Training & Evaluation System of Intelligent Oral Phonics Based on Speech Recognition Technology

https://doi.org/10.3991/ijet.v13i04.8469

Zhaoxia Yin
Beihua University, Jilin, China
nancyin2007@163.com

Abstract—The majority of Chinese people are still bound up in "dumb" English today. The English learning software is ubiquitous in our lives, but most of them merely focus on English literacy without pronunciation evaluation and corrective feedback enabled. How to improve the oral English learning efficiency and quality has more and more become a hotspot of people's common concern. The maturity of Speech Recognition Technology (SRT) has kicked off a new mode of oral English learning, which allows the learning software enable pronunciation evaluation and feedback function. This paper probes into the speech signal extraction and pattern matching in SRT. For the sake of ease learning, the Android mobile phone platform is introduced for learner whereby to propose a rating method based on Adaptive Parameters (AP), create a mouth shape correction, and design intelligent English oral phonics training and evaluation system. This paper describes the system implementation process in detail and gives a test demonstration for the system's availability.

Keywords—speech recognition technology (SRT); oral English; pronunciation evaluation; pronunciation feedback

1 Introduction

In the context of globalization, English, as one of the internationalized communicative languages, has appealed to more and more Chinese learners. While the oral English is a major means for achieving English oral communication and measuring conversational competence. Therefore, how to improve the oral English learning efficiency and quality has become a focus of frontier studies. With the rapid development and the popularization of computer technology, many English learning software relying on computer platforms emerges greatly. however, most of them focus on the improvement of English literacy. Individual software has simple repeat after functions, but lack good design and application in the terms of oral pronunciation evaluation and corrective feedback [1], which has also become a bottleneck in spoken English learning intelligence.

In recent years, SRT has gradually become matured, which makes the communication between human and machine come true, and allows the learning software enable the evaluation and feedback mechanisms. More and more people are therefore prone

to use SRT for spoken English articulation training studies. In particular, the rapid development of Internet technology and the massive popularization of smartphones have made it possible for SRT's application in the mobile and intelligent spoken English training systems [2]. A new model for oral English learning is initiated.

Speech recognition is the key to success in the design of the intelligent spoken English pronunciation training and evaluation system. This paper, based on the analysis of SRT elementary theory and key algorithm, and relying on the Android mobile phone as a platform, designs a training and evaluation system for intelligent spoken English pronunciation, thus to realize the intelligence of spoken English articulation, which not only facilitates more English learners but also plays an important role in improving the oral English learning efficiency.

2 Designing System Speech Recognition Algorithm

2.1 SRT

The SRT can be simply interpreted as a machine-recognizable input converted from human language by a certain technology so that the human-machine communication has triumphed [3]. With the wide spreading of the Internet, SRT has started to be applied in various areas. The advancement of computer technology applies more and more people to studying computer aided speech learning based on speech recognition.

Speech recognition is the key to success in the design of intelligent spoken English pronunciation training and evaluation system. The three types of speech recognition modes in SRT are shown in the Table 1 [4].

Table 1. Comparison of various voice recognition methods

Speech recognition method	Features
Based on the channel model of the method	Although it started early, it did not reach the practical stage due to the complexity of biology model and phonetic knowledge
Pattern matching method	More mature, commonly used techniques are Dynamic Time Warping (DTW), Hidden Markov (HMM) and Vectorization (VQ), etc.
Artificial neural network approach	To achieve more complicated, is currently in the experimental phase

Currently, the speech recognition systems mostly adopt SRT more mature based on the pattern matching. The flow chart of speech recognition system based on pattern matching is shown in Fig. 1, including the speech signal preprocessing, feature extraction and pattern matching, etc. [5]. This paper focuses on profound exploration and analysis on speech signal extraction and the pattern matching.

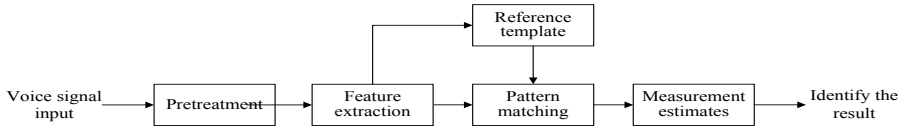

Fig. 1. Pattern recognition based on the identification of flow chart

2.2 Extraction of speech signal feature

It calculates and extracts a few of parameters that can reflect the signal feature to give an effective description for language signals. The three commonly used feature parameters and their characteristics are described as shown in the Table 2 [6].

