JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 17, No. 3, September 2023, pp. 93-99

A Study on the Recognition of English Pronunciation Features in Teaching by Machine Learning Algorithms

Wei Xiong*
School of International Studies and Trade, Jiangxi University of Engineering, Xinyu, China
vw10pn@yeah.net

Abstract

A better understanding of students' English pronunciation features would be a useful guide for teaching spoken English. This paper first analyzed the English pronunciation features and extracted Mel-frequency cepstral coefficients (MFCC) features from the pronunciation signal. Then, the support vector machine (SVM) method was used to identify the cases of incorrect and correct pronunciation. To further improve the recognition effect, deep features were extracted using deep brief network (DBN) as the input of the SVM, and the parameters of both DBN and SVM were optimized by the sparrow search algorithm (SSA). Experiments were conducted on the dataset. The results showed that the MFCC-SSA-SVM algorithm had better recognition performance than the MFCC-SVM algorithm. The DBN-SVM algorithm had higher recognition correctness and accuracy than the MFCC-SSA-SVM algorithm, while the SSA-DBN-SVM method had 88.07% correctness and 85.49% accuracy, indicating the best performance. The results demonstrated the reliability of the proposed method for English pronunciation feature recognition; therefore, it can be applied in practical spoken language teaching.

Category: Natural Language Processing

Keywords: Machine learning; English pronunciation; Feature recognition; Pronunciation error; Support vector machine

I. INTRODUCTION

Under the influence of globalization, English, as a widely used language, is increasingly becoming a second language learning option for many people [1]. English learning has also always been a very important element in university teaching [2]. However, due to the limitations of the learning environment and teaching conditions, many English as a second language learners have significant deficiencies in English speaking [3]. In oral English learning, correct pronunciation is very important, but in the current learning environment, it is difficult for teachers to teach students individually; consequently, many students do not have sufficient knowledge of their pronunciation characteristics, whether correct or incorrect. With the application of computers and other technologies in language teaching [4], students can learn the differences between their English pronunciation and the standard pronunciation by repeatedly listening and comparing; consequently, they can correct and improve. Speech recognition is one of the key technologies in computer-assisted spoken language teaching [5]. With the advancement of algorithms such as machine learning, more methods have been applied in speech recognition [6]. To study mild cognitive impairment

Open Access **http://dx.doi.org/10.5626/JCSE.2023.17.3.93** **http://jcse.kiise.org**

Received 01 June 2023; **Accepted** 30 August 2023
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

recognition, Toth et al. [7] extracted parameters such as pauses, rhythm, and utterance length from speech signals and used machine learning algorithms to identify healthy controls and mild cognitive impairment patients, achieving an F1-score of 78.85%. Based on Mel-frequency cepstral coefficients (MFCC) features of the speech signal, Yousaf et al. [8] used a hidden Markov Model to identify 15 deaf children aged 7–13 years and achieved an accuracy of 97.9%. Ravanelli et al. [9] designed a recurrent neural network (RNN)-based method for speech recognition. Through experiments on a modified RNN model, it was found that the model improved the recognition accuracy. Pakoci et al. [10] studied the speech recognition of Serbian language and adjusted the parameters of a neural network using an n-gram language model. They carried out an experiment and found that an eight-layer deep neural network with 625 neurons had the best performance. This paper focuses on the identification of English pronunciation features through machine learning algorithms, to understand students' pronunciation errors and thus achieve better results in teaching spoken English.

II. ENGLISH PRONUNCIATION FEATURE EXTRACTION

Before English pronunciation features are extracted, pre-processing is required.

A. Pre-emphasis

Pre-emphasis is used to balance the high- and low-frequency components of the signal, which is achieved by a digital filter:

$$H(z) = 1 - az^{-1}, \tag{1}$$

where a is the pre-emphasis factor, $0.9 < a < 1.0$.

