Design of Students' Spoken English Pronunciation Training System Based on Computer VB Platform

https://doi.org/10.3991/ijet.v14i06.10154

Yanju Jin
University of Science and Technology Liaoning, Anshan, China
miranda_jyj@163.com

Abstract—Spoken English communication is most commonly used in the international communication. However, the accuracy of spoken English pronunciation is the key factor to restrict English learners in China. For the current situation that spoken English proficiency is generally low in China, this paper aims to design a spoken English pronunciation training system that will provide guidance and help for English learners' spoken pronunciation. The Visual Basic platform is used in the design of the system. This paper first conducts an in-depth study on the related theories of voice recognition, discusses the correction algorithm of voice scoring and pronunciation, and puts forward more practical and convenient AP-based scoring method, providing full theoretical support for the design of the system. Then through the function analysis and design of the spoken English pronunciation training system, this paper realizes the system design of scoring and correcting errors of English spoken pronunciation based on the VB platform. The system boasts the basic functions, including English phonetic symbols and word pronunciation to follow, real-time voice evaluation, and pronunciation error correction. According to the test, the similarity of the system with the experts is over 90% in scoring and its efficiency of pronunciation error correction reaches 80%, which plays a certain role in improving spoken English of English learners.

Keywords—Spoken English; VB platform; voice recognition; AP-based voice scoring; system design

1 Introduction

With the development of the global integration process and the implementation of "the Belt and Road" and "going global" strategies, people in China have more and more opportunities to communicate directly with foreigners in spoken English. Thus, there are more and more people learning English [1]. However, spoken English pronunciation is the shortcoming of most Chinese English learners, which is the key factor to prevent Chinese English learners from participating in English communication.

Lots of translation software can provide basic auxiliary pronunciation functions, such as word pronunciation, for the inaccurate spoken English pronunciation. However, they cannot provide English voice evaluation and pronunciation error correction to different English learners [2]. At present, voice recognition technology has got a mature development and has been applied in mobile devices and computers. The development of voice recognition technology and progress of pronunciation feedback technology have made it possible to conduct voice recognition and pronunciation error correction.

The spoken English proficiency is generally low in China because there is a big difference in pronunciation between English and Chinese, teachers who can accurately pronounce are insufficient in domestic schools and the environment for spoken English practice is far from enough [3]. To improve the spoken English pronunciation level, relevant scholars have started to carry out studies on computer-assisted pronunciation learning based on voice recognition technology. However, their applications are so small that they are not suitable for promotion. Besides, there also exist some problems in voice evaluation and error correction method [4].

This paper conducts an in-depth study on the related theories of voice recognition technology and puts forward AP-based scoring algorithm for the inaccurate scoring of voice recognition technology in spoken language learning. In addition, combined with Visual Basic platform, it provides a set of feasible technical solution for spoken English training. Detailed analysis and design have been conducted on the system's input and output modules, scoring module, voice feedback module, and user interface, etc. Finally, the design of the system for spoken English pronunciation training is realized and has achieved good results in the actual test. The design of this system can not only provide new assisting methods for spoken English learners, but play a certain role in promoting voice recognition scoring and error correction technology in spoken English evaluation.

2 VB Computer Technology and System Language Recognition Algorithm

2.1 VB computer technology

Visual Basic (VB) is a universal object-based programming language developed by Microsoft. It is a structured, modular, object-oriented visual programming language that includes event-driven development environment [5].

Visual Basic originated from the BASIC programming language. VB has a graphical user interface (GUI) and a rapid application development (RAD) system that makes it easy to connect a database with DAO, RDO, and ADO, or to create ActiveX controls for efficient generation of type-safety and object-oriented application programs. Programmers can easily use the components provided by VB to quickly set up an application program. When a traditional program designs language programming, it usually designs the interface of the application program (such as appearance and location of the interface) by writing a program, where the actual effect of the interface cannot be seen. However, in Visual Basic 6.0, Object-Oriented Programming is used to encapsulate programs and data as objects, each of which is visual [6]. In the interface design, developers can use the Visual Basic 6.0 toolbox to "draw" different types of objects, such as window, menu, and command buttons on the screen, as well as set the properties for each object. The only thing that developers need to do is to code for the objects in the event process, so the efficiency of the programming can be enormously improved.

