A Probe into Spoken English Recognition in English Education Based on Computer-Aided Comprehensive Analysis

https://doi.org/10.3991/ijet.v15i03.12937

Yongjuan Wang, Peng Zhao (✉)
Yanching Institute of Technology, Langfang, China
zhaopeng@yit.edu.cn

Abstract—At present, computer-aided spoken English learning is becoming increasingly popular among learners. The computer-aided comprehensive analysis technology can evaluate and correct learner's spoken pronunciation, thereby improving their pronunciation. Based on computer-aided comprehensive analysis, this paper aims to explore the automatic recognition and scoring methods of spoken English in English education. For this, it studies the effective matching of the feedback information with the known pronunciation scoring results, and then develops a computer evaluation plug-in consisting of different modules such as user login, English spoken speech acquisition and recognition, voice evaluation, speech broadcast, and spoken dialogue. The research results show that the computer evaluation plug-in matches and compares the extracted feature parameters of input speech with the standard features, scores the spoken language input by the learner, and gives the correct pronunciation so that the learner can get feedback in time. For different stages of English learning, the focus of recognition technology and the spoken recognition algorithms applied also vary. The research findings provide theoretical and technical support for oral English recognition, error correction and scoring.

Keywords—Computer, spoken English, pronunciation, automatic recognition, feature parameters

1 Introduction

English is a universal language for global communication. With the continuous acceleration of the global integration, the demand for English learning has grown rapidly. Spoken language recognition in English education has always been a difficult point for English learning [1]. At present, affected by the mother tongue, Chinese native learners are used to practicing or recognizing English using the Chinese pronunciation method despite the difference in pronunciation features between Chinese and English, which results in some difficulties in recognizing English pronunciation [2-3]. The recognition technology of spoken English began in the 1950s. However, with the continuous expansion of the English field, the constraints on spoken language recognition such as small-vocabularies, speaker-independent and isolated-words need to be relaxed, and the expansion of the vocabulary makes it difficult to select and establish the templates of spoken language recognition [4]. Now there have been many methods of spoken language recognition in English education. The commonly used ones are methods based on vocal tract models and phonetic knowledge, the template matching method, and the artificial neural networks etc. [5].

Following the rapid development and widespread application of computer technology, the computer's powerful data analysis and processing capabilities and colorful multimedia functions have greatly enhanced the English learning efficiency, and computer-aided speech recognition and language learning have increasingly attracted more attention [6-7]. The computer-aided spoken English recognition system mainly studies how to measure the indicators such as intonation, stress, speed of sound, and prosody. It can detect and correct a given spoken pronunciation error by comprehensively evaluating the quality of pronunciation [8]. Spoken English recognition can help correct wrong pronunciations, and improve the accuracy of English pronunciation, which lays the foundation for further learning [9]. The computer-aided comprehensive analysis can be performed to recognize spoken English without relying on specific time and place. It can perform real-time monitoring of pronunciation errors in a targeted manner, and ensure an effective English teaching with computer's powerful computing and data analysis capabilities [10]. In view of the above, this paper attempts to explore the automatic recognition and scoring methods of spoken English based on the comprehensive computer analysis technology, and effectively pairs the feedback information with the known pronunciation scoring results. This study provides theoretical and technical support for spoken English recognition, error correction and scoring.

2 Spoken English Recognition Algorithms

2.1 Selecting feature parameters

Spoken English pronunciation is related to phonetic symbols and phonemes etc. The pronunciation signals are analog signals that change over time. Generally, preprocessing is required during the recognition of spoken English. Common preprocessing methods include pre-emphasis, windowing and framing, etc. [11]. Among them, the windowing of signals is realized by weighting a movable finite-length window. Let x (n) and w (n) be the spoken language signals and window functions, respectively, then the windowing process can be expressed as:

$$xw(n)=x(n)w(n-w) \tag{1}$$

Currently, there are three commonly used window functions:Rectangular window, Hanning window and Hamming window. Their function expressions are shown below:

Rectangular window:

$$w(n)=\begin{cases} 1 & 0 \le n \le N-1 \\ 0 & other \end{cases} \tag{2}$$

Hanning window:

$$w(n)=\begin{cases} 0.5 + 0.5\cos[\pi n(N-1)] & 0 \le n \le N-1 \\ 0 & other \end{cases} \tag{3}$$

Hamming window:

$$w(n)=\begin{cases} 0.54 + 0.4\cos[\pi n/(N-1)] & 0 \le n \le N-1 \\ 0 & other \end{cases} \tag{4}$$

where, N is the window size.