For voice signal $x(n)$, the pre-emphasis process is:

$$y(n) = x(n) - ax(n-1). \tag{2}$$

B. Framing and Windowing

Before extracting features, the signal needs to be divided into frames, and then features are extracted from the short speech frames. Meanwhile, to maintain the continuity of adjacent signals, there is an overlap between frames, which becomes frame shifts. The number of frames of a signal can be expressed as:

$$f(n) = \left|\frac{n-s}{l-s}\right|, \tag{3}$$

where n is the number of signal points and l and s refer to the frame length and frame shift, respectively.

After framing, windows need to be added to the signal. Commonly used window functions include:

1) Rectangular window: $w(n) = \begin{cases} 1, 0 \le n \le N-1 \\ 0, \text{else} \end{cases}$;

2) Hamming window:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, \text{else} \end{cases};$$

3) Hanning window:

$$w(n) = \begin{cases} 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right], 0 \le n \le N-1 \\ 0, \text{else} \end{cases}.$$

Among the above window functions, the rectangular window is easy to lead to the loss of signal details and the Hanning window is the most suitable for random signals. The smoothing ability of the Hanning window is good [11]; therefore, the Hanning window is chosen for windowing.

C. Endpoint Detection

Double threshold endpoint detection method is a commonly used and effective method [12], which determines the signal endpoints by the following two features:

1) Short-time energy: $E(n) = \sum_{m=-\infty}^{\infty}[x(n)w(n-m)]^2$, where $w(n)$ is the window function. Let $h(n) = w^2(n)$, then $E(n) = \sum_{m=-\infty}^{\infty} x^2(n)h(n-m) = x^2(n)h(n)$.

2) Short-time average zero-crossing ratio: $Z(n) = \sum_{m=-\infty}^{\infty} sgn[x(m)] - sgn[x(m-1)]$, where sgn is the sign function:

$$sgn[x(n)] = \begin{cases} 1, x(n) \ge 0 \\ -1, x(n) < 0 \end{cases}. \tag{4}$$

After pre-processing, features are extracted for subsequent English pronunciation recognition. In speech recognition, linear predictive cepstral coefficients (LPCC) [13] and MFCC [14] are commonly used. MFCC fully simulates the auditory characteristics of the human ear [15] and has low computational complexity; therefore, MFCC is chosen as the English pronunciation feature in this paper.

For speech signal $x(n)$, a fast Fourier transform (FFT) is needed [16]:

$$x(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}nk}, k = 0,1,2,\cdots,N-1, \tag{5}$$

where N is the frame length.

The MFCC feature uses Mel frequencies. The conversion from linear to Mel frequencies can be written as:

$$Mel(f) = 2597lg\left(1 + \frac{f}{700}\right). \tag{6}$$

Modulo operation is performed on $x(k)$ to obtain signal

amplitude spectrum $|x(k)|$. Then, it is filtered by a triangular filter bank:

$$F(l) = \sum_{k=f_o(l)}^{f_h(l)} w_l(k)|x(k)|, l = 1,2,\cdots,L, \quad (7)$$

$$w_l(k) = \begin{cases} \frac{k-f_o(l)}{f_c(l)-f_o(l)}, f_o(l) \le k \le f_c(l) \\ \frac{f_h(l)-k}{f_h(l)-f_c(l)}, f_c(l) \le k \le f_h(l) \end{cases}, \quad (8)$$

$$f_o(l) = \frac{o(l)}{f_s/N}, \quad (9)$$

$$f_h(l) = \frac{h(l)}{f_s/N}, \quad (10)$$

$$f_c(l) = \frac{c(l)}{f_s/N}, \quad (11)$$

where L stands for the number of filters, $w_l(k)$ stands for the filter coefficient of the filter, $o(l)$, $c(l)$, and $h(l)$ are lower, center, and upper limit frequencies of the filter, respectively, and f_s is the sampling rate.

The final MFCC obtained is:

$$M(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log F(l) \cos\left[\left(l-\frac{1}{2}\right)\frac{i\pi}{L}\right], i = 1,2,\cdots,Q, \quad (12)$$

where Q is the MFCC order.