This paper chooses VB computer platform as human-computer interface of spoken English pronunciation training system. By embedding the relevant voice recognition algorithms and evaluation feedback methods, this platform will serve as a simple, practical and efficient platform for spoken English evaluation after the completion of the relevant programming processes.

2.2 Design of language recognition algorithm

Introduction of recognition algorithm: The recognition of the pronunciation of English learners is the prerequisite of the intelligent pronunciation training system, and only after the voice can be recognized can further evaluation and feedback be conducted. The essence of voice recognition is machine's recognition of natural voice of humans [7].

At present, there are three common voice recognition algorithms and their characteristics are shown in Table 1.

Table 1. Comparison of various voice recognition methods

Voice recognition method	Characteristic
Method based on channel model	The acoustic model and voice knowledge are too complicated, no reach practical stage
Pattern matching method	Has reached the practical stage, commonly used technology has a dynamic time regulation (DTW), hidden Markov (HMM) and vector quantization (VQ)
Artificial neural network method	Too complex to achieve, is still in the experimental research stage

As shown in the table, artificial neural network and sound channel model have not yet reached the practical stage, so the pattern matching technology is used for automatic voice recognition. The basic flow chart of the technology is shown in Figure 1.

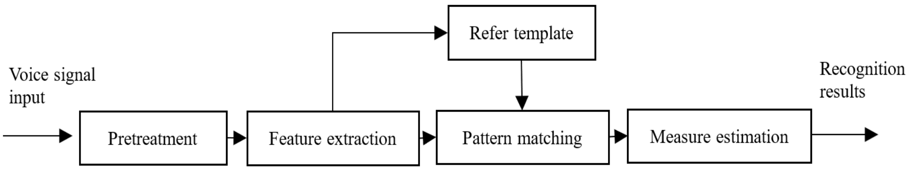


Fig. 1. The flowchart based on pattern matching recognition

T=he main processes of voice recognition include pretreatment, feature extraction, pattern matching and measure estimation. The input voice signal can be achieved the recognition results through these processes.

Petreatment of voice signal: Voice signal is an analog signal with unstable amplitude, and computer recognition can be carried out only after its digital conversion.

According to the Nyquist theorem, the signal is sampled with the sampling rate being greater than or equal to twice of the signal bandwidth. The frequency of voice signal is generally 300 ~ 3400Hz. This pape

r uses 8000Hz sampling rate and 16-bit digitalizing bits.

Pre-emphasis of voice signal: The universal law of energy signal loss: Each time the frequency of the signal increases by twice, the amplitude of the power spectrum will drop by 6dB. Therefore, according to the corresponding proportion, the signal is emphasized by the first-order high-pass filter, and the transfer function is:

$$H(z) = 1 - a * z^{-1} \tag{1}$$

In the form of time domain, the pre-emphasized signal $S_2(n)$ is:

$$S_2(n) = s(n) - a * s(n-1) \tag{2}$$

Where, *a* is the pre-emphasis coefficient. Generally, the value is close to 1, and the value in the system is 0.97.

Framing windowing processing of voice signal: Next, the voice signal is framed and windowed by using the time of 256 sampling points as a frame. Then the method of continuous framing is adopted to process the voice signal. In order to reduce the inconsistency of the boundary signal, a window function is usually multiplied. The commonly used window functions include a rectangular window function and a Hamming window function, and their expressions are as follows:

Rectangular window function:

$$w(n) = \begin{cases} 1, 0 \leq n \leq (N-1) \\ 0, \quad n = \text{other value} \end{cases} \tag{3}$$

Hamming window function:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n/(N-1)\right], 0 \leq n \leq (N-1) \\ \qquad 0, \quad n = \text{other value} \end{cases} \tag{4}$$

The suitable window function is a parameter of the short-term feature of the signal. The selection of the window function mainly considers two aspects, namely window shape and window length. The system uses a rectangular window function in the time domain endpoint, and uses a Hamming window function in the short time-frequency transform processing.