The level of spoken English pronunciation by English learners is feedback to the learner from the computer, so the computer must have a clear reference or standard in the measurement process [12]. During the English spoken language recognition, feature comparison is used to compare learner's spoken language features with those of standard English for scoring and error analysis in spoken language [13]. The feature parameters selected during this process include pitch period and short-term average magnitude, which mainly refer to phonetic symbols and phonemes in English [14].

2.2 Scoring mechanism and scoring parameters adjustment in feature comparison of spoken language

When processing spoken English by machine, it is difficult to compare simple input speech features with standard speech features, because the language has regional and accent features, and the input signal has a large randomness [15]. It’s assumed that the feature vector of the standard spoken language template is a, and the feature vector of the input spoken language is b. In the recognition algorithm, the distance between features needs to be calculated, but it cannot be directly used as the score of the pronunciation. This study attempts to explore the relationship between them, as shown in Equation 5:

$$\text{Score}=\frac{100}{1+a(dist)^{b}} \tag{5}$$

Figure 1 shows the scoring functions by the feature comparison. It can be seen that as the distance increased, the scoring value decreased. The actual scoring is shown in Equation 5, in which the feature vectors of the standard spoken language template and the input spoken language were used. In the actual calculation, the weighted average of the two should be taken. The adjusted formula for the parameters is shown in Equation 6.

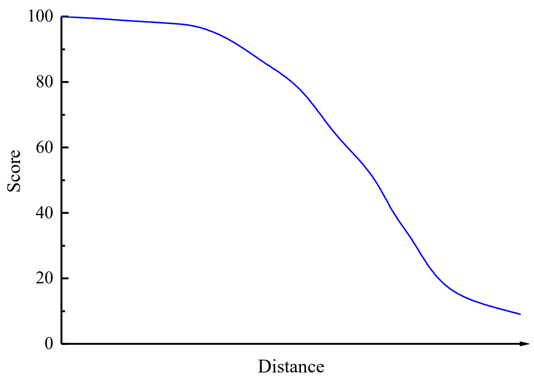


Fig. 1. Feature comparison scoring function

$$\text{Score}=\text{w1}\cdot\frac{100}{1+a_1(dist_1)^{b_1}}+\text{w2}\cdot\frac{100}{1+a_2(dist_2)^{b_2}} \tag{6}$$

where, w_1 and w_2 are the weights of two feature parameters, and $w_1+w_2=1$.

3 Automatic Recognition and Scoring Methods of Spoken English

3.1 Speech recognition technology in spoken English learning

In spoken English learning based on speech recognition, pronunciation is the key. For different stages of English learning, the focus of recognition technology also varies. Some recognition algorithms of spoken English are specifically used to study the pronunciation errors of beginners, and some are to the entire English pronunciation skills [16]. In English education, both teacher or parent are not willing to let students use the spoken English learning system oriented to accuracy recognition. Figure 2 shows the functional modules of the spoken English learning system including the login module, speech recognition, speech evaluation, voice announcement, and spoken dialogues, etc., where the speech recognition module was subdivided into speech collection, speech data pre-processing, feature extraction of speech data, etc.; speech evaluation was subdivided into speech recognition, phoneme scoring, and comprehensive scoring. Figure 3 shows a flow chart of spoken English recognition. After preprocessing the spoken language signal, the acoustic parameters were analyzed; then, its measure was estimated by the distortion measure, and the recognition result was determined. Figure 4 shows a flow chart of the spoken English recognition modules, in which the learner's spoken pronunciation and standard pronunciation were first extracted respectively; then the reference phoneme model was used for forced alignment, to score the phonemes; the correction suggestions or comprehensive scores were finally given. In short, the main modules of the spoken language recognition system include five parts: feature value extraction, factor recognition, factor correlation, pronunciation evaluation, and error detection.

