Semi-Continuous Hidden Markov Model Optimized Pronunciation Pattern Recognition in English Education

Ruolin Shi[1], Xuebing Zhang[2], Rui Chen [3], and Xinyu Zhang[4]

[1] Cangzhou Normal University, Cangzhou, Hebei 061000, China, zlj@caztc.edu.cn
[2] Cangzhou Normal University, Cangzhou, Hebei 061000, China, srl891216@163.com
[3] Cangzhou Normal University, Cangzhou, Hebei 061000, China, chenruihebei@126.com
[4] Cangzhou Normal University, Cangzhou, Hebei 061000, China, bjwuzg@126.com

Corresponding author: Xuebing Zhang, srl891216@163.com

Abstract. The purpose of this paper was to study the improved recognition algorithm and use the DTW in speech recognition to automatically recognize the learner's English pronunciation, realize basic recognition and scoring, and provide English learners with more feasible pronunciation information feedback. This paper focused on the theoretical issues of signal preprocessing, frame decomposition, feature selection and calculation, feature matching, etc. in the process of speech recognition, so as to theoretically clarify the speech recognition ideas and key technical issues and provide a theoretical basis for the future research and development of speech recognition technology in hardware or software. The research shows that the proposed algorithm has certain effects and can provide theoretical reference for subsequent related research.

Keywords: semi-continuous hidden Markov; spoken English; matching technique; spoken recognition
DOI: https://doi.org/10.14733/cadaps.2020.S1.1-11

1 INTRODUCTION

In China, many college graduates can't even carry out very simple English communication. Even if many people can communicate easily in English, they are not standard in English pronunciation and have a heavy local spoken language. Therefore, the study of spoken English has gradually become a major obstacle to Chinese English learning. In the process of learning spoken English, real-time feedback on the pronunciation of English learning is also very necessary. Correct pronunciation feedback can greatly improve the learning efficiency of English learners and improve the enthusiasm of learners to learn spoken English and the initiative of spoken English learning [1].

In foreign countries, the Stanford Research Institute in the United States designed the voice interactive learning system, which is referred to as VILTS. This system can dynamically generate segments of speech in real time through a continuous speech recognition system (STI) and can score people's voices. The principle of the system is based on HMM, and the speech is scored from HHM similarity, speech rate, speech period and other aspects. At that time, the system is very representative in the direction of speech scoring [2]. Fragkou P used the HMM posterior probability and HMM similarity to remove the speech rate assessment parameters and found that the factor-based posterior probability score is close to the similarity of the artificial score. Later, He Y improved the system to improve the accuracy of the score [3]. He Y evaluated the speech at the level of the phonetic factor, so that it can accurately judge the specific location of the tester's speech. The purpose is to evaluate the similarity between the words spoken by the tester and the target speech [4]. Carrie E used automatic speech recognition to study non-native speech and finds many shortcomings in CAPT (Computer Aided Pronunciation Training), such as lack of understanding of speech recognition system and improper setting of courses. Finally, the following conclusions were drawn. If the appropriate method is used on the basis of the previous one and the pronunciation of the pronunciation is incorrectly detected, the system can get a good scoring effect [5]. ISLE (Interactive Spoken Language Education) was born in 2017, 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, Boutsen F R et al. used the formants to extract. However, their research only detected the pause of sound, but did not carry out related research on the quality estimation of reading aloud [6]. 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 [7]. Another research institution that conducts pronunciation quality scoring is SRI (Stanford International Research Institute), and their research results have been reasonably applied in the EduSpeak and WebGrader series of software [8]. The ISLE (Interactive Spoken Language Education) system is a spoken English pronunciation practice system specially designed for German English learners. The system has a pronunciation evaluation function to detect the position of the learner's pronunciation error. However, this system is not ideal for error detection and final feedback [9]. 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 [8]. Ordinate in the United States develops Versant, which is a system 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 [10].