Voice signal endpoint detection: Endpoint detection is to find out the starting and ending point of each segment of the voice signal element by use of digital erection techniques and related algorithms. Voice endpoint is the key to the accuracy of signal detection [8]. This paper adopts the endpoint detection method combining short-time energy with short-term zero-crossing rate. The method is simple with small calculation and high reliability.

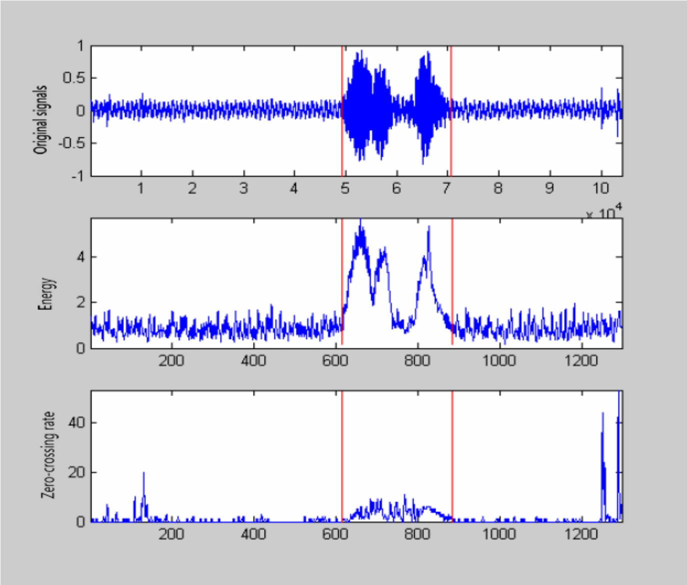


Fig. 2. The voice signal endpoint detection

It can be seen from Figure 2—diagram of voice endpoint detection that combination of short-time energy with short-term zero-crossing rate playa a very good role in determining the starting and ending point of each segment of the voice signal, making good preparation for further processing of signal [9].

Feature extraction of voice signal: The feature processing of voice signal is to calculate and extract the key parameters that reflect the features of the signal. The features of voice signal are effectively described by a small amount of parameters to facilitate the subsequent processing. Feature extraction is conducted on signal. MFCC feature extraction is a more commonly used feature extraction method, which can not only reflect the features of voice, but better reduce noise. MFCC feature parameters extraction steps are shown in Figure 3.

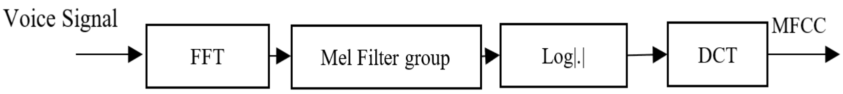


Fig. 3. MFCC feature parameter extraction process diagram

MFCC converts the frequency domain to the Mel frequency domain to better smooth the voice spectrum and reduce the effect of harmonic wave. As a result, the input voice will not be affected by tone or volume, which is suitable for spoken English pronunciation recognition.

Pattern matching of voice signal: In the voice recognition evaluation, the similarity between voice to be evaluated and the reference standard voice is reflected by comparing the difference between their feature parameters. However, due to the difference in length of pronunciation and speech, the two cannot directly match each other Thus, a matching discriminated method is used to carry out pattern matching for the feature parameters. Dynamic time warping (DTW) is a nonlinear normalization method that combines time warping with distance-gap calculation. And the distance between the vectors expresses the matching similarity between the template to be tested and the reference template eigenvector. The larger the distance is, the smaller the matching similarity is. The distance between the eigenvectors T (n) and R (m) is usually expressed by Euclidean distance:

$$d[T(n), R(m)] = \sum_{i=1}^{p} (t_i - r_i)^2 \tag{5}$$

Where, t_i and r_i represent the ith dimension eigenvector of T (n) and R (m) respectively, and p is the order of the eigenvector. DTW needs time warping function m=w (n). The time axis n of the template to be tested is non-linearly mapped to the time axis m of the reference template to obtain the minimum distance of the whole matching.