III. SUPPORT VECTOR MACHINE-BASED RECOGNITION ALGORITHM

A. Support Vector Machine Algorithm

Support vector machine (SVM) is a kind of binary classification model [17], which has the advantages of simple structure and easy implementation, and a wide range of applications in data classification [18] and prediction [19]. The correctness or incorrectness of English pronunciation is also a kind of binary classification problem, so this paper implements the recognition of English pronunciation features by an SVM.

For data set (x_i, y_i), $i = 1, 2, ..., N$, $y_i \in \{-1, 1\}$, its classification decision function can be written as: $\langle w \cdot x \rangle + b = 0$, where w is the weight value and b is the bias. To make the classification interval $\frac{\|w\|^2}{2}$ maximum, the classification problem is transformed into:

$$\min_{w,b} \frac{\|w\|^2}{2}, \quad (13)$$

$$s.t.\, y_i[\langle w \cdot x_i \rangle + b] \ge 1. \quad (14)$$

Then, the classification decision function is written as:

$$f(x) = sgn(w \cdot x + b). \quad (15)$$

When linearly inseparable, relaxation variable ζ_i, and penalty function C are introduced. Then, $\min_{w,b,\zeta} \frac{\|w\|^2}{2} + C\sum_{i=1}^{N} \zeta_i$ is obtained. The nonlinear problem is converted to a linear problem using the Lagrangian transformation. It is solved:

$$maxL(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j \cdot y_i y_j k(x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i, \quad (16)$$

$$\sum_{i=1}^{N} \alpha_i\, y_i = 0, \quad (17)$$

where α is the Lagrangian multiplier and $k(x_i \cdot x_j)$ is the kernel function, the radial basis function (RBF) here:

$$k(x_i, x_j) = exp\left(-\gamma\left\|(x_i - x_j)\right\|^2\right). \quad (18)$$

The SVM is applied to the recognition of English pronunciation features to obtain the MFCC-SVM algorithm, i.e., the MFCC features extracted from the signal are used as the input to the SVM, and the recognition of incorrect and correct English pronunciation can be achieved by training the algorithm.

B. Deep Belief Network

To further improve the effectiveness of SVM in recognizing English pronunciation features, a method combining a deep belief network (DBN), i.e., the DBN-SVM method, is designed to further extract the hidden information in English pronunciation features as the input of SVM.

DBN has good performance in deep feature extraction [20] and has good applications in image recognition [21], and text classification [22], among others. DBN is obtained by stacking several restricted Boltzmann machines (RBM) [23]. It is assumed that there is a visible layer v and hidden layer h, then the energy function of the RBM can be written as:

$$E(v, h|\theta) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_i h_i - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j, \quad (19)$$

where $\theta = \{a_i, b_j, w_{ij}\}$ refers to the parameter of RBM, a_i is the bias of the i-th neuron, b_j is the bias of the j-th neuron, and w_{ij} is the link weight of the i-th and j-th neurons.

DBN can obtain more abstract and deeper features through learning of features, and its training consists of two main steps:

1) Positive pre-training: The input data is used as the visible layer to generate the hidden layer, and the contrast scatter (CD) algorithm is used to update the parameters of the RBM. The cycle is repeated to

obtain the parameter set of the forward DBN.

2) Reverse tuning: The optimal model is obtained by top-down tuning of the RBM using a reverse back-propagation network and fine-tuning the parameters obtained in the first step using the errors.

C. Parameter Optimization Method

Both DBN and SVM have some parameters that need to be determined. To improve the effectiveness of English pronunciation feature recognition, this paper uses the sparrow search algorithm (SSA) to find the optimal parameters of DBN and SVM.

