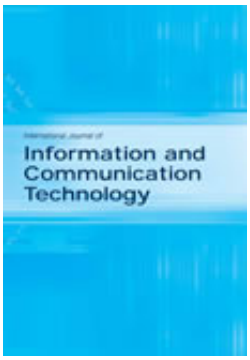
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Enhancing oral English self-study: a speech knowledge recognition algorithm approach

Yuanyuan Zhang

Article History:

Received:	10 April 2024
Last revised:	20 May 2024
Accepted:	23 May 2024
Published online:	02 August 2024

Enhancing oral English self-study: a speech knowledge recognition algorithm approach

Yuanyuan Zhang

International Education School of Jiangsu Maritime Institute,
Nanjing, Jiangsu, 211170, China
Email: psr20120202@163.com

Abstract: In today's globally connected world, English proficiency is vital for effective communication. This paper introduces a novel approach utilising a speech knowledge recognition algorithm to evaluate oral English self-study. Through a comprehensive analysis comparing self-study with standard oral English, aspects such as acoustics, rhythm, and perception are assessed. By integrating both subjective and objective evaluations, the proposed algorithm provides a robust framework for assessing oral English proficiency. The ultimate goal is to improve the efficiency and efficacy of English self-study. Simulation results affirm the effectiveness of the speech knowledge recognition algorithm in evaluating oral English proficiency.

Keywords: speech knowledge recognition algorithm; oral English score; error detection; prosodic correction.

Reference to this paper should be made as follows: Zhang, Y. (2024) 'Enhancing oral English self-study: a speech knowledge recognition algorithm approach', *Int. J. Information and Communication Technology*, Vol. 25, No. 5, pp.40–49.

Biographical notes: Yuanyuan Zhang studied at China Petroleum University, Beijing in 2001. He obtained his Bachelor's degree in 2005 and Master's degree in 2007. He graduated in 2007 and worked at the Jiangsu Maritime Institute. He published more than 20 papers in provincial-level or above journals, mainly in the field of English language teaching.

1 Introduction

Oral English proficiency is a critical skill in various contexts, including education, professional development, and social interactions. In educational settings, it enables effective communication with peers and educators from diverse linguistic backgrounds. Professionally, it is often a determinant factor for career advancement and international collaboration. Moreover, in social interactions, the ability to speak English fluently can significantly enhance cross-cultural understanding and global citizenship. With the continuous advancement of China's social economy, foreign exchange has become an essential conduit (Williams et al., 2016; Toledo et al., 2015). In comparison to traditional English learning, greater emphasis is now placed on comprehension in reading and writing, often at the expense of oral proficiency. Consequently, individuals may excel in

writing but struggle with verbal communication, hindering their ability to meet practical communication needs (Yu et al., 2016). Fortunately, the development of computer technology has facilitated extensive resource sharing on the internet, offering significant convenience for English learners. Voice technology enables the sharing of standardised pronunciation resources, allowing users to select suitable resources for comparison and improvement of oral pronunciation (Jeanjean et al., 2015; Oyewole, 2017). Previous studies have explored the use of speech recognition technology for language learning and pronunciation improvement, highlighting the potential of such systems to provide personalised feedback and facilitate autonomous learning.

However, it is crucial to acknowledge the complexity surrounding speech recognition. Variations in accents pose a significant challenge, as individuals may have diverse accents even when uttering the same sentence or word. Additionally, accurate recognition of English accents involves complexities beyond word recognition, extending to sentence structure and pauses (Tantiwich and Sinwongsuwat, 2019; Lavid and Moratón, 2015). Industry experts have endeavoured to assess English pronunciation quality, employing speech signal detection and feature extraction techniques. Despite efforts to analyse oral English pronunciation signals through intelligent signal processing, limitations persist due to challenges in feature extraction and detection (Mante-Estacio and Bernardo, 2015).

Advancements in artificial intelligence technology have introduced novel detection methods such as wavelet analysis, time-frequency analysis, and Fourier feature extraction. However, these methods often entail complex computations, which can result in extended waiting times, potentially diminishing users' interest in self-study systems for oral English. Prolonged waiting times may ultimately impede English learning and oral practice (Born et al., 2016). Accurate pronunciation is paramount for effective communication, as incorrect pronunciation or word usage can impede communication efficiency. Hence, systems facilitating the correction of oral English pronunciation are essential (Allan, 2018; Arizavi et al., 2019; Afshar, 2021).