3 Pronunciation Feedback Evaluation Methods and Technologies

The scoring method based on the adaptive parameter (AP) in this paper provides scoring feedbacks for spoken English pronunciation. AP-based scoring algorithm is shown formula (6):

$$\text{score} = \frac{100}{1+x(d)^y} \tag{6}$$

Where, x and y are adaptive parameters whose value is uncertain and they can conduct adaptive changes according to the computer or hardware settings [10]. AP-based scoring algorithm is shown in Figure 3.

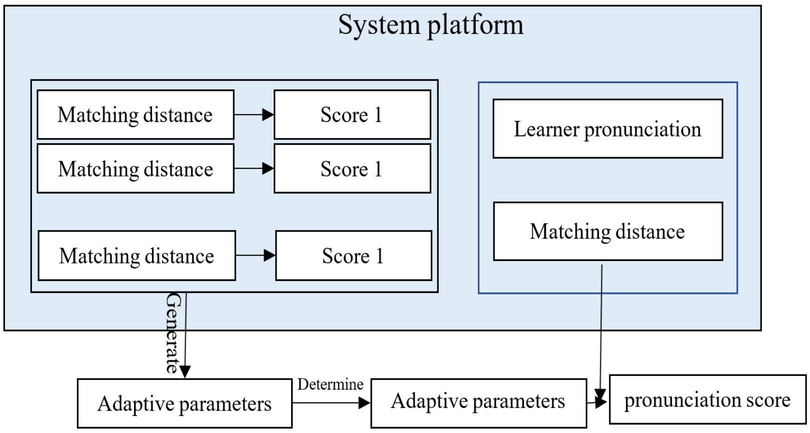


Fig. 4. Schematic diagram of AP-based scoring algorithm

The spoken English scoring system generates the adaptive parameters through a separate scoring parameter generation module before scoring. Learners pronounce for different voices, and experts score the learner's pronunciation by experience. The scores of experts correspond to the MFCC frame matching distance one by one. And the set of MFCC frames is $A = \{d_1, d_2, d_3, \ldots d_i \ldots d_n\}$, the set of scores of experts is $B = \{s_1, s_2, s_3 \ldots s_i \ldots s_n\}$. The corresponding relationship of n pairs of data is shown in formula (6) (Saastamoinen, J et al., 2005):

$$\begin{cases} s_1 = \frac{100}{1+x(d_1)^y} \\ s_2 = \frac{100}{1+x(d_2)^y} \\ \ldots\ldots \\ s_n = \frac{100}{1+x(d_n)^y} \end{cases} \tag{7}$$

The estimated value of parameters x and y is calculated by three samples for scoring. Such systematic scoring and expert scoring have a high degree of similarity, which makes the system more accurate and valuable for scoring of spoken pronunciation [11].

4 Design and Realization of Students' Spoken English Pronunciation Training System

The main function of the system is to realize the learning and training of pronunciation of English phonetic symbols and words in the form of multimedia output and to provide evaluative feedback on spoken pronunciation of English learners while guiding the learners to continuously train and improve spoken pronunciation level. The system boasts the basic functions, including pronunciation demonstration, pronunciation to follow, pronunciation contrast, pronunciation scoring and pronunciation result image output [12].

4.1 System function module analysis and design

Design of I/O module and mode settings of system input and output are conducted. The system chooses audio record to record the voice signal and uses audio track to output the corresponding voice signal. The final audio format of the system is as follows. Sampling frequency: 8000Hz; sampling sound channel: mono; sampling bit: 16 bits. The system demonstrates how to pronounce for all all phonetic symbols.

Scoring module design uses AP pronunciation scoring technology, including scoring parameters generation and pronunciation scoring. The voice signal is first processed according to the order of Section 2.2 and then the pronunciation is scored according to the score adaptive parameters x and y.

The main content of the feedback module design is to graphically describe the contrast between the standard reference pronunciation and the learner's practice of speaking pronunciation so as to facilitate the qualitative reflection of the differences in pronunciation between the two.