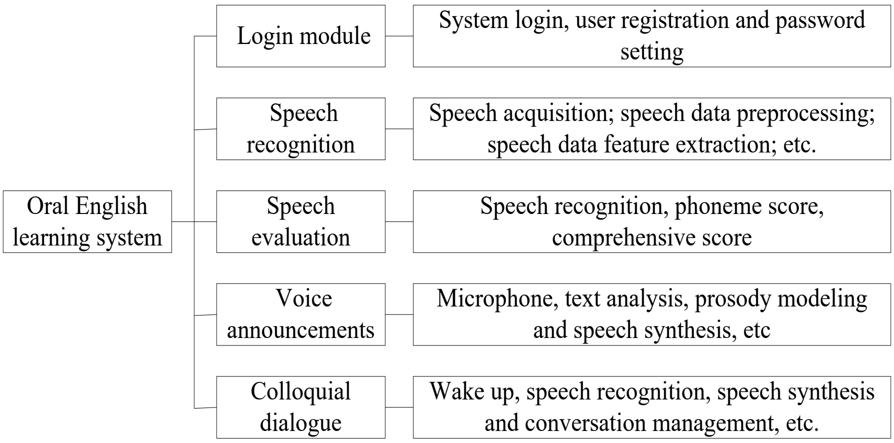

Oral English learning system
Login module
System login, user registration and password setting
Speech recognition
Speech acquisition; speech data preprocessing; speech data feature extraction; etc.
Speech evaluation
Speech recognition, phoneme score, comprehensive score
Voice announcements
Microphone, text analysis, prosody modeling and speech synthesis, etc
Colloquial dialogue
Wake up, speech recognition, speech synthesis and conversation management, etc.

Fig. 2. Function module of spoken English learning system

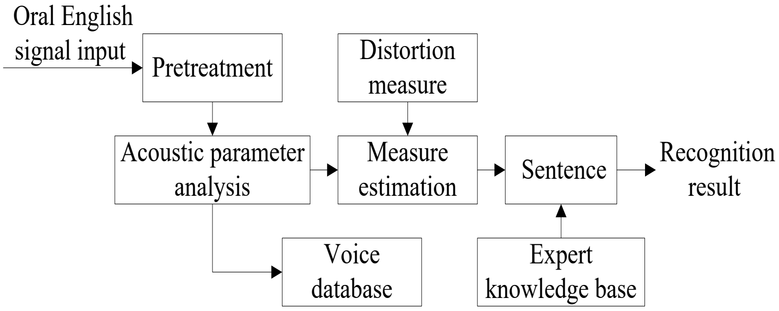


Fig. 3. Flow chart of spoken English recognition

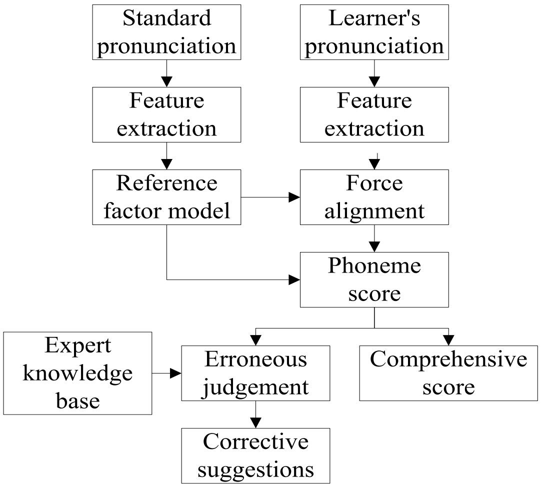


Fig. 4. Flow chart of spoken English recognition module

3.2 Application of speech recognition scoring system in spoken English

The speech recognition scoring system is one of the most important applications of computer-aided language learning. The quality of the scoring system is strongly dependent on speech recognition technology. Figure 5 shows the flow chart of the scoring system for spoken English recognition. It started with extracting feature parameters of a speech signal, and then cut the speech signal. In the end, the scoring was made by voice recognition and tone recognition of a single syllable. When using the corresponding scoring method, the system recognizes the spoken language and produces related results. Spoken language recognition can score the entire sentence, individual words, individual phonemes, or the overall rhythm. Figure 6 shows the flow chart of the scoring system. The scoring process of the entire system can be divided into two parts. The first is to extract the spoken language features and select a suitable recognition algorithm for eliminating the false scores caused by differences in standard pronunciation; second, different score units at phoneme-level should be decomposed according to the spoken language features to eliminate noises. Figure 6 shows the operation flow of spoken English recognition process, which fulfils the speech recognition, phoneme scoring, and pronunciation evaluation at a time.