Addressing these limitations and demands, this paper introduces a speech knowledge recognition algorithm. The objective of this study is to develop a speech knowledge recognition algorithm that can enhance the self-study of oral English. This paper provides a comprehensive overview of the proposed algorithm, its implementation, and its evaluation. The structure of the paper is as follows: after the introduction, we review the existing literature on oral English self-study methods and speech recognition technologies. We then present the details of our speech knowledge recognition algorithm, followed by a comparison with existing systems. Finally, we conclude with the findings and potential applications of our research. By examining key indicators of oral English pronunciation evaluation and integrating assessments from acoustic, rhythmic, and perceptual perspectives, the paper constructs a scoring mechanism for speech recognition algorithms. This approach promotes self-study and self-correction of oral English, enhancing the effectiveness and efficiency of English teaching while fostering student engagement and enjoyment in learning English (Techentin et al., 2021 Lee et al., 2017; Mishchenko, 2019). To enhance the speech recognition algorithm's ability to understand various accents and dialects, future research could focus on collecting and analysing speech data from different regions worldwide. By constructing a database that encompasses a wide range of accents and dialects, the algorithm can better learn and adapt to these differences. Additionally, researchers can explore transfer learning

techniques, which enable the algorithm to apply knowledge learned in one language environment to another, thereby improving its generalisation capabilities.

2 Voice scoring mechanism and scoring technology

2.1 Speech score based on speech knowledge recognition feature comparison

The primary feature of voice knowledge identification scoring lies in comparing a user's spoken English pronunciation with the standard English pronunciation as a reference. This process involves extracting characteristic parameters from the spoken English speech under evaluation and comparing them with those of standard spoken English. Through data normalisation and adjustment, utilising speech recognition technology facilitates a corresponding comparison of knowledge. The similarity between the two sets of data is then calculated, and an evaluation index is employed for scoring (Wang, 2021; Kang, 2021; Wang et al., 2021).

In terms of spoken English characteristics, distinctions are primarily drawn from sound intensity, phonemes, speed, and speech content. Sound intensity is quantified by the relevant amplitude of the speech signal, while phonemes are depicted through trajectory curves. Speed is discerned by the signal's pace of change, and the final evaluation result is derived through a comprehensive assessment of each element. Given the potential impact of biased training data on algorithm performance, researchers can employ data augmentation techniques to expand the training set by simulating various pronunciation conditions. For instance, adding background noise, varying pitch and speech rate can enhance model robustness. Meanwhile, ensemble learning methods, such as random forests or gradient boosting trees, can combine the predictive outcomes of multiple models to improve the recognition accuracy of speech features from underrepresented linguistic groups. Considering the challenges of resource-constrained environments, researchers can explore the development of speech recognition algorithms with lower resource requirements to operate on devices with lower specifications. Additionally, the development of offline or lightweight application versions can enable users with unstable or limited internet access to effectively use self-study tools.

2.2 Voice score based on acoustic model

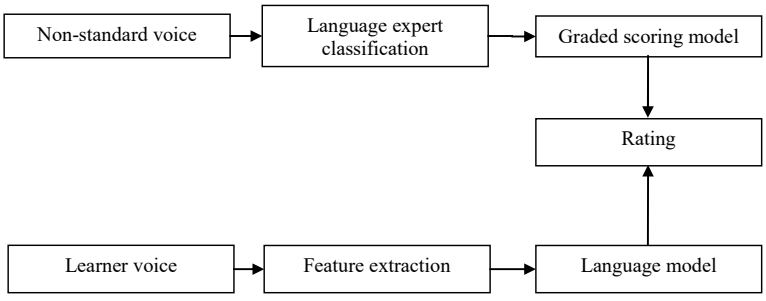
For acoustic models, standard spoken English is not required for comparison and reference (Lin et al., 2021; Cao and Hao, 2021). The specific process is shown in Figure 1.

Assuming that the basic scoring unit is a phoneme and Γ_i is the start time of the i^{th} phoneme, the scoring formula is shown in formula (1):

$$l_i = \sum_{t=t_i}^{t_{i+1}-1} \log\left(p\left(s_t \left| s_{t-1}\right.\right) p\left(x_t \left| s_t\right.\right)\right) \tag{1}$$

where x_t and s_t are the states of the observation vector and HMM at time T respectively; $P(s_t| s_{t-1})$ is the probability of transition; $P(x_t| s_t)$ is the output probability distribution of state s_t.