SSA is an optimization algorithm based on the foraging behavior of sparrows [24]. In the *d*-dimensional space, suppose there is a sparrow population:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1d} \\ s_{21} & s_{22} & \cdots & s_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nd} \end{bmatrix},$$

and its fitness value is:

$$F_x = \begin{bmatrix} f([s_{11} & s_{12} & \cdots & s_{1d}]) \\ & \vdots & & \\ f([s_{n1} & s_{n2} & \cdots & s_{nd}]) \end{bmatrix}.$$

In the population, the discoverer is responsible for finding the direction of food, the joiner is responsible for finding more food in the location found by the discoverer, and the watcher monitors the behavior of other individuals and makes contention. The equation for the discoverer's location update can be written as:

$$S_{i,j}^{t+1} = \begin{cases} S_{i,j}^{t} \cdot exp\left(\frac{-i}{\alpha T_{max}}\right), if\ A < ST \\ S_{i,j}^{t} + Q \cdot L, if\ A \geq ST \end{cases}. \quad (20)$$

$A < ST$ indicates that the surroundings of the finder are safe, while $A \geq ST$ indicates that the surrounding area is not safe and the location needs to be changed. The meanings of the other parameters are as follows: t stands for the number of iterations, α is a random number in (0,1], T_{max} represents the maximum number of iterations, A is an alarm threshold value in [0,1], ST represents the safety threshold value in [0.5,1], Q represents a random number with a normal distribution, and L represents a horizontal matrix with element 1.

The formula for updating the position of the joiner can be written as:

$$S_{i,j}^{t+1} = \begin{cases} Q \cdot exp\left(S_{worst}^{t} - S_{i,j}^{t}\right), if\ i > \frac{n}{2} \\ S_{p}^{t+1} + \left|S_{i,j}^{t+1} - S_{p}^{t+1}\right| \cdot B^{+} \cdot L, if\ i \leq \frac{n}{2} \end{cases}, \quad (21)$$

where S_{worst}^{t} stands for the worst position globally, S_{p}^{t+1} is the optimal location of the discoverer, and $B^{+} = B^{T}(BB^{T})^{-1}$.

The monitor's position update formula is:

$$S_{i,j}^{t+1} = \begin{cases} S_{best}^{t} + \beta \cdot \left|S_{i,j}^{t} - S_{best}^{t}\right|, if\ f_i \neq f_g \\ S_{i,j}^{t} + C \cdot \left(\frac{\left|S_{i,j}^{t} - S_{worst}^{t}\right|}{(f_i - f_w) + \varepsilon}\right), if\ f_i = f_g \end{cases}, \quad (22)$$

where S_{best}^{t} indicates the globally optimal position, β indicates the step control parameter, f_i indicates the fitness value of sparrow i, f_g indicates the fitness value of the optimal individual, C is the direction of movement of the sparrow, whose value range is [-1,1], f_w is the fitness value of the worst individual, and ε is the smallest constant that avoids the denominator being zero.

SSA finds the optimal parameters by continuously updating the sparrow positions and outputs the global optimum when the maximum number of iterations is reached.

IV. RESULTS AND ANALYSIS

A. Experimental Setup

The dataset used for the experiment was L2-ARCTIC [25], which included English speaking audio of 24 people. The speaking content came from the CMU-ARCTIC dataset and was processed by phoneme balance. Every audio was approximately one hour long. Three doctors of philosophy in linguistics performed phoneme pronunciation annotation and mispronunciation annotation on the audios. The pronunciation errors include the following three types.

1) Replacement: The standard phoneme pronunciation is replaced by the wrong phoneme pronunciation.
2) Deletion: The pronunciation of a standard phoneme is missing.
3) Insertion: An extra standard phoneme is pronounced.

To further expand the training set, the TIMIT dataset [26] was used, which included English speaking audios of 630 people. There was a total of 6,300 sentences. The experimental dataset is shown in Table 1.

The evaluation indicators included:

$$correct = \frac{N - A - B}{N}, \quad (23)$$

Table 1. Experimental dataset

	Number of speakers	Number of sentences
Training set		
TIMIT	630	6,300
L2-ARCTIC	18	2,699
Test set		
L2-ARCTIC	6	900

$$accuracy = \frac{N-A-B-C}{N}, \quad (24)$$

where N is the total number of phonemes, and A, B, and C are the pronunciation errors of replacement, deletion, and insertion, respectively.

B. Analysis of Results

First, the number of hidden layers of DBN and the number of nodes in each layer were determined. The accuracy of recognition in different cases was calculated by layer-by-layer experiments, and the results are presented in Table 2.