Domestically, Darabseh A evaluated the parameters of pronunciation and nonlinear attenuation. Finally, the experimental results show the similarity of each sentence in speech and the logarithmic probability of syllables [11]. At the National Acoustics Academic Conference, China's Dong Bin proposed to evaluate the attributes of speech. Finally, this algorithm can evaluate speech phonemes very well, which is better than the posterior probability of HMM model in accuracy. Xia C [12] et al. evaluated the fluency of English reading, but they did not conduct in-depth research on indicators such as the quality of spoken language and the completeness of reading. Klubička F et al. [13] did some research work on English reading and retelling testers respectively, which solved the problem that the acoustic model of the target pronunciation and probability space was confusing, and finally improved the performance of pronunciation quality evaluation. Elaraby M et al. [14] adopted a new algorithm called GMM-UBM model. This model was finally applied to the pronunciation quality assessment, and finally got a good effect on the speech score, which is close to the expert score. Nzimande E et al. [15] adopted a linguistic approach in the logarithmic posterior probability, which made the machine score very close to the manual score, and the correlation reached 0.795. In terms of accuracy, it is 9% better than the previous algorithm.

From the research status at home and abroad, we can see that the research on speech evaluation has been carried out for a long time at home and abroad. The domestic start of speech evaluation is later than that of foreign countries, but it has also achieved good research results. In these research results, many software also has shortcomings. For example, from the scope of application, they are limited to some fixed English sentences of 121 sentences. The learner's own learning goals and content are not within this scope, and it is difficult to meet the learner's learning requirements. From the application field, many software are only used for business or personal leisure time learning, but there are few English speech evaluation systems used in middle school classrooms. In order to solve the shortcomings of the above-mentioned oral English learning software, this paper 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 the diversity of the use environment determines that the system needs to consider the noise impact in the environment. Since the system relies only on the traditional MFCC eigenvalues, its noise immunity is not the best. 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. The differential calculation used the following formula:

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

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.

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

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

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

(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.

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.

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

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

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

2.2 Improved Viterbi alignment algorithm

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 guaranteed the recognition rate. The key operations found 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.

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

Table 3: Main calculation time ratio.

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 represented by a grid diagram, and 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.

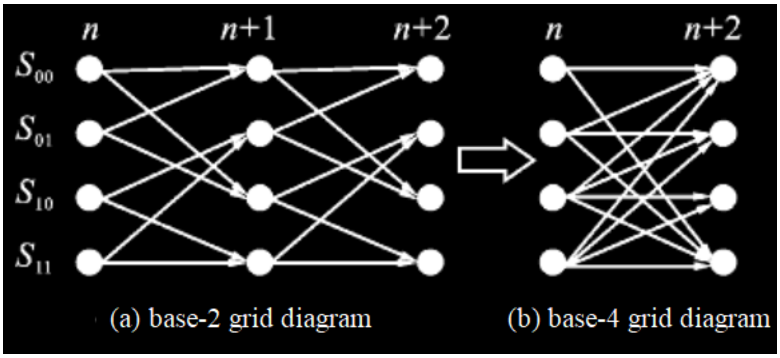

Figure 1: Viterbi grid diagram.

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

Table 4: Comparison of different feedforward steps

While selecting the base-4 structure, in order not to reduce the accuracy, it is necessary to study a clipping algorithm of a suitable decoding path to delete a path with a 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 key. If the threshold is chosen properly, the amount of calculation can be greatly reduced, and the performance will not drop significantly. The threshold F_c is:

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

Among them, 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 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 get the highest and lowest scores in all paths of the current frame and use the split point method to take 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 $P_{\min}(S_t)$ and $P_{\max}(S_t)$ 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.

Speech text is known for specific characters in spoken English pronunciation learning. Therefore, it can be used as a priori knowledge to delete the case of 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 complete the 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 on 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. Among them, time domain feature analysis includes: short-term energy analysis, short-term average amplitude analysis, short-term zero-crossing rate analysis, endpoint detection, and frequency domain feature analysis includes a series of MFCC based feature analysis.

(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, and give the score of the learner's pronunciation.