To facilitate spoken English learners to use the system, a simple boundary system interface is very important [13]. Figure 5 shows the structure of the user interface.

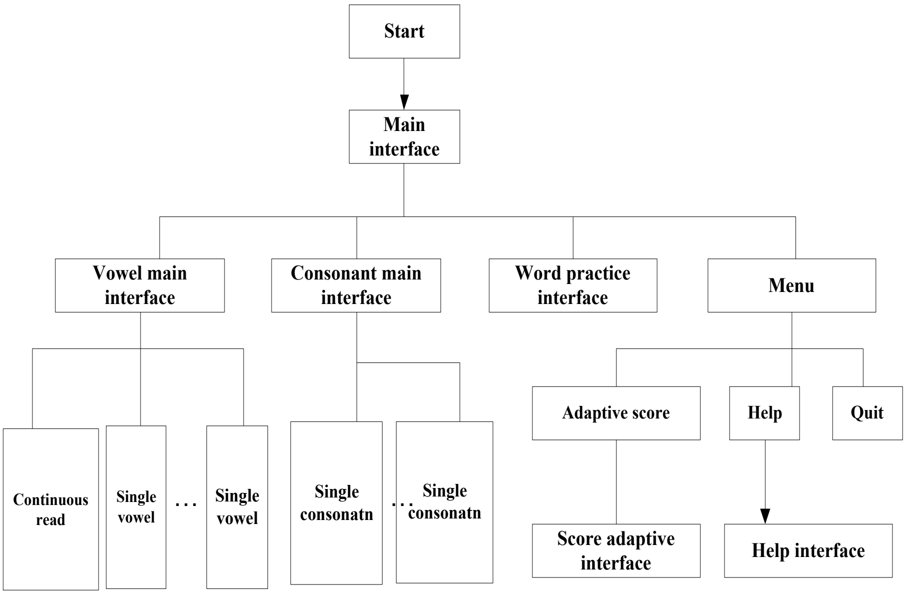


Fig. 5. User interface diagram

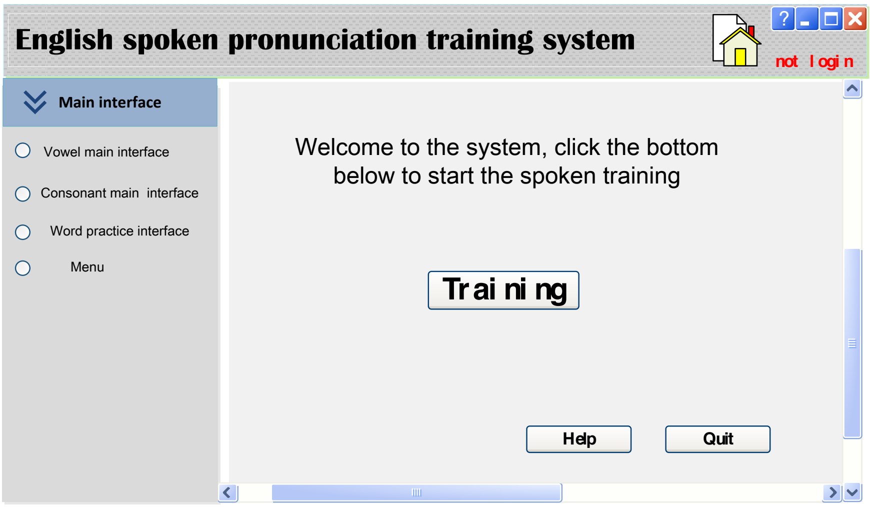
We can enter the main interface of the system after starting the interface. According to the learners' different learning contents, special practices of "vowel phonetic", "consonant phonetic" and "pronunciation of words" are provided on the main interface [14]. At the same time, the system provides self-adaptive scoring and system help interface to assist learners in completing the pronunciation training and scoring of words better.

4.2 Realization of system

Operating environment of system: Hardware: i3 and above processor, 2G memory, 500G hard disk. Operating system and software: window XP, window 7 and other window operating system, Visual Basic 6.0 or above, Office and other common office software.