Table 2. Each characteristic parameter and its characteristic

Characteristic Parameters	Parameter characteristics
LPCC	Small amount of calculation, easy to implement, the effect of general
MFCC	Good recognition performance and anti-noise ability
ASCC	Recognition of the band in the voice of good results

MFCC () has good speech recognition and anti-noise performances. Since the change in spoken English pronunciation and intonation will not affect the recognition effect, it is applicable to the intelligent English pronunciation training and evaluation system as a basic requirement. The computation process of MCC is shown in Fig.2 [7].

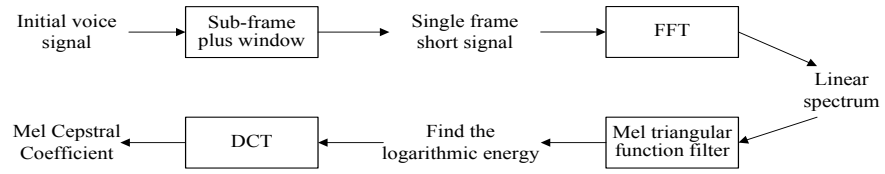

Fig. 2. MCC parameter extraction process

The specific procedure is given as follows:

Convert the initial speech signal into a single-frame short-term signal (n) after pre-processing.

Apply FFT (fast Fourier) butterfly transform algorithm to convert the short-term signal x(n) into the linear spectrum X(k), the formula is given as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{(-j\frac{2\pi kn}{N})}, 0 \le k \le N-1 \quad (1)$$

Where N is the window length.

X(k) is filtered using the Mel-Triangle filter function to solve the logarithmic energy S(i). The Mel-Triangle filter function is shown in the formula (2), and the logarithmic energy S(i) is shown in the formula (3).

$$H_i(k)=\begin{cases}0, k \le f(i) k \ge f(i+2)\\ \dfrac{k-f(i)}{f(i+1)-f(i)}, f(i)<k \le f(i+1)\\ \dfrac{f(i+2)-k}{f(i+2)-f(i+1)}, f(i+1)<k<f(i+2)\end{cases} \tag{2}$$

Where, $\sum_{n=0}^{M-1} H_i(k)=1$, f(i): the center frequency of filter, 0≤i≤M，M: the number of filters.

$$s(i)=\ln\left[\sum_{k=0}^{N-1}\left|X(k)^2\right| H_i(k)\right], 0 \le i \le M \tag{3}$$

DCT is used to discretely transform the logarithmic energy S(i) into a cepstrum domain, i.e., a signal feature parameter MFCC (p is a parameter order).

$$C(n)=\sum_{i=0}^{M-1} s(i)\cdot\cos\left(\frac{n\pi(i+0.5)}{M}\right), n=1.2.\mathrm{L}\ \ p \tag{4}$$

2.3 Speech signal pattern matching

Dynamic time warping (DTW) and Hidden Markov Method (HMM) are the two commonly used approaches for matching the speech signal pattern. Given that the algorithm shall be simple and the intelligent spoken English pronunciation training and evaluation system gets more practical, this paper adopts the DTW algorithm. The schematic diagram of DTW algorithm is shown in Fig. 3 [8].

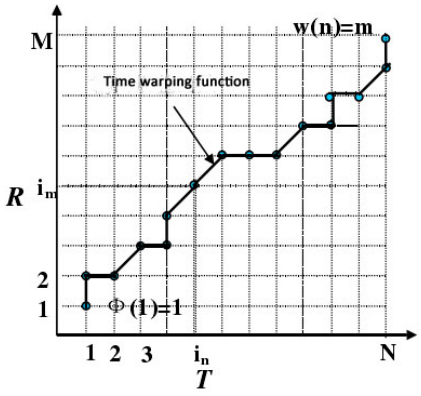

Fig. 3. DTW algorithm schematic diagram

First, the standard speech template and the test template as referred are based to perform computation, respectively. The frame matching distance matrix is available at last, from where the optimal path is found, i.e. the cumulative distance and the mini-

mum (matching distance) of the optimal path function. This path shall depart from the starting point (1, 1) to the destination (N, M) via every intersection. The standard speech template and the test template are compared by this minimum matching distance. The comparative result may reflect the similarity of language features.

3 System Design

3.1 System analysis and design

The development of mobile Internet technology and the popularization of smart phones greatly facilitate people's life and learning. This paper takes the widely used Android mobile phone as the application platform of intelligent spoken English pronunciation training and evaluation system so that English learner can study spoken English anytime, anywhere.