3.2 Data collection and function realization

MATLAB itself provides a certain amount of audio processing capabilities. MATLAB provides wav file reading and writing functions and sound card recording and playback functions. These functions can be used to implement some simple speech signal processing. It mainly includes 6 functions in the table 5.

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

Table 5: Sound processing functions.

When using MATLAB software to process speech signals, the recorded speech can be processed and played back using the sound processing functions in Table 5 to quickly verify the reliability and data 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 properties of the audio device, monitor it with the recorder program provided in the accessory, and adjust the volume to fit. The following picture shows the original speech signal of za.wav.

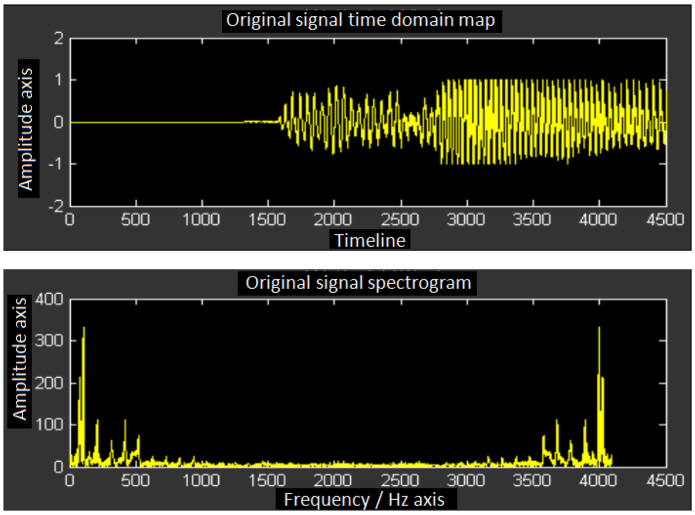

Figure 2: The original signal of the word "za."

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 is very inconvenient to operate. 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 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. The figures 3-5 showed the pre-emphasis of the word "za", the waveform after the framed window and the spectrum.

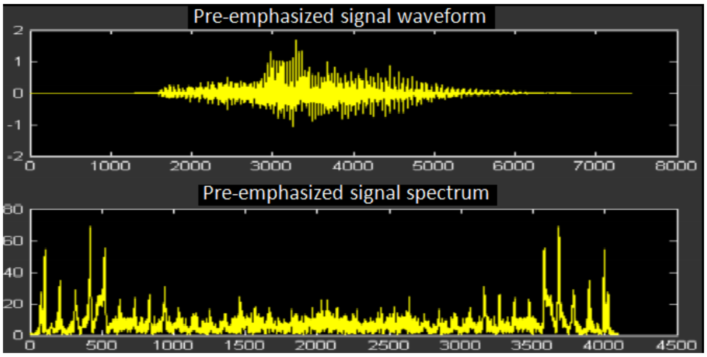

Figure 3: Pre-emphasis of “za.wav.”

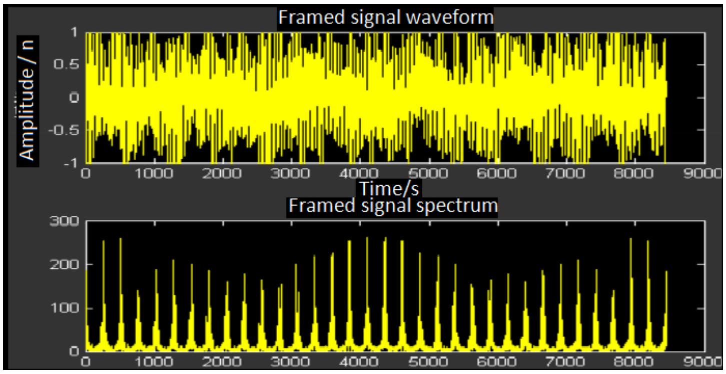

Figure 4: Framing of "za.wav."

At runtime, the program will stop and wait for the user's voice input, and the interval between each query is 0.1 seconds. When the endpoint detection is successful, the speech waveform, short-time energy and zero-crossing rate are plotted, as shown in Figure 6.