As shown in Table 2, it was found that when the number of hidden layers was 3, the recognition accuracy was the highest, reaching 83.69%, while the accuracy decreased by 0.48% when the number of hidden layers was increased to 4. Therefore, in the following experiments, the number of hidden layers of DBN was set to 3, and the number of nodes in each layer was 10-10-10.

Then, the SVM-based recognition methods using MFCC as the feature were compared. The performance of the MFCC-SVM and MFCC-SSA-SVM methods in recognizing English pronunciation features is shown in Fig. 1.

As shown in Fig. 1, first, in terms of the correct rate of English pronunciation feature recognition, the correct rate of the MFCC-SVM method was 77.49%, and after optimization by SSA, the correct rate of the MFCC-SSA-SVM method reached 84.58%, which was an improvement of 7.09% compared with the MFCC-SVM method. Then, in terms of the accuracy of English pronunciation feature recognition, the accuracy of the MFCC-SVM method was 75.68%, and after optimization by SSA, the accuracy of MFCC-SSA-SVM reached 80.12%, which was an improvement of 4.44% compared with MFCC-SVM. These findings indicated that the optimization of the parameters of SVM using SSA could improve the correct rate and accuracy of English pronunciation feature recognition.

Then, the recognition effect of the SVM was compared when using DBN to extract features, and the results are shown in Fig. 2.

As shown in Fig. 2, first, when the parameters of DBN and SVM were optimized without using SSA, the correct rate of the DBN-SVM method was 84.55% and the accuracy was 81.27%, both of which were higher than 80%. As shown in Figs. 1 and 2, the correct rate and accuracy of the DBN-SVM method were improved by 0.58% and 1.15%, respectively, compared with the MFCC-SSA-SVM method, which indicated that the DBN feature extraction had a good effect on improving the recognition of English pronunciation features. Then, after parameter optimization by SSA, the correct rate of the SSA-DBN-SVM method reached 88.07%, which was 2.91% higher than that of the DBN-SVM method, and the accuracy reached 85.49%, which was 4.22% higher than that of the DBN-SVM method. The results showed that using SSA for parameter optimization and extracting deep features by DBN before using SVM for identification was the most effective.

Table 2. Recognition accuracy under different numbers of hidden layers

	Hidden layer				
	1	2	3	4	5
Number of nodes in each layer	10	10-10	10-10-10	10-10-10-10	10-10-10-10-10
Recognition accuracy (%)	80.12	82.77	83.69	83.21	82.07

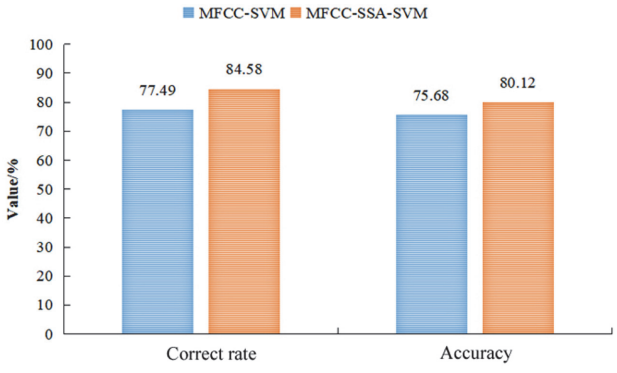

Fig. 1. Comparison of the SVM-based recognition methods with MFCC as a feature.

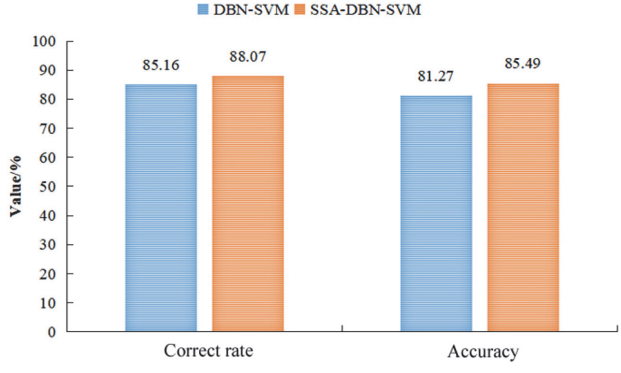

Fig. 2. Comparison of the recognition effect between the SVM methods when using DBN to extract features.