The requirement analysis of the current spoken English learners shows that SRT-based intelligent spoken English pronunciation training and evaluation system shall mainly include the following several core functions:

Audio record and playback module: it, as the basis of the system, is an imperative part for achieving human-computer interaction. The headset set in Android system is used as a recording and playback device.

Video-based pronunciation scoring module: it, as one of the core functions of the system, uses SRT to rate and evaluate learners' spoken English pronunciation, so as to let them recognize their mistakes, help them improve the level of spoken English.

Pronunciation formant graph display module: it can achieve the key function of spoken English learning feedback. Learners may correct their pronunciation errors based on the comparison between the audio formant standard speech template and their pronunciation graph.

3.2 Pronunciation rating method and process design

Pronunciation rating method. This paper draws on the standard speech template reference to achieve the intelligent spoken English pronunciation training and evaluation system for English learners. In order to improve the veracity and reliability of scoring, this paper proposes a rating method based on adaptive parameter (AP) which adapts to different Android mobile phone applications. AP's rating algorithm [9]:

$$score = \frac{100}{1 + x(d)^y}$$
$$d = \frac{D(N,M)}{N} \tag{5}$$

Where, x and y are the adaptive parameters; a separate rating parameter generation module is configured on each device. Before the pronunciation rating, the learner performs the adaptive training for the rating parameter generation. Expert rate it by experience. The best values of x, y are generated by the least squares. D is the frame

match distance; N is test template frame length; D (N, M) is the distance between the standard speech template and the test template.

Designing the speech rating process. A system pronunciation rating process is shown in Fig. 4. After the system preprocesses the speech signal, MCC feature extraction and DTW pattern matching are done to calculate the frame matching distance between the standard speech template and the tester's speech template. If user uses the system for the first time, the expert rating is also required to generate adaptive parameters. Only after the rating function is determined can the systems make a rating. This process runs only once. The system will automatically save the relevant parameters. User can directly get the pronunciation scores when reusing it.

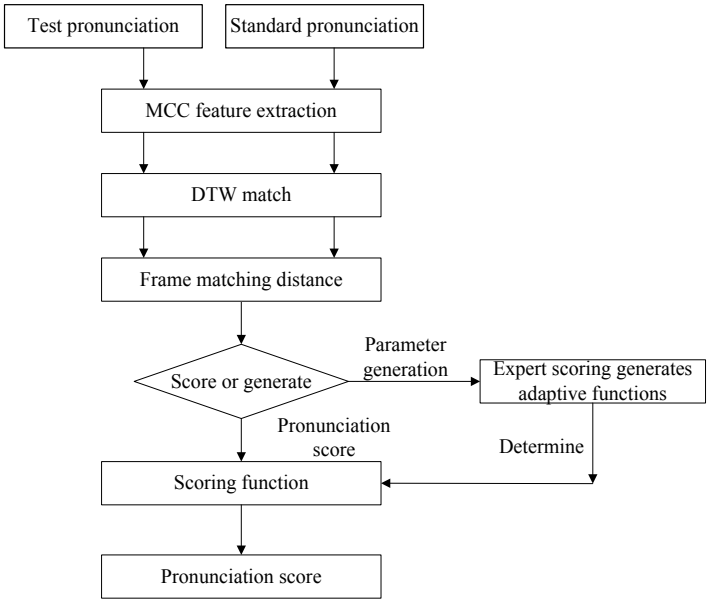

Fig. 4. System pronunciation score process

3.3 Speech feedback and lip reclamation

The relationship between pronunciation formant and orolingual shapes. Chinese people are accustomed to pronounce spoken English using Chinese articulation manner, but the two differ greatly. The study suggests that the most leading problems in spoken English round up to vowels. Chinese articulation rests in the front of the oral cavity while English is "Back vowel pronunciation" [10].

In order to truly enable the feedback function of the intelligent spoken English speech training system, this paper proposes a phonetic formant graphics lip reclamation based on the relationship between the speech mouth-shape and formant [11]. The schematic diagram of the relationship between the speech formant and mouth-shape is shown in Fig. 5 [15]. The vowel pronunciation has three formants, i.e. F1, F2 and F3. the F1 spectrum is the highest, and has the basic characteristics of the speech formant.

This paper adopts the F1 formant as the basis for articulation quality evaluation. The formant frequency can reflect the positions of the oral cavity and the tongue. The higher the frequency is, the lower the tongue is and the larger the mouth opens.