Figure 1 Speech scoring process based on acoustic model



For each frame of segment I of the phoneme Qi, the frame-based posterior probability $P(q_i|\, x_t)$ of the phoneme Qi is calculated as shown in formula (2):

$$p\left(q_i \left| x_i \right.\right) = \frac{p\left(x_t \left| q_t \right.\right) p\left(q_t\right)}{\sum_{q=1}^{M} p\left(x_t \left| q \right.\right) p(q)} \tag{2}$$

The posterior probability logarithm score based on frames is obtained by summing all frames in paragraph I, as shown in formula (3):

$$p_i = \sum_{q=\Gamma}^{\Gamma+1-1} \log\left(p\left(q_i \left| x_t \right.\right)\right) \tag{3}$$

The score of likelihood based on speech knowledge recognition is shown in formula (4):

$$M_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg\left[P\left(s_t \left| s_{t-1} \right.\right) P\left(o_t \left| s_t \right.\right)\right] \tag{4}$$

The score based on the posterior probability of speech knowledge recognition is shown in formula (5):

$$M_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg P\left(q \left| o_t \right.\right) = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg \frac{P\left(o_t \left| q \right.\right) P(q)}{P\left(o_t \left| q_i \right.\right) P\left(q_i\right)} \tag{5}$$

2.3 Analysis of spoken English pronunciation signals

The quality assessment and signal feature extraction are carried out by using the spoken English signal analysis method. The original input feature sequence of spoken English pronunciation is set as $x = [x(0), \cdots, x(N-1)]$, and the knowledge recognition transformation of the spoken English pronunciation feature sequence of X is shown in formula (6):

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j \frac{2\pi}{N} nk\right) \tag{6}$$

In the formula, $0 \le K \le n-1$, $X = DFT\{X\}$ is used to represent the DFT of finite time series X of oral English pronunciation features, as shown in formula (7):

$$X = [X(0), \cdots, X(N-1)] \tag{7}$$

For integer N_0, N_1 intelligent voice input signal $v_0(n)$, $v_1(n)$ length is: $C_j(k) = [x(t), \varphi_{j,k}(t)]$.

Voice control and multimedia control technology are adopted to analyse speech features combined with expert system analysis method. The relationship between the input and output of the automatic evaluation system for spoken English pronunciation quality is shown in formula (8):

$$Z^N = g \cdot X^N + W^N \tag{8}$$

where $Z^N = (z_1, z_2, \cdots, z_N)^H$, $X^N = (x_1, x_2, \cdots, X_N)^H$, $W^N = (w_1, w_2, \cdots, w_N)^H$ and $N = 1, 2, \ldots$ are random variables, and the relationship between linear relations is shown in formula (9):

$$H\left(P_e^E\right) + P_e^E \log\left|S^N\right| \ge H\left(S^N \left| \widehat{S}_E^N \right.\right) \ge NR - N\varepsilon \tag{9}$$

2.4 Signal filtering pre-processing

The signal component phase rotation technique is adopted for linear coding, and the rotational moment of inertia of the output speech signal is obtained as shown in formula (10):

$$angle\left(gX^N\right) = \left(angle\left(X^N\right) + \varphi_g\right) \bmod (2\pi) \tag{10}$$

The positive correlation characteristic quantity of speech signal output is shown in formula (11):

$$gX^N = |g| R^N \tag{11}$$

According to formulas (10) and (11), it is obtained as shown in formula (12):

$$Z^N = |g| R^N + W^N \tag{12}$$

3 Correct pronunciation mistakes

3.1 Phoneme error detection

Error detection relies on phoneme recognition outcomes and phoneme associations, leveraging the textual context of the learned speech. By employing a two-step process that forcefully associates one phoneme recognition, a method can be devised to detect three types of errors outlined above. Figure 2 and Figure 3 illustrate the association and recognition results (the first three words) and error detection for an example sentence ‘Good to see you again’.