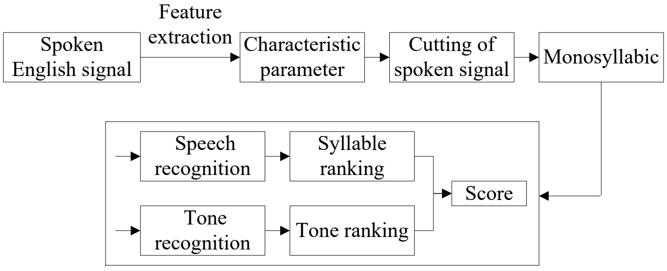


Fig. 5. The flow chart of spoken English recognition scoring system

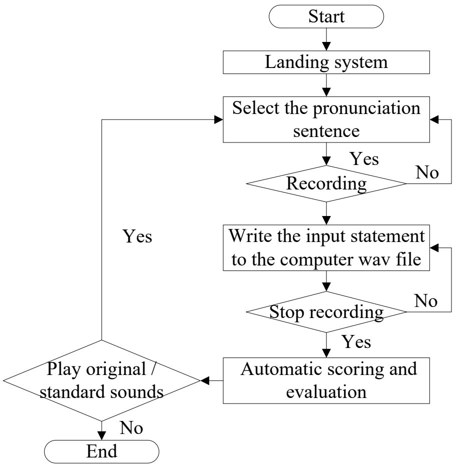

Start
Landing system
Select the pronunciation sentence
Yes
Recording
No
Yes
Write the input statement to the computer wav file
Stop recording
No
Yes
Automatic scoring and evaluation
Play original / standard sounds
No
End

Fig. 6. Flow chart of scoring system

4 Design and Implementation of Spoken English Recognition System Based on Computer-Aided Comprehensive Analysis

4.1 Detailed design and implementation of spoken English recognition system

The core of spoken English education is the design and implementation of spoken English recognition system. Whether it is an English sentence or a word, the recognition process is based on a single English vocabulary. The single standard spoken speech samples were collected to provide data for the training templates and training codebooks of the system. Based on the computer-aided comprehensive analysis, the spoken English recognition system in this paper involved sampling frequency, status number, and isolated-word speech number, and a sequence of isolated word speech strings related to English. Figure 7 shows the core framework of the spoken English recognition system. It first read the input audio file, and performs pre-processing, feature extraction, and feature matching of the voice signals; then, it combined reading and display of computer multimedia learning materials to perform phonetic learning and words learning, pronunciation recognition and pronunciation correction.

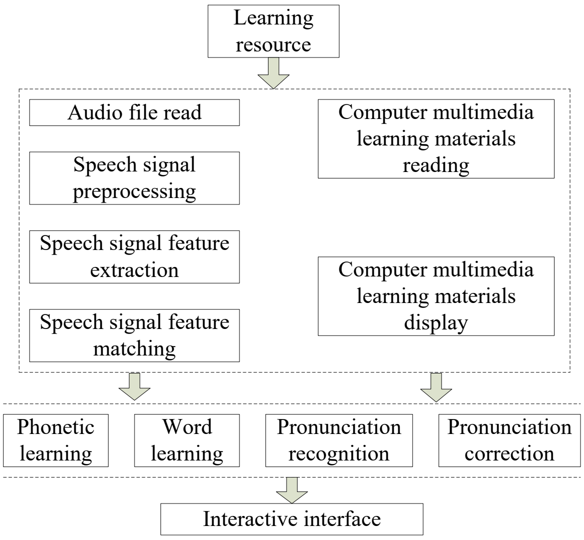

Learning resource
Audio file read
Speech signal preprocessing
Speech signal feature extraction
Speech signal feature matching
Computer multimedia learning materials reading
Computer multimedia learning materials display
Phonetic learning
Word learning
Pronunciation recognition
Pronunciation correction
Interactive interface