V. CONCLUSION

This paper mainly explored the recognition methods of English pronunciation features in teaching and designed several SVM-based recognition methods. The experimental analysis results showed that the SSA-DBN-SVM method had the highest correct rate and accuracy, 88.07% and 85.49%, respectively, which proved the role of parameter optimization with SSA and DBN feature extraction. The proposed method is expected to significantly improve the recognition effect of English pronunciation features, thus helping teachers to better understand students' pronunciation and provide targeted instructions in teaching.

Conflict of Interest(COI)

The authors have declared that no competing interests exist.

REFERENCES

1. M. Bennett, "Book Review: Improving library services in support of international students and English as a second language learners. Edited by Leila Jun Rod-Welch. Chicago: ALA, 2019," *The Library Quarterly*, vol. 90, no. 4, pp. 586-589, 2020. https://doi.org/10.1086/710266
2. H. Zhang, "On cultural teaching in college English," *World Scientific Research Journal*, vol. 6, no. 3, pp. 205-209, 2020. https://doi.org/10.6911/WSRJ.202003_6(3).0028
3. J. Xue, B. Li, R. Yan, J. R. Gruen, T. Feng, M. F. Joanisse, and J. G. Malins, "The temporal dynamics of first and second language processing: ERPs to spoken words in Mandarin-English bilinguals," *Neuropsychologia*, vol. 146, article no. 107562, 2020. https://doi.org/10.1016/j.neuropsychologia.2020.107562
4. X. Y. Chai and G. Subramaniam, "The use of communication strategies in mobile asynchronous chat," *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 11, no. 2, pp. 33-50, 2021. https://doi.org/10.4018/IJCALLT.2021040103
5. Q. Cao and H. Hao, "Optimization of intelligent English pronunciation training system based on Android platform," *Complexity*, vol. 2021, article no. 5537101, 2021. https://doi.org/10.1155/2021/5537101
6. M. Ravanelli and M. Omologo, "Automatic context window composition for distant speech recognition," *Speech Communication*, vol. 101, pp. 34-44, 2018. https://doi.org/10.1016/j.specom.2018.05.001
7. L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, M. Pakaski, and J. Kalman, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130-138, 2018. https://doi.org/10.2174/1567205014666171121114930
8. K. Yousaf, Z. Mehmood, T. Saba, A. Rehman, M. Rashid, M. Altaf, and Z. Shuguang, "A novel technique for speech recognition and visualization based mobile application to support two-way communication between deaf-mute and normal peoples," *Wireless Communications and Mobile Computing*, vol. 2018, article no. 1013234, 2018. https://doi.org/10.1155/2018/1013234
9. M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92-102, 2018. https://doi.org/10.1109/TETCI.2017.2762739
10. E. Pakoci, B. Popovic, and D. J. Pekar, "Improvements in Serbian speech recognition using sequence-trained deep neural networks," *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 53-76, 2018.
11. N. Tomen and J. C. van Gemert, "Spectral leakage and rethinking the kernel size in CNNs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada 2021, pp. 5118-5127. https://doi.org/10.1109/ICCV48922.2021.00509
12. G. Wang and W. Wang, "Bowel sound signal identifcation of whole abdomen based on voice endpoint detection," *Chinese Journal of Medical Instrumentation*, vol. 43, no. 2, pp. 90-93, 2019. https://doi.org/10.3969/j.issn.1671-7104.2019.02.004
13. Y. B. Wang, D. G. Chang, S. R. Qin, Y. H. Fan, H. B. Mu, and G. J. Zhang, "Separating multi-source partial discharge signals using linear prediction analysis and isolation forest algorithm," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 6, pp. 2734-2742, 2020. https://doi.org/10.1109/TIM.2019.2926688
14. T. Y. Badgujar and V. P. Wani, "Wavelet transform and mel-frequency cepstral coefficient-based feature extraction of the sheet metal trimming process to study burr formation," *International Journal of Mechatronics and Manufacturing Systems*, vol. 15, no. 1, pp. 20-36, 2022. https://doi.org/10.1504/IJMMS.2022.122906
15. S. R. Hasibuan, R. Hidayat, and A. Bejo, "Speaker recognition using mel frequency cepstral coefficient and self-organising fuzzy logic," in *Proceedings of 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2020, pp. 52-55. https://doi.org/10.1109/ISRITI51436.2020.9315337
16. A. Eghtesad, K. Germaschewski, and M. Knezevic, "Coupling of a multi-GPU accelerated elasto-visco-plastic fast Fourier transform constitutive model with the implicit finite element method," *Computational Materials Science*, vol. 208, article no. 111348, 2022. https://doi.org/10.1016/j.commatsci.2022.111348
17. B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *European Journal of Operational Research*, vol. 265, no. 3, pp. 993-1004, 2018. https://doi.org/10.1016/j.ejor.2017.08.040
18. H. Wimalarathna, S. Ankmnal-Veeranna, C. Allan, S. K. Agrawal, P. Allen, J. Samarabandu, and H. M. Ladak, "Comparison of machine learning models to classify auditory brainstem responses recorded from children with auditory processing disorder," *Computer Methods and Programs in Biomedicine*, vol. 200, article no. 105942, 2021. https://doi.org/10.1016/j.cmpb.2021.105942
19. S. Khademolqorani, "Quality mining in a continuous production line based on an improved genetic algorithm fuzzy support vector machine (GAFSVM)," *Computers & Industrial Engineering*, vol. 169, article no. 108218, 2022. https://doi.org/10.1016/j.cie.2022.108218
20. H. Jang, S. M. Plis, V. D. Calhoun, and J. H. Lee, "Task-specific feature extraction and classification of fMRI volumes using