Figure 2 Results of phoneme association (see online version for colours)

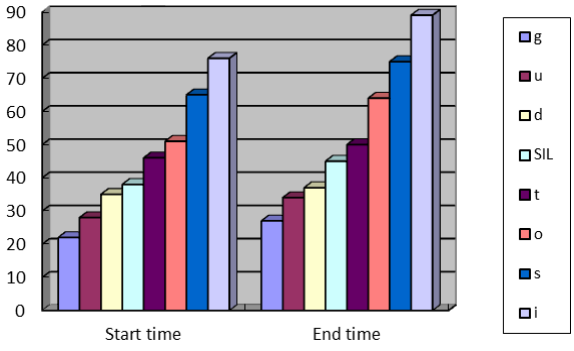
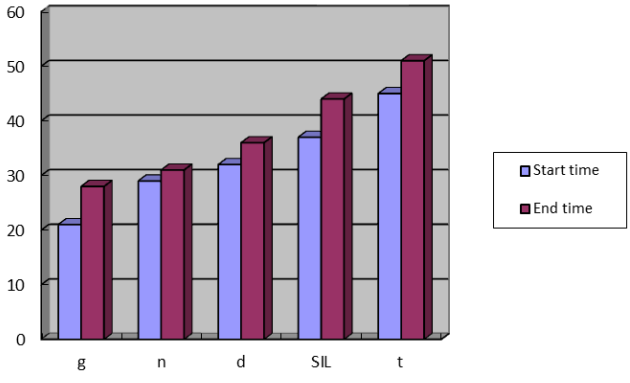


Figure 3 Results of phoneme recognition (see online version for colours)



Among the errors: 'ʊ' became ə five times, and 's' became θ three times. From the correct phoneme 'tʊ' to 'tən', there are mispronouncing 'ə' and inserting 'n'; From 'si:' to 'tθI:' there is insertion of 't' and misreading of 'θ'.

3.2 Correct feedback

Corrective feedback involves utilising phonemes for correction and providing corresponding feedback. Correction feedback for phoneme pronunciation: Through data collection, phonemes are compared, and among the 40 factors analysed, up to 400 cases of incorrect pronunciation may be identified. For instances of missing or added phonemes, appropriate feedback can be retrieved from a corrected knowledge base, enabling learners to undertake self-correction and improvement. Correction feedback for sentence prosody: This aspect aims to allow learners to hear the correct rhythm of English spoken in their own voices, providing encouragement and facilitating easier identification of issues compared to the original pronunciation. It involves modifying the prosody of words, incorporating features such as fundamental frequency and duration. However, both reference speech and learning speech require manual annotation. For automatic modification algorithms, time correlation results can serve as a form of automatic annotation. The modification process comprises two main components: Correction of correlation annotation: when both reference and learning speeches are

accurately annotated, the fundamental frequency of the synchronous overlay result closely resembles that of the reference speech. However, the accuracy of phoneme time correlation results may not reach the level of manual labelling, leading to discrepancies between synthetic speech and expectations, which may sound unnatural. Algorithms described earlier can effectively detect instances where learners miss phonemes. When this occurs, the associated phoneme in the text will be deleted and re-associated, correcting the phoneme association result. Speech synthesis: In speech synthesis, the pitch synchronisation overlay algorithm is widely employed for prosody modification. This method boasts high computational efficiency and preserves the learner's timbre information by splicing waveforms in the time domain while simultaneously adjusting pitch and duration. The performance of our proposed algorithm is compared with existing speech recognition systems and language learning tools. This comparison evaluates the algorithm's effectiveness in terms of accuracy, user engagement, and learning outcomes. The results of this analysis provide insights into the unique advantages of our approach and its potential to address the challenges faced by current oral English self-study solutions. To provide personalised pronunciation guidance, future research can focus on developing machine learning algorithms that analyse users' pronunciation patterns and offer customised improvement suggestions. Moreover, natural language processing techniques can be utilised to mimic language teachers' feedback, providing users with more natural and personalised guidance. The establishment of online communities and peer-learning platforms can facilitate mutual assistance and experience sharing among users.

4 Simulation experiment

4.1 Data collection

Three-hundred sentences of daily English dialogues were collected from existing English texts, and ten independent learners were chosen to recite them. The learners were instructed to maintain the original pronunciation level as closely as possible, ensuring minimal basic errors. For the standard reference speech, an English-speaking teacher was selected to provide voice recordings, which served as the benchmark for comparison. The acoustic model used for scoring was made publicly available for use.