Audio input device: general anti-noise microphone.

Realization of system: Related controls are set by VB language and are embed with voice recognition technology and AP adaptive scoring function, and corresponding expert voice pronunciation is stored for comparison with input voice. After completing the programming of the corresponding controls, the system functions and interface design are achieved relying on the user interface structure diagram.


English spoken pronunciation training system
not login
Main interface
○ Vowel main interface
○ Consonant main interface
○ Word practice interface
○ Menu
Training
Help
Quit

Welcome to the system, click the bottom below to start the spoken training

Fig. 6. The main interface of the system

Unregistered users can click on the training button to start training and are familiar with the system's operating functions and voice training effects after entering the main interface. Users who are satisfied with the system can save training records and set personalized voice pronunciation training.

As shown in Figure 7, it is a set of training samples of words for the learners to conduct pronunciation comparison and pronunciation scoring after entering the pronunciation. For the words with lower pronunciation scores, users can read after the pronunciation to improve the accuracy of pronunciation by training.

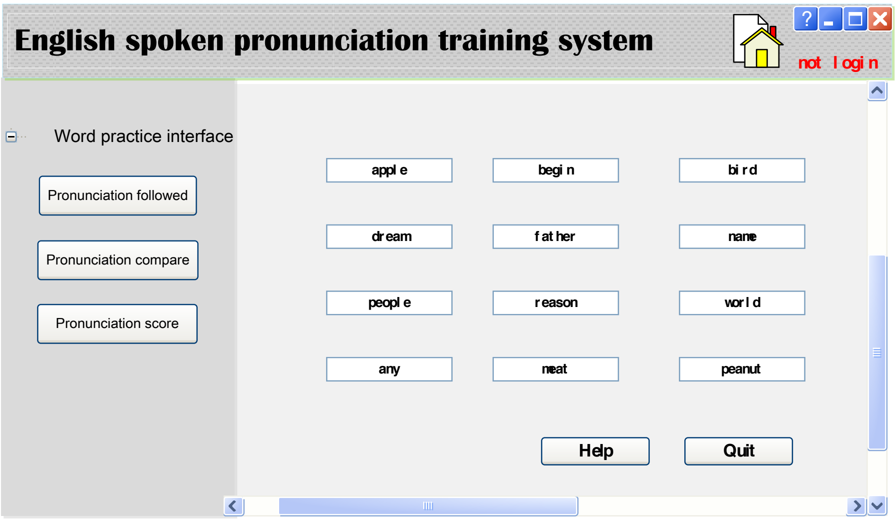


Fig. 7. The word practice interface

4.3 System test of spoken English pronunciation training system

20 vowels, 24 consonants, and 12 words are taken as a test sample. As shown in table 2, the results of the system are tested from three aspects: pronunciation success rate, scoring accuracy and error correction efficiency.

Table 2. The test results

Score type	Vowel	Consonant	Word
Pronunciation success rate	100%	91.6%	100%
Score accuracy	95.52%	75.26%	91.68%
Error correction efficiency	85%	81%	75%

It can be seen from the table that the system has good test results that can meet learners' needs of spoken English pronunciation training.

5 Conclusion

The accuracy of spoken English pronunciation is an urgent need for Chinese English learners to improve for the lack of special and systematic training and guidance of spoken pronunciation. With the improvement of internationalization, spoken English has attracted more and more attention. Through a deep analysis of voice recognition technology and phonetic pronunciation scoring algorithm combined with the computer VB platform, this paper achieves the design of students' spoken English pronunciation training system based on the VB platform by fully considering the system's functional requirements. The system design has the following conclusions and significance:

- The AP-based voice-scoring algorithm in this paper has high accuracy, which is is more suitable for spoken English pronunciation.
- The system provides a convenient spoken English training platform that will significantly improve the accuracy of spoken English pronunciation.