a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks," *NeuroImage*, vol. 145, pp. 314-328, 2017. https://doi.org/10.1016/j.neuroimage.2016.04.003
21. H. Yalcin, "Plant recognition based on deep belief network classifier and combination of local features," in *Proceedings of 2021 29th Signal Processing and Communications Applications Conference (SIU)*, Istanbul, Turkey, 2021, pp. 1-4. https://doi.org/10.1109/SIU53274.2021.9477879
22. J. Gao, J. Yi, W. Jia, and X. Zhao, "Improved deep belief network to feature extraction in Chinese text classification," in *Proceedings of 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018, pp. 283-287. https://doi.org/10.1109/ICSESS.2018.8663827
23. S. S. Athalye and G. Vijay, "Taylor series-based deep belief network for automatic classification of diabetic retinopathy using retinal fundus images," *International Journal of Imaging Systems and Technology*, vol. 32, no. 3, pp. 882-901, 2022. https://doi.org/10.1002/ima.22691
24. S. Zhang, L. Zhang, T. Gai, P. Xu, and Y. Wei, "Aberration analysis and compensate method of a BP neural network and sparrow search algorithm in deep ultraviolet lithography," *Applied Optics*, vol. 61, no. 20, pp. 6023-6032, 2022. https://doi.org/10.1364/AO.462436
25. G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: a non-native English speech corpus," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*, Hyderabad, India, 2018, pp. 2783-2787.
26. J. H. Hansen, A. Stauffer, and W. Xia, "Nonlinear waveform distortion: assessment and detection of clipping on speech data and systems," *The Journal of the Acoustical Society of America*, vol. 144, no. 3_Supplement, pp. 1871-1871, 2018. https://doi.org/10.1121/1.5068230

Wei Xiong https://orcid.org/0009-0005-0381-8371

Wei Xiong was born in February 1977, graduated from China Central Radio & TV University, majoring in English. He received a master's degree from Hunan Normal University with a major in educational management. Since July 2007, He has been engaged in English teaching as a lecturer in Jiangxi University of Engineering. His main research directions are English teaching and language teaching.