J Inf Process Syst, Vol.20, No.4, pp.491~500, August 2024
https://doi.org/10.3745/JIPS.02.0217
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Creation of a Voice Recognition-Based English Aided Learning Platform

Hui Xu*

Abstract

In hopes of resolving the issue of poor quality of information input for teaching spoken English online, the study creates an English teaching assistance model based on a recognition algorithm named dynamic time warping (DTW) and relies on automated voice recognition technology. In hopes of improving the algorithm's efficiency, the study modifies the speech signal's time-domain properties during the pre-processing stage and enhances the algorithm's performance in terms of computational effort and storage space. Finally, a simulation experiment is employed to evaluate the model application's efficacy. The study's revised DTW model, which achieves recognition rates of above 95% for all phonetic symbols and tops the list for cloudy consonant recognition with rates of 98.5%, 98.8%, and 98.7% throughout the three tests, respectively, is demonstrated by the study's findings. The enhanced model for DTW voice recognition also presents higher efficiency and requires less time for training and testing. The DTW model's KS value, which is the highest among the models analyzed in the KS value analysis, is 0.63. Among the comparative models, the model also presents the lowest curve position for both test functions. This shows that the upgraded DTW model features superior voice recognition capabilities, which could significantly improve online English education and lead to better teaching outcomes.

Keywords

DTW, English, Endpoint Detection, MFCC, Online Teaching, Voice Recognition

1. Introduction

Due to the unique nature of language learning, English teaching relies more on oral teaching and training. Therefore, in the teaching environment of online learning, the teaching of English subjects also requires higher input quality of voice information [1-3]. Compared to face-to-face English oral teaching, online oral teaching has more obvious characteristics of personalization and portability. Students can repeatedly listen to and follow up on English oral teaching materials according to their own needs, and can also adopt more diverse multimedia cognitive methods for learning [4-6]. And this further enhances the requirements of English teaching platforms for the quality of oral speech input. Once there is distortion or inaccuracy in the input of oral teaching information, it can lead to students following incorrect learning examples in the process of autonomous learning, which in turn leads to a decrease in the quality of students' learning [7-9].

Therefore, this study aims to design an English assisted learning platform based on speech recognition

Manuscript received June 12, 2023; first revision August 11, 2023; accepted August 26, 2023.
* **Corresponding Author:** Hui Xu (happy317xuhui@163.com)
College of Basic Education, Zhumadian Preschool Education College, Zhumadian, China

technology from the perspective of the quality of speech information input in online English oral teaching, providing reliable guarantees for students' autonomous English learning. This study aims to address the issue of traditional teaching methods leading to students' lack of interest in learning English phonetic symbols and poor teaching effectiveness. In order to provide reliable guarantees for students' autonomous English learning, this project aims to create an English learning platform based on speech recognition technology from the perspective of the quality of speech information input in online English oral teaching. By designing and developing an English speech learning tutoring platform based on speech recognition technology, and utilizing speech recognition technology to provide correct and incorrect pronunciation feedback, this study combines the development of mobile learning mechanisms and mobile internet technology to implement a learning platform through speech recognition interaction.

2. Related Works

With Prastikawati's team [7] using digital English language teaching technology to a new generation of students who are increasingly dependent on digital products, research on online English language teaching has improved recently. By analyzing students' perspectives and the results of their use of the English learning platforms, the study evaluated the efficacy of digital English language teaching platforms. The authors of [8] enhanced students' independent learning through campus-based applications by using online social media as a language teaching tool. Via an unstructured data survey, the study examined the pre- and post-learning status of the students. The study's findings demonstrated that online platforms for autonomous learning can assist students in honing their skills while lightening the strain on teachers. Sun et al. [9] integrated artificial intelligence algorithms to an English teaching aid system, resulting in a hybrid decision tree algorithm and neural network. A teaching quality assessment model for online English instruction developed by Huang [10] used Bayesian and hybrid Gaussian learning approaches to carefully choose and label samples in order to achieve classification and, as a result, a high level of teaching quality evaluation.

In hopes of reducing the relative error rate in voice recognition by adjusting to the speaker’s stress and pitch range, Bell et al. [11] suggested an end-to-end adaptive voice recognition system that combines hidden Markov models with neural network models. The study's findings demonstrated the algorithm's efficacy. Haeb-Umbach et al. [12] created a distant recognition method that enables automatic Voice Recognition even when the microphone is not nearby. When recording from a distance, this method can handle distortion.

Recent studies have shown that, despite their rapid development, voice recognition and online English education technologies are not yet fully integrated. In reality, speech input is one of the most natural forms of input used in the teaching process, and spoken English in particular, is highly dependent on input. In order to enhance the quality of teaching information input through language recognition technology and to assist online English teaching, this study blends voice recognition technology with online teaching technology.

3. Design of a Voice Recognition-Based English-Assisted Learning Platform

3.1 Pre-processing and Identification of Voice Signal

Speech recognition technology uses speech as the input method, and machines recognize and understand speech content through speech. This technology converts the input voice signal into corresponding text commands, enabling the machine to control through voice. Speech feature extraction and speech pattern matching are the two main modules in Fig. 1, which reflect the recognition concepts used by almost all speech recognition systems when processing speech input.