Fig. 7. The core framework of spoken English recognition system

In the initial spoken language recognition, the speech data collected by the computer were original signals, which need to be pre-processed step by step. This process included sampling and quantization of analog signals, format conversion of digital signals, mute processing at the beginning and end, and pre-emphasis of voice data. Figure 8 shows a model of spoken English learning and recognition based on computer-aided comprehensive analysis. It can be seen from Figure 8 that there are three ways to learn spoken English: graphic learning, video learning, and audio learning, among which the video learning method is to use pronunciation articulation, sentence pronunciation demonstration and guidance, and imitate pronunciation; while the audio learning is by listening to standard pronunciation and learner pronunciation to form contrasts in pronunciation so that learners can adjust pronunciation in time according to feedback.

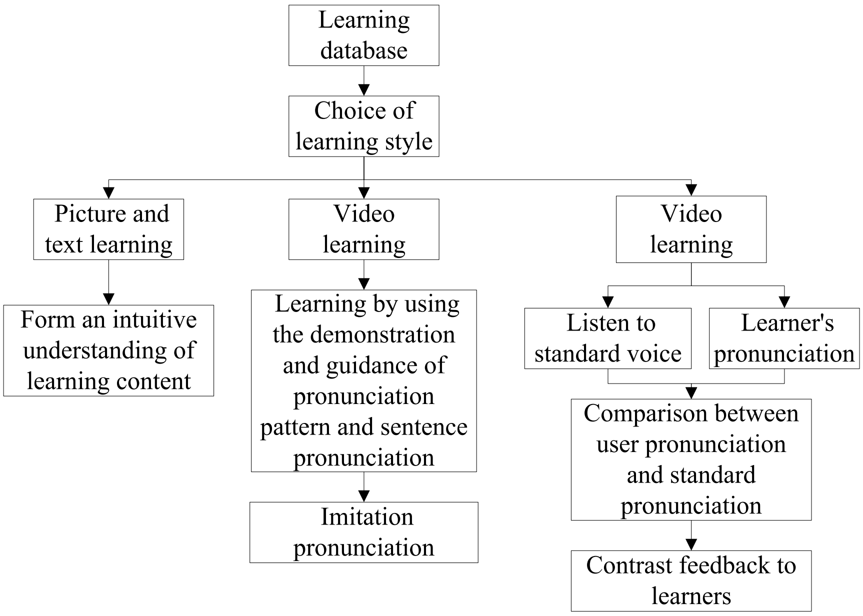
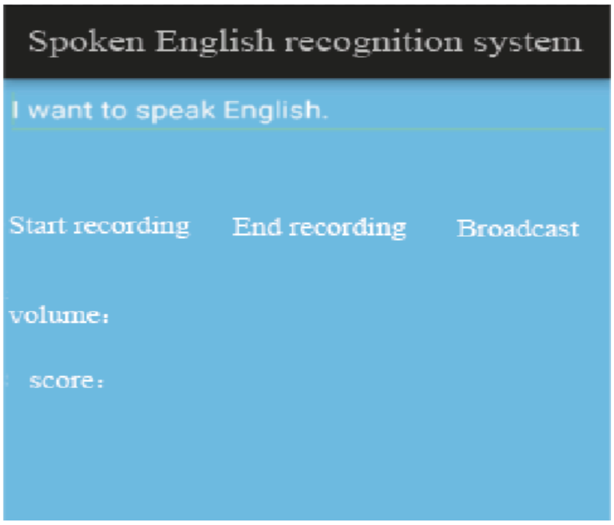

Learning database
Choice of learning style
Picture and text learning
Video learning
Video learning
Form an intuitive understanding of learning content
Learning by using the demonstration and guidance of pronunciation pattern and sentence pronunciation
Imitation pronunciation
Listen to standard voice
Learner's pronunciation
Comparison between user pronunciation and standard pronunciation
Contrast feedback to learners

Fig. 8. A model of spoken English learning and recognition based on computer-aided comprehensive analysis