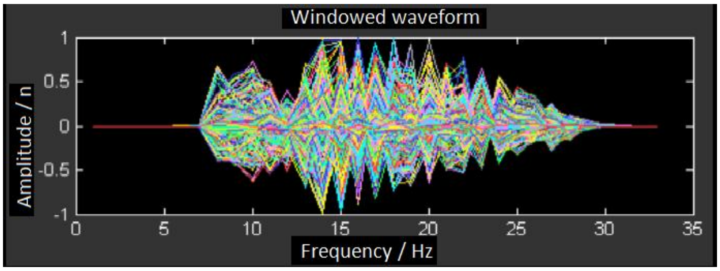

Figure 5: Windowing of "za.wav."

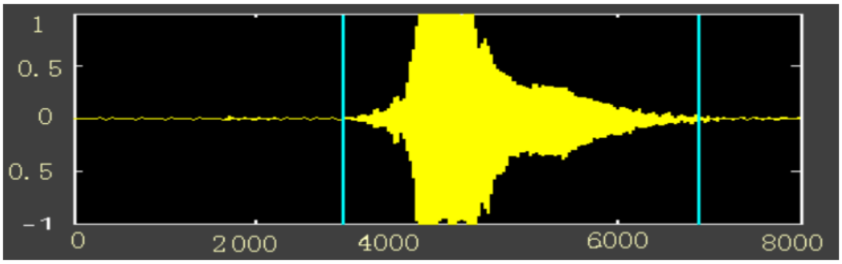

Figure 6: Endpoint detection graph of the word "za.wav."

4 ANALYSIS AND DISCUSSION

The data acquisition interface is shown in Figure 7. With the advancement of science and technology and the rapid development of the economy and the tremendous improvement of the performance of global intelligent terminals, major manufacturers have increased the software development of intelligent platforms, and various voices have appeared. However, the general speech recognition is to provide users with the function of interpersonal interaction control, and the application research and development for English learning needs are still lacking.

The existing PC-based intelligent English learning software has been able to provide computer-assisted learning technology, so that learners can get the intelligent function of pronunciation quality score in time. If we want to port such 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. Aiming at the hardware and software limitations of embedded systems, this paper studied a set of English learning system based on continuous speech recognition technology running smoothly on the platform. The system uses the speech recognition technology to effectively evaluate the learner's pronunciation quality and feedback the user's pronunciation information. The system development is based on Carnegie Mellon University's SPHINX as the core recognition engine of the whole system, which has advantages in 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 guaranteed the recognition rate. The experiments found that key operations include: Gaussian calculation, HMM calculation (Viterbi path finding) and Gaussian mixture model calculation.

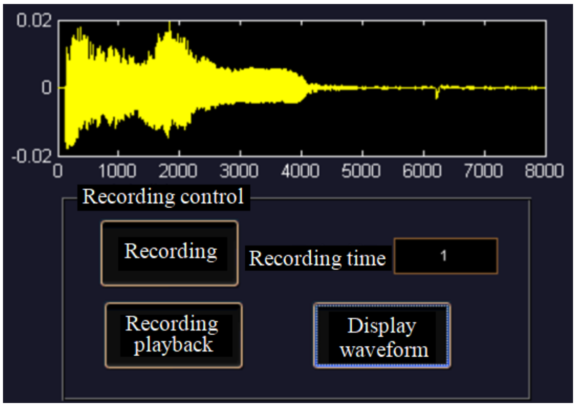

Figure7: Data acquisition interface.

The disadvantage of HMM is that the establishment of a statistical model relies on a large speech library. This occupies a large amount of work in actual 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 capacity of SRAM. In view of the scarcity of 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 $I/0$ 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 get faster system startup speed.

Embedded systems are very fast in integer and Boolean calculations, but weak in floating-point operations (like 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 conducted in English. The use of intelligent portable terminals to provide users with 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 object of the thesis is the non-specific English spoken speech recognition method under the PC system. At the same time, this paper improved the detection method of speech endpoints, improved the recognition algorithm of speech recognition under PC, and used DTW algorithm to match the template. In addition, the endpoint detection method proposed in this paper improved 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 can know that the speech recognition algorithm of this paper achieves a high recognition rate.