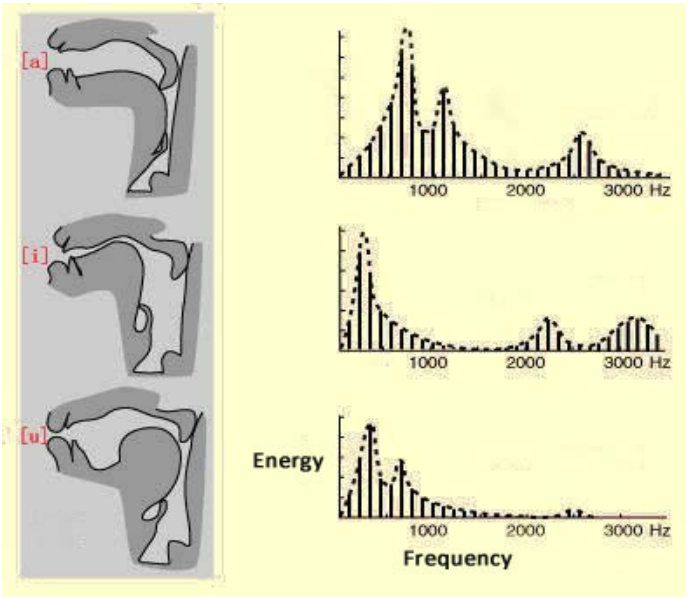

Fig. 5. Different pronunciation resonance diagram

Formant-based lip reclamation. Based on the relationship between the articulation formant and the orolingual shapes, the standard speech template and the tester template can be compared to expose the disparity between the articulation formants for the purpose of the orolingual reclamation to the tester. The comparison of formants is shown in Fig. 5 [12], where the black and red lines represent the formants of the learner and the standard speech templates, respectively. The disparity between the two helps us analyze the similarity between the learner and the standard articulations based on which to correct the mouth shape. As shown in Fig. 6, it follows that the learner should narrow the mouth shape and raise up the tongue to coincide with standard articulation according to the F1 formant mentioned above. Based on the relationship between the articulation formant and the orolingual shapes, the standard speech template and the tester template can be compared to expose the disparity between the articulation formants for the purpose of the orolingual reclamation to the tester. The comparison of formants is shown in Fig. 5 [12], where the black and red lines represent the formants of the learner and the standard speech templates, respectively. The disparity between the two helps us analyze the similarity between the learner and the standard articulations based on which to correct the mouth shape. As shown in Fig. 6, it follows that the learner should narrow the mouth shape and raise up the tongue to coincide with standard articulation according to the F1 formant mentioned above.

Designing speech feedback module process. Speech feedback module mainly extracts the formants from the tester and standard speech templates after pretreatment, FFT transformation, and presents a graphic display for learners so that they analyze gap between their articulation mouths and standard. The pronunciation feedback module process is shown in Fig. 7.