4.2 Experimental results

Detection errors due to incorrect pronunciation or missing phonemes can have a significant impact. During the testing process, errors in the selection of phonemes are more prevalent compared to existing standard oral English. Among these errors, there are 705 types of law-related errors and 670 types related to reading leakage issues. Figure 4 illustrates the phonemes least prone to misreading, as well as those most susceptible to misreading.

Based on the findings depicted in Figure 5, it is evident that diphthongs and consonants are particularly prone to mispronunciation. This can be attributed to the inherent differences in pronunciation between these phonetic elements, as well as variations in pronunciation habits between Chinese and English. Additionally, consonants, characterised by their brief duration, are often challenging to enunciate

clearly. Consequently, it is advisable to establish distinct thresholds for different spoken languages to address these challenges effectively.

Figure 4 The least mispronounced phonemes (see online version for colours)

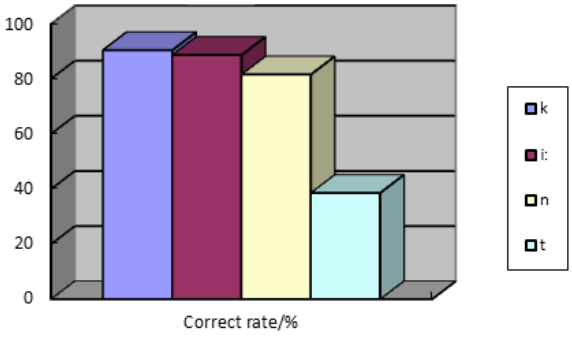
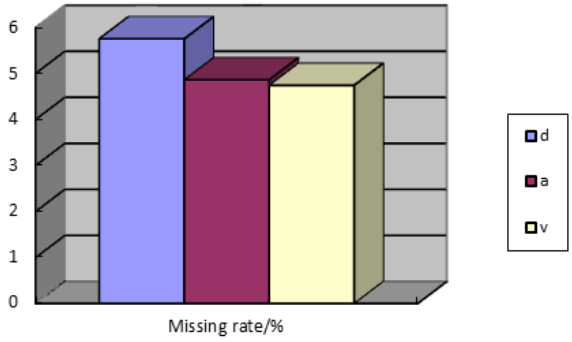


Figure 5 Phonemes are the easiest to skip (see online version for colours)



As indicated in Figure 5, based on statistical findings, the most frequently overlooked factor may not be captured due to incorrect reading but rather due to incomplete pronunciation or inadequate duration, leading to omission by the computer. Hence, it is imperative to permit such occurrences in the actual error detection process to enhance the corresponding accuracy rate.

5 Conclusions

As China continues to embrace globalisation, proficiency in oral English has become increasingly essential for effective communication with the international community. Addressing this growing demand, this paper introduces a knowledge recognition algorithm designed to enhance oral English proficiency. By organising spoken English self-study systems and constructing evaluation indices based on acoustics, rhythm, and comprehension, this study evaluates both subjective and objective factors. Leveraging phonetic knowledge recognition algorithms, the paper establishes a scoring mechanism aimed at promoting autonomous learning of oral English and improving learning effectiveness and accuracy. Through simulation experiments, the efficacy of the speech

knowledge recognition algorithm is demonstrated, showcasing its ability to support the evaluation of spoken English self-learning systems. When dealing with personal data, privacy and data security are of utmost importance. To protect user privacy, researchers can adopt end-to-end encryption techniques to ensure the security of data during transmission and storage. Furthermore, differential privacy techniques can protect individual privacy by injecting noise into the data while maintaining its statistical properties. User consent and transparent data processing procedures are also key to enhancing user trust. The conclusion will further emphasise the potential of the speech knowledge recognition algorithm proposed in this study to enhance the efficiency and accuracy of English oral self-study. Through simulation experiments, we have demonstrated the effectiveness of the algorithm in evaluating spoken English self-learning systems. Future work will focus on translating these findings into practical teaching practices and exploring how to further optimise the algorithm to meet a broader range of learning needs and environments.

Acknowledgements

The research is supported by: The 2021 Jiangsu Province Higher Education Reform Research Project 'Research and Teaching Practice on the Construction of English Curriculum System for the Training of Internationalised Skilled Talents' (No. 2021JSJG682).

The 2021 annual project of the 14th Five Year Plan for Education Science in Jiangsu Province, titled 'Research on the Path of Internationalised Talent Cultivation in Higher Vocational Colleges under the Background of the' Double High Plan 'Construction' (No.: D/2021/03/86).