6 ORCID

Ruolin Shi, https://orcid.org/0000-0002-2781-2979
Xuebing Zhang, https://orcid.org/0000-0003-0021-9267
Rui Chen, https://orcid.org/0000-0001-7555-3521
Xinyu Zhang, https://orcid.org/0000-0003-1775-5218

REFERENCES

[1] Jingyu, C.; Ying, L.: Research on English pronunciation training based on intelligent speech recognition, International Journal of Speech Technology, 2018, https://doi.org/10.1007/s10772-018-9523-8

[2] Saon, G.; Kurata, G.; Sercu, T.; et al: English conversational telephone speech recognition by humans and machines, 2017, https://doi.org/10.21437/Interspeech.2017-405

[3] Fragkou, P.: Applying named entity recognition and co-reference resolution for segmenting English texts, Progress in Artificial Intelligence, 6(3), 2017, 1-22, https://doi.org/10.1007/s13748-017-0127-3

[4] He, Y.; Kwon, M. Y.; Legge, G. E.: Common constraints limit Korean and English character recognition in peripheral vision, Journal of Vision, 18(1), 2018, https://doi.org/10.1167/18.1.5
[5] Carrie, E.; Mckenzie, R.: American or British? L2 speakers' recognition and evaluations of accent features in English, Journal of Multilingual & Multicultural Development, 2018, 39:págs. 313-328, https://doi.org/10.1080/01434632.2017.1389946
[6] Boutsen, F. R.; Dvorak, J. D.; Deweber, D. D.: Prosody and spoken word recognition in early and late Spanish–English bilingual individuals, Journal of Speech Language and Hearing Research, 60(3), 2017, 712, https://doi.org/10.1044/2016_JSLHR-H-15-0274
[7] Sidgi, L. F. S.; Shaari, A. J.: The effect of automatic speech recognition EyeSpeak software on Iraqi students' English pronunciation: A Pilot Study, Advances in Language and Literary Studies, 8(2), 2017, 48, https://doi.org/10.7575/aiac.alls.v.8n.2p.48
[8] Zou, S.: Designing and practice of a college English teaching platform based on artificial intelligence, Journal of Computational and Theoretical Nanoscience, 2017, https://doi.org/info:doi/10.1166/jctn.2017.6133
[9] Fragkou, P.: Applying named entity recognition and co-reference resolution for segmenting English texts, Progress in Artificial Intelligence, 6(3), 2017, 1-22, https://doi.org/10.1007/s13748-017-0127-3
[10] Dat, N. D.; Phu, V. N.; Tran, V. T. N.; et al: STING algorithm used english sentiment classification in a parallel environment, International Journal of Pattern Recognition and Artificial Intelligence, 31(07), 2017, 25, https://doi.org/10.1142/S0218001417500215
[11] Darabseh, A.; Siami-Namini, S.; Siami Namin, A.: Continuous authentications using frequent english terms, Applied Artificial Intelligence, 2018, 1-35, https://doi.org/10.1080/08839514.2018.1447535
[12] Xia, C.; Jing, Z.: English translation of classical Chinese poetry, Orbis Litterarum, 73(4), 2018, https://doi.org/10.1111/oli.12184
[13] Klubička, F.; Toral, A.; Sánchez-Cartagena, V. M.: Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian, Machine Translation, 1, 2018, 1-21, https://doi.org/10.1007/s10590-018-9214-x
[14] Elaraby, M.; Tawfik, A. Y.; Khaled, M. et al: Gender aware spoken language translation applied to English-Arabic, 2018, https://doi.org/10.1109/ICNLSP.2018.8374387
[15] Nzimande, E.: Domestication versus foreignisation in the English translation of Mthembu's UMamazane, Language Matters, 49(1), 2018, 149-163, https://doi.org/10.1080/10228195.2018.1444080