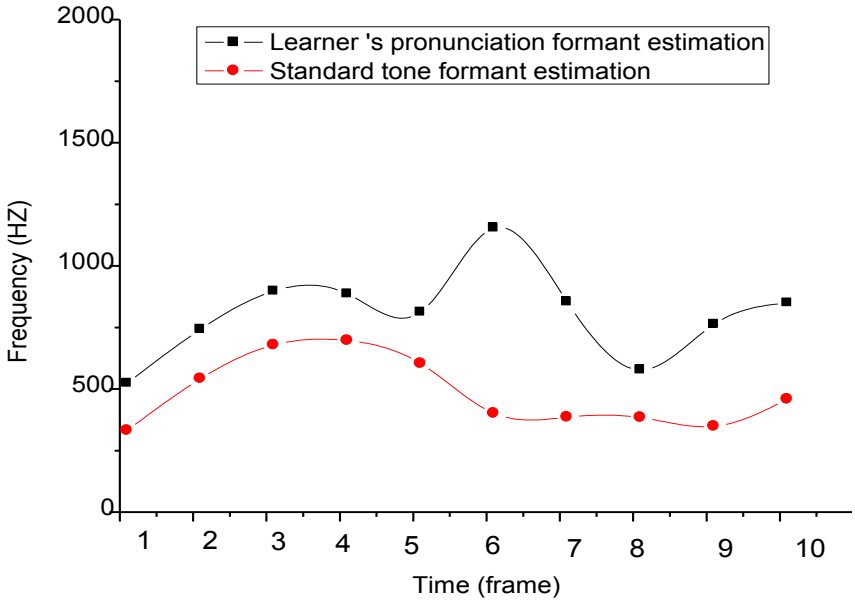

Fig. 6. Pronunciation formant comparison chart

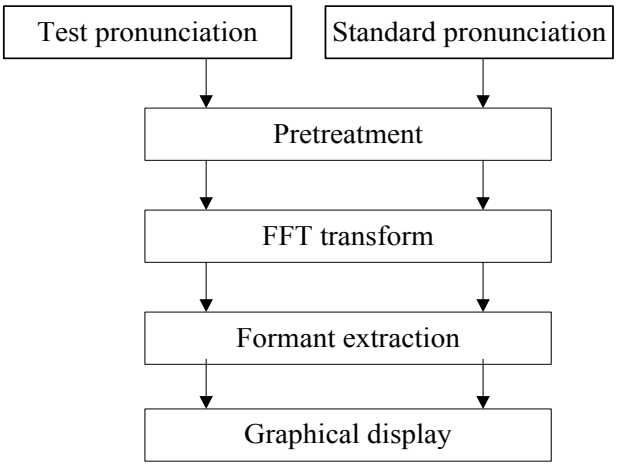

Fig. 7. Formant contrast flow chart

Designing the user interface structure. The system uses the Android system mobile phone as a running platform to facilitate user operation and learning. The user interface is designed to be simple and concise, see Fig. 8 for the user interface structure. After user enters the main interface, he or she may select vowels, consonant phonogram and word pronunciation practice functions, click on the button to run the functions. There is phonics demo, pronunciation, practice, repeat after and other functions set on the phonogram practice interface. The beginner may find the rating parameter adaptive function in the menu, where also has Help and Launch functions.

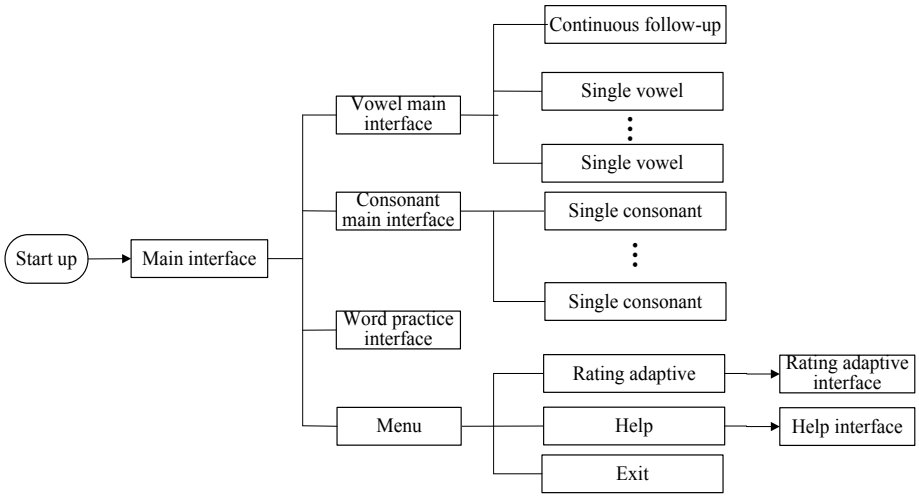

Fig. 8. User interface structure

4 System Implementation

SRT-based intelligent spoken English pronunciation training and evaluation system is developed under the Eclipse integration environment [13], and runs on the Android system mobile phone platform in real machine. In this paper, the main interface of the system and a single phonetics pronunciation exercises are taken as an example to introduce the implementation of system functions.

The system interface is implemented through extended Activity [14]. As shown in Fig. 9, this is the main interface of the system, where there is vowel, consonant and word pronunciation function keys respectively in the left upper side.

Fig. 9. System main interface

When clicking on the Vowel phonetics practice function key, the system enters the individual interface for vowel phonetics practice as shown in Fig. 10. The main interface for vowel phonetics practice is shown in Fig. 11. The learner may click on an array of function keys such as articulation demo, repeat after, contrast and evaluation on demand to enter the corresponding function interface. By clicking on the Pronunciation evaluation, the system will pop up the Scores dialog, as shown in Fig. 12. If learner wants to correct his or her pronunciation, click the View formant button, the system pops up a contrast map about his or her formants with the standard articulation, as shown in Fig. 13, based on which learners can adjust the orolingual shapes using the above method, and repeat pronunciation test and comparison.

