews in the restaurant industry. Journal of Cultural Economy, 9, pp. 458–475.
3. BI Anqi, DONG Aimei, and WANG Shitong, (2016). A dynamic data stream clustering algorithm based on probability and exemplar. Journal of Computer Research and Development, 53(5), pp. 1029–1042.
4. Chen, S., Moh'd, A., Nourashrafeddin, S., & Milios, E. (2018). Active High-Recall Information Retrieval from Domain-Specific Text Corpora based on Query Documents. In Proceedings of the ACM Symposium on Document Engineering 2018 (p. 12). ACM.
5. Centeno R., Hermoso R., Fasli M. (2015). On the inaccuracy of numerical ratings: Dealing with biased opinions in social networks. Information Systems Frontiers, 17, pp. 809–825.
6. Choetkiertikul, M., Dam, H. K., Tran, T., Pham, T., & Ghose, A. (2018). Predicting components for issue reports using deep learning with information retrieval. In Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings (pp. 244–245). ACM.
7. Cattivelli F S, Sayed A H., (2011). Distributed detection over adaptive networks using diffusion adaptation. IEEE Transactions on Signal Processing, 59(5), pp. 1917–1932.
8. Chutani A, Sethi S P., (2012). Optimal advertising and pricing in a dynamic durable goods supply chain. J of Optimization Theory and Applications, 154(2), pp. 615–643.
9. Ghaderi A., Mohammadpour H., Ginn H. (2015). High impedance fault detection in distribution network using time-frequency based algorithm. IEEE Trans Power Deliv, 30 (3), pp. 1260–1268.
10. Lin C., Chen H., Wu Y. (2014). Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection. Expert Systems with Applications, 41 (15), pp. 6611–6621.
11. Marcos-Pablos, S., & García-Peñalvo, F. J. (2019). Information retrieval methodology for aiding scientific database search. Soft Computing, 1–10.
12. Nishida, K., Saito, I., Otsuka, A., Asano, H., & Tomita, J. (2018). Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 647–656). ACM.
13. PATEL V M, NGUYEN H V, VIDAL R., (2015). Latent space sparse and low-rank subspace clustering. IEEE Journal of Selected Topics in Signal Processing, 9(4), pp. 691–701.
14. Sun Li-juan, Chen Xiao-dong, Han Chong, Guo Jian, (2015). New Fuzzy-Clustering Algorithm for Data Stream. Journal of Electronics and Information, 37(7), pp. 1620–1625.
15. Shi, F., Chen, L., Han, J., & Childs, P. (2017). A data-driven text mining and semantic network analysis for design information retrieval. Journal of Mechanical Design, 139(11), 111402.
16. Ullah A. Zeb W. (2015). The impact of emotions on the helpfulness of movie reviews. Journal of Applied Research and Technology, 13, pp. 359–363.
17. Yuan, K. (2017). Multi-dimensional Formula Feature Modeling for Mathematical Information Retrieval. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1381–1381). ACM.
18. Yu, J., Lu, Y., Qin, Z., Zhang, W., Liu, Y., Tan, J., & Guo, L. (2018). Modeling text with graph convolutional network for cross-modal information retrieval. In Pacific Rim Conference on