References

Afshar, H.S. (2021) 'Task-related focus-on-forms foreign language vocabulary development: focus on spoken form and word parts', *System*, Vol. 96, No. 3, pp.102–106.

Allan, R. (2018) 'Lexical bundles from one century to the next: an analysis of language input in English teaching texts', *Journal of Historical Pragmatics*, Vol. 19, No. 2, pp.167–185.

Arizavi, S., Yazdan, C. et al. (2019) 'To use or not to use the shorter forms: a corpus-based analysis of the apologetic expressions 'sorry and I'm sorry' in American spoken English discourse', *Corpus Pragmatics*, Vol. 3, No. 1, pp.21–47.

Born, C.D., Divaris, K., Zeldin, L.P. et al. (2016) 'Influences on preschool children's oral health-related quality of life as reported by English and Spanish-speaking parents and caregivers', *Journal of Public Health Dentistry*, Vol. 5, No. 3, pp.1–8.

Cao, Q. and Hao, H. (2021) 'Optimization of intelligent English pronunciation training system based on android platform', *Complexity*, Vol. 2021, No. 4, pp.1–11.

Jeanjean, T., Stolowy, H., Erkens, M. et al. (2015) 'International evidence on the impact of adopting English as an external reporting language', *Journal of International Business Studies*, Vol. 46, No. 2, pp.180–205.

Kang, J. (2021) 'Automatic translation of spoken English based on improved machine learning algorithms', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 3, No. 11, pp.736–740.

Lavid, J. and Moratón, L. (2015) 'Intersubjective positioning and thematisation in English and Spanish: a contrastive analysis of letters to the editor', *Nordic Journal of English Studies*, Vol. 14, No. 1, pp.289–319.

Lee, G.G., Lee, H.Y., Song, J. et al. (2017) 'Automatic sentence stress feedback for non-native English learners', *Computer Speech & Language*, Vol. 41, No. 4, pp.29–42.

Lin, L., Liu, J., Zhang, X. et al. (2021) 'Automatic translation of spoken English based on improved machine learning algorithm', *Journal of Intelligent and Fuzzy Systems*, Vol. 40, No. 2, pp.2385–2395.

Mante-Estacio, M.J. and Bernardo, A. (2015) 'Illusory transparency in bilinguals: does language of text affect bilingual readers' perspective taking in reading?', *Current Psychology*, Vol. 34, No. 4, pp.744–752.

Mishchenko, O. (2019) 'Thematic magazines: alternative method to control learning outcomes of future foreign language teachers', *Journal of Education and e-Learning Research*, Vol. 6, No. 4, pp.69–73.

Oyewole, O. (2017) 'Influence of mother tongue in the teaching and learning of English language in selected secondary schools in Ondo State, Nigeria', *Journal of Education and Practice*, Vol. 6, No. 2, pp.109–118.

Tantiwich, K. and Sinwongsuwat, K. (2019) 'Thai university students' use of yes/no tokens in spoken interaction', *English Language Teaching*, Vol. 12, No. 3, pp.1–10.

Techentin, C., Cann, D.R., Lupton, M. et al. (2021) 'Sarcasm detection in native English and English as a second language speakers', *Canadian Journal of Experimental Psychology*, Vol. 5, No. 5, pp.190–198.

Toledo, P., Eosakul, S.T., Grobman, W.A. et al. (2015) 'Primary, spoken language and neuraxial labor analgesia use among hispanic medicaid recipients', *Survey of Anesthesiology*, Vol. 60, No. 4, pp.152–160.

Wang, N., Zhang, X. and Sharma, A. (2021) 'A research on HMM based speech recognition in spoken English', *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, Vol. 14, No. 2, pp.79–84.

Wang, Y. (2021) 'Detecting pronunciation errors in spoken English tests based on multifeature fusion algorithm', *Complexity*, Vol. 2, No. 4, pp.62–70.

Williams, J.T., Darcy, I. and Newman, S.D. (2016) 'The beneficial role of L1 spoken language skills on initial L2 sign language learning: cognitive and linguistic predictors of M2L2 acquisition', *Studies in Second Language Acquisition*, Vol. 3, No. 4, pp.1–18.

Yu, P., Pan, Y., Li, C. et al. (2016) 'User-centred design for Chinese-oriented spoken English learning system', *Computer Assisted Language Learning*, Vol. 29, No. 5, pp.984–1000.