6 Acknowledgement

2016 Higher Education Teaching Reform Project Affiliated to Liaoning Province (2016) 23; 2017 Teaching Reform Project Affiliated to University of Science and Technology Liaoning (XJGZD201703)

7 References

[1] Passera, S., Kankaanranta, A., Louhiala-Salminen, L. (2017). Diagrams in contracts: fostering understanding in global business communication. IEEE Transactions on Professional Communication, 99, 1-29. https://doi.org/10.1109/TPC.2017.2656678

[2] Hamada, H., Nakatsu, R. (1988). Evaluation of English pronunciation based on the static and dynamic spectral characteristics of words spoken by Japanese. Journal of the Acoustical Society of America, 84(S1), S113. https://doi.org/10.1121/1.2025688

[3] Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal of the Acoustical Society of America, 107(2), 989-999. https://doi.org/10.1121/1.428279

[4] Kim, S.J. (2012). A study on the oral health behavior of international students in Korean universities - with a focus on Chinese students -. Journal of Korean society of Dental Hygiene, 12(1), 17-26. https://doi.org/10.13065/jksdh.2012.12.1.017

[5] Shan F.H., Zhao L.Q., Yang F. (2018). A novel semantic matching method for chatbots based on convolutional neural network and attention mechanism, Revue d'Intelligence Artificielle, 32(S1), 103-114. https://doi.org/10.3166/ria.32.s1.103-114

[6] Wang S., Hu Y.Z. (2018). Binocular visual positioning under inhomogeneous, transforming and fluctuating media, Traitement du Signal, 35(3-4), 253-276. https://doi.org/10.3166/ts.35.253-276

[7] Seng D.W., Zhang H.Q., Fang XJ., Zhang X.F., Chen J. (2018). An improved fingerprint image matching and multi-view fingerprint recognition algorithm, Traitement du Signal, 35(3-4), 341-354. https://doi.org/10.3166/ts.35.341-354

[8] Miao Y.S., Wu H.R., Zhu H.J., Song Y.L. (2018). Localization accuracy of farmland wireless sensor network localization algorithm based on received signal strength indicator, Ingénierie des Systèmes d'Information, 23(5), 69-80. https://doi.org/10.3166/isi.23.5.69-80

[9] Srikanth B., Kumar H., Rao K.U.M. (2018). A robust approach for WSN localization for underground coal mine monitoring using improved RSSI technique, Mathematical Modelling of Engineering Problems, 5(3), 225-231. https://doi.org/10.18280/mmep.050314

[10] Tsang, W., Salgo, I., Gajjar, M., Freed, B., Weinert, L., & Nathan, S. (2013). Comparison of sts mortality score with a three-parameter echocardiographic score in the prediction

of aortic stenosis mortality. Journal of the American College of Cardiology, 61(10), E881-E881. https://doi.org/10.1016/S0735-1097(13)60881-5

[11] Saastamoinen, J., Karpov, E., Hautamäki, V., Fränti, P. (2005). Accuracy of MFCC-based speaker recognition in series 60 device. Eurasip Journal on Advances in Signal Processing, 2005(17), 1-12. https://doi.org/10.1155/ASP.2005.2816

[12] Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P., O'Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Communication, 55(2), 237-251. https://doi.org/10.1016/j.specom.2012.08.007

[13] Suraya, H., Yusof, N.A., Ijab, M.T., Leong, O., Hong, J.J. (2008). The design and development of an open and flexible e-training system for the creation of learning organizations. Veterinary Record, 127(1), 357-78.

[14] Simonsen, D., Popovic, M.B., Spaich, E.G., Andersen, O.K. (2017). Design and test of a microsoft kinect-based system for delivering adaptive visual feedback to stroke patients during training of upper limb movement. Medical & Biological Engineering & Computing, 362, 1-9. https://doi.org/10.1007/s11517-017-1640-z

8 Author

Yanju Jin currently works at the School of Business Administration, University of Science and Technology Liaoning. Yuran does research in Business Administration, Supply chain management and Transport Economics. Their current project is 'Business Model Innovation' Email id**:** miranda_jyj@163.com

Article submitted 2019-01-15. Resubmitted 2019-02-28. Final acceptance 2019-02-28. Final version published as submitted by the authors.