Fig. 10. Vowel pronunciation main interface

Fig. 11. Initial interface

Fig. 12. Pronunciation scores

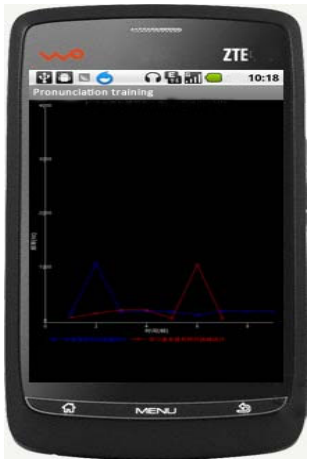

Fig. 13. Formant estimation comparison

5 Conclusion

This paper explores the SRT theory and proposes a SRT-based intelligent spoken English pronunciation training and evaluation system for the purpose of improving the spoken pronunciation efficiency of current English learners with specific results as follows:

Based on SRT, the adaptive parameters (AP), rating method and formant map, which adapts to the spoken English pronunciation evaluation and feedback, are acquired;

The Android system is used as the application platform to design the SRT-based intelligent spoken English pronunciation training and evaluation system in detail;

The system is designed to run on the real machine, and allows expanded functions on the system's main interface and individual pronunciation practices by the test of system availability, which is helpful to improve the learners' level of the spoken English pronunciation to a certain extent.

6 References

[1] Chen, C.M., Chung, C.J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. Computers & Education, 51(2), 624-645. https://doi.org/10.1016/j.compedu.2007.06.011

[2] 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. https://doi.org/10.1121/1.428279

[3] Demenko, G., Wagner, A., Cylwik, N. (2010). The use of speech technology in foreign language pronunciation training. Archives of Acoustics, 35(3), 309-329. https://doi.org/10.2478/v10168-010-0027-z

[4] Felps, D., Bortfeld, H., Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. Speech Communication, 51(10), 920-932. https://doi.org/10.1016/j.specom.2008.11.004

[5] Flege, J.E., Bohn, O.S., Jang, S.Y. (1997). Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics, 25(4), 437-470. https://doi.org/10.1006/jpho.1997.0052

[6] Kawai, G., Hirose, K. (2000). Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology. Speech Communication, 30(2), 131-143. https://doi.org/10.1016/S0167-6393(99)00041-2

[7] Kloosterman, S.H. (1994). Design and implementation of a user-oriented speech recognition interface: the synergy of technology and human factors. Interacting with Computers, 6(1), 41-60. https://doi.org/10.1016/0953-5438(94)90004-3

[8] Molina, C., Yoma, N.B., Wuth, J., Vivanco, H. (2009). ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech Communication, 51(6), 485-498. https://doi.org/10.1016/j.specom.2009.01.002

[9] Peng, L., Setter, J. (2000). The emergence of systematicity in the English pronunciations of two Cantonese-speaking adults in Hongkong. English World-Wide, 21(1), 81-108. https://doi.org/10.1075/eww.21.1.05pen

[10] Sandberg, J., Maris, M., Geus, K.D. (2011). Mobile English learning: an evidence-based study with fifth graders. Computers & Education, 57(1), 1334-1347. https://doi.org/10.1016/j.compedu.2011.01.015

[11] Sato, K., Kato, M., Kosaka, T. (2013). An investigation of vowel substitution rules in the automatic evaluation system of English pronunciation. Journal of the Acoustical Society of America, 133(5), 3247. https://doi.org/10.1121/1.4805215

[12] Tatsuhito, H., Makoto, K., Ban, H. (2015). An English vocabulary learning support system for the learner's sustainable motivation. Springer plus, 4(1), 99. https://doi.org/10.1186/s40064-015-0792-2

[13] Williamson, D.T., Draper, M.H., Calhoun, G.L., Barry, T.P. (2005). Commercial speech recognition technology in the military domain: results of two recent research efforts. International Journal of Speech Technology, 8(1), 9-16. https://doi.org/10.1007/s10772-005-4758-6

[14] Yukari, H. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. Computer Assisted Language Learning, 17(3-4), 357-376. https://doi.org/10.1080/0958822042000319629

[15] Goto, M., Takata, S., Uekawa, Y. (1988). Microprocessor based English speech training system. IEEE Transactions on Consumer Electronics, 34(3), 824-834. https://doi.org/10.1109/30.20190

7 Author

Zhaoxia Yin received the B.A. degree in English language teaching from Ji Lin Normal University, Siping, China, in 1992, and the M.A degree in foreign and applied linguistics from Beihua University in 2007. She is currently working toward the Ph. D. degree at Kyungnam University. Her research interests include language teaching, translation studies and practice.

Article submitted 16 October 2017. Resubmitted 29 November 2017. Final acceptance 23 February 2018. Final version published as submitted by the author.