4.2 Development of a computer evaluation plug-in

Spoken English learning is established through stimulus and response. It is a process of identification and feedback. With the widespread application of computer-aided learning or teaching tools, students can actively choose and process in the learning process, thereby constructing internal mental representations, that is, during the learning process using computer-aided systems, learners can use spoken language recognition system to build their own knowledge system. In order to implement the spoken language recognition system based on speech recognition technology, the author developed a computer evaluation plug-in. This plug-in can better help learners to standardize the English pronunciation. The whole plug-in module includes user login, speech acquisition and recognition of spoken English, speech evaluation, voice broadcast, and spoken dialogue, of which speech acquisition includes the main parameters of sampling rate, audio data format, and audio source etc. Figure 9 shows the interface for the spoken English acquisition and recognition. The acquisition module of the entire plug-in includes start recording, end recording and broadcast. The entire internal evaluation module is given by a detailed comparison between the spoken language recognition algorithm and the phoneme level scoring unit. Figure 10 shows the functional operation flow of the spoken language input module. The input module is controlled by three buttons: microphone on, microphone off and voice broadcast. The spoken language input and broadcast form a complete feedback and correction process of spoken English.


Spoken English recognition system

I want to speak English.

Fig. 9. Spoken English acquisition and recognition interface

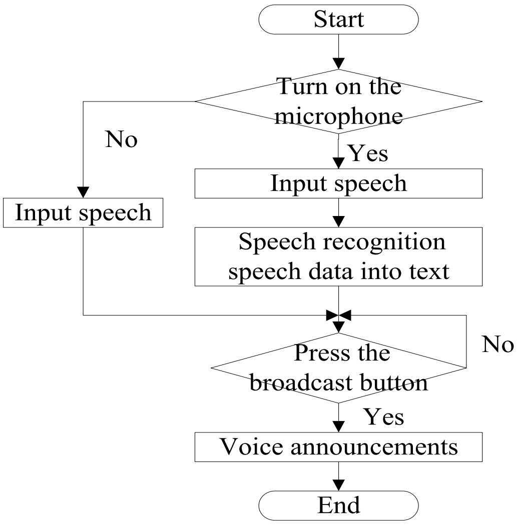


Fig. 10. Function operation process of oral input module

5 Conclusion

Based on the computer-aided comprehensive analysis, this paper explores the automatic recognition and scoring methods of spoken English, and effectively matches the feedback information with the known pronunciation scoring results. The specific conclusions are as follows:

- The functional modules of the spoken English learning system include a login module, speech recognition, speech evaluation, speech broadcast, and spoken dialogue etc.; speech evaluation includes speech recognition, phoneme scoring, and comprehensive scoring.
- The quality of the scoring system is strongly dependent on speech recognition technology. First, the spoken language features must be extracted and a suitable recognition algorithm must be selected to eliminate false scores caused by differences in pronunciation. Second, different score units at phoneme-level should be decomposed according to the spoken language features to eliminate noises.
- There are three ways to learn spoken English: graphic learning, video learning, and audio learning. This paper develops a computer evaluation plug-in. This plug-in can better help learners to standardize the English pronunciation. The whole plug-in module includes user login, speech acquisition and recognition of spoken English, speech evaluation, voice broadcast, and spoken dialogue.

6 References

[1] Moore, R., Caines, A., Graham, C., Buttery, P. (2015). Incremental Dependency Parsing and Disfluency Detection in Spoken Learner English. International Conference on Text, Speech, and Dialogue, 9302: 470-479. https://doi.org/10.1007/978-3-319-24033-6_53

[2] Maeda, N., Yoneyama, K. (2015). Foreign accentedness of English sentences spoken by Japanese EFL learners and Japanese teachers of English: a first report. The Journal of the Acoustical Society of America, 138(3): 1946. https://doi.org/10.1121/1.4934155

[3] Li, L. (2014). A data-based investigation into reliability and validity of computer-assisted oral English test. Applied Mechanics and Materials, 543-547, 4494-4497. https://doi.org/10.4028/www.scientific.net/amm.543-547.4494

[4] Zhang, T., Hasegawa-Johnson, M., Levinson, S. E. (2006). Extraction of pragmatic and semantic salience from spontaneous spoken English. Speech Communication, 48(3-4): 437-462. https://doi.org/10.1016/j.specom.2005.07.007

[5] Rodríguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., Bordel, G. (2016). Kalaka-3: a database for the assessment of spoken language recognition technology on youtube audios. Language Resources and Evaluation, 50(2): 221-243. https://doi.org/10.1007/s10579-015-9324-5

[6] Raguram, L. S. B., Shanmugam, V. M. (2017). Deep belief networks for phoneme recognition in continuous Tamil speech–an analysis. *Traitement du Signal*, 34(3-4): 137-151. https://doi.org/10.3166/ts.34.137-151

[7] Lopez, C., Dhouib, M. T., Cabrio, E., Zucker, C. F., Gandon, F., Segond, F. (2018). SMILK, linking natural language and data from the web. *Revue d'Intelligence Artificielle*, 32(3): 287-312. https://doi.org/10.3166/ria.32.287-312

[8] Zhang, J., Xu, J., Bao, X. G., Zhou, R., Yan, Y. H. (2017). Weighted phone log-likelihood ratio feature for spoken language recognition. Qinghua Daxue Xuebao/Journal of Tsinghua

University, 57(10): 1038-1041, 1047. http://dx.doi.org/10.16511/j.cnki.qhdxxb.2017.25.042

[9] Navratil, J. (2001). Spoken language recognition-a step toward multilinguality in speech processing. Speech & Audio Processing IEEE Transactions on, 9(6): 678-685. https://doi.org/10.1109/89.943345

[10] Siniscalchi, S. M., Reed, J., Svendsen, T., Lee, C. H. (2013). Universal attribute characterization of spoken languages for automatic spoken language recognition. Computer Speech & Language, 27(1): 209-227. https://doi.org/10.1016/j.csl.2012.05.001

[11] Ferrer, L., Lei, Y., Mclaren, M., Scheffer, N. (2016). Study of senone-based deep neural network approaches for spoken language recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(1): 105-116. https://doi.org/10.1109/taslp.2015.2496226

[12] Tong, R., Ma, B., Li, H., Chng, E. S. (2009). A target-oriented phonotactic front-end for spoken language recognition. IEEE Transactions on Audio, Speech and Language Processing, 17(7): 1335-1347. https://doi.org/10.1109/tasl.2009.2016731

[13] Zhu, D.L., Li, H.Z., Ma, B., Lee, C. H. (2008). Optimizing the performance of spoken language recognition with discriminative training. IEEE Transactions on Audio, Speech and Language Processing, 16(8): 1642-1653. https://doi.org/10.1109/tasl.2008.2005319

[14] Ng, R. W. M., Nicolao, M., Hain, T. (2017). Unsupervised crosslingual adaptation of tokenisers for spoken language recognition. Computer Speech & Language, 46: 327-342. https://doi.org/10.1016/j.csl.2017.05.002

[15] Zhang, W. Q., Liu, W. W., Li, Z. Y., Shi, Y. Z., Liu, J. (2014). Spoken language recognition based on gap-weighted subsequence kernels. Speech Communication, 60: 1-12. https://doi.org/10.1016/j.specom.2014.01.005

[16] Shabnam, G. F., Shaghayegh, R., Yasser, S. (2018). Spoken language recognition using a new conditional cascade method to combine acoustic and phonetic results. International Journal of Speech Technology, 21(3): 649-657. https://doi.org/10.1007/s10772-018-9526-5

7 Authors

Yongjuan Wang was born on 17th, October, 1982 in HanDan City, Hebei Province, China. She received B.A. degree of English from Hebei University in 2006 and M.A degree in Foreign Linguistics and Applied Linguistics from University of International Business and Economics in 2015. Her M.A. thesis is Needs Analysis of Business English Majors and Its Implications. She started her teaching career in 2007 in Yanching Institute of Technology. Since 2007, she has authored than 10 papers in English teaching and literature, compiled two teaching materials and a CET-4 Writing tutorial， participated in a provincial and a municipal research projects.

Peng Zhao was born on 22nd, September, 1981 in HanDan City, Hebei Province, China. He received his B.E. in Computer Science and Technology from North China Institute of Science and Technology in 2007 and Master Degree by research in Computer Technology from Beijing University of Chemical Technology in 2014. He started his career in 2007 in Yanching Institute of Technology. He has authored more than 10 papers, compiled a number of teaching materials, presided over several school-level and municipal research projects and participated in a provincial level research project.

Article submitted 2019-11-30. Resubmitted 2020-01-09. Final acceptance 2020-01-09. Final version published as submitted by the authors.