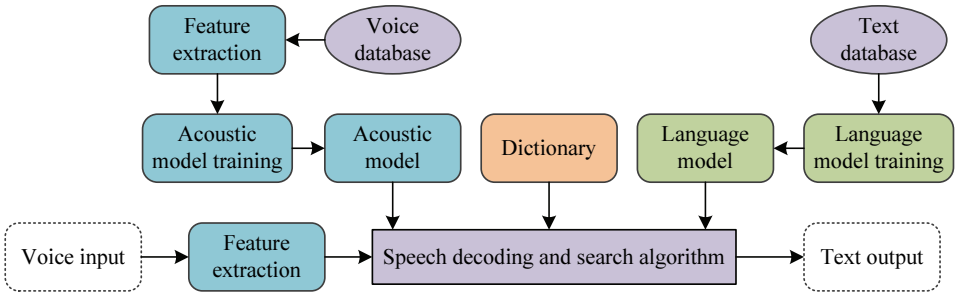

Fig. 1. Voice recognition system principle.

To achieve the greatest similarity impact, standard speech is compared with the feature parameters that were retrieved during the speech signal pre-processing step and can capture the substance of speech. The speech signal pre-processing module typically consists of five components: endpoint detection, framing, pre-emphasis, pre-emphasis, and speech signal digitization. The process of digitizing entails quantizing and sampling a spoken input to create a discrete digital signal. The standard procedure involves passing the voice signal through a first order digital filter with transfer coefficients stated in equation that has a 6 dB/oct boost of high frequency characteristics:

$$H(z) = 1 - \alpha z^{-1}. \tag{1}$$

In Eq. (1), the pre-emphasis factor α has a value between and (0.9, 1.0). When processing frames, a specific window function is employed to process the voice signal in order to lessen the impact of the time between frames. The rectangular window, the Hamming window, and the Hanning window are frequently used window types. Equation displays the expression for a rectangle window:

$$W_r = \begin{cases} 1, (0 \le n \le N-1) \\ 0, (Others) \end{cases}. \tag{2}$$

The expression for the Hamming window is shown in Eq. (3):

$$W_{hm} = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, (Others) \end{cases}. \tag{3}$$

The two methods usually used are short-time energy detection and short-time over-zero rate detection. Short-time energy is a response to the change of speech energy with time, let the short-time energy of the first n-frame x_n of the speech signal is E_n, its calculation equation is as in Eq. (4). The reaction to a change in speech energy over time is referred to as short-time energy. Let's suppose the calculation equation for the short-time energy E_n of the speech signal's initial n-frame x_n is Eq. (4):

$$E_n = \sum_{m=0}^{N-1} x_n^2(m), 0 \le m \le N-1. \tag{4}$$

In Eq. (4), N denotes the frame length. In contrast, the number of times a discrete signal with two different sampling signs crosses the time axis is equal to the over-zero rate of a continuous speech signal, i.e. the number of times the time domain signal crosses the time axis. The equation gives the definition of the short time over-zero rate of a speech signal:

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} \left| \operatorname{sgn}[x_n(m)] - \operatorname{sgn}[x_n(m-1)], \right. \tag{5}$$

where sgn $[x]$ represents the sign function. The recognition performance, noise resistance, and robustness of Mel frequency cepstral coefficient (MFCC) feature parameters are good, so it is selected for speech recognition comparison. The MFCC can fully characterize the auditory features of the human ear and simulate the perception of speech at different frequencies.

The relationship between MFCC and frequency is shown in Eq. (6):

$$f_{MFCC} = 2595 \log_{10} \left(1 + \frac{f}{700}\right). \tag{6}$$

Generally, it is converted into energy distribution in the frequency domain for analysis, and different energy distributions represent the characteristics of different speech. Each frame must undergo a fast Fourier transform in the hope of obtaining the total energy on the frequency spectrum shown in Eq. (7):

$$x(i,k) = FFT[x_i(m)]. \tag{7}$$

By taking the square of the mode of the speech signal's spectrum as in equation, the spectral energy of the speech signal is obtained:

$$E(i,k) = |x(i,k)|^2 \tag{8}$$

That according Eq. (9), the frequency domain energy spectrum $E(i,k)$ of each frame is multiplied by the frequency domain response $H_m(k)$ of the Mel filter to determine the energy passing through the filter.

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k) \tag{9}$$

The MFCC coefficients are obtained by the discrete cosine transform (DCT) as in Eq. (10):

$$MFCC\ (i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} log\ [S(i,m)] cos\ \left(\frac{\pi n(2m-1)}{2M}\right), n = 1,2,\cdots,L \quad (10)$$

where L stands for the MFCC coefficient step in Eq. (10), which accepts values in the range $[12,16]$. The energy of the Mel filter, $S(i,m)$, is obtained from step 4. The m-th filter's symbol is m; i stands for the i-th frame; and n stands for the spectral line following DCT.

3.2 Design of an English-Assisted Learning Platform based on Improved DTW Recognition Algorithm

The dynamic time warping (DTW) algorithm is a classical voice recognition algorithm, the basic principle of which is to elongate or shorten the unknown quantity until it is the same length as the reference template. Assume that the U sequence is the reference template and the V sequence is the test template, and that there are two time sequences $U = u_1, \dots u_i, \dots u_l$, $V = v_1, \dots v_j, \dots v_h$, of lengths l and h, respectively. As a result, the value of each point in the series is the feature value of each frame in the speech sequence. When l and h are not equal, then the two speech sequences need to be aligned. In order to compare the two speech segments, the two time series need to be aligned and a matrix grid of $l * h$ constructed. By plotting each frame of the standard signal on the horizontal axis of the rectangular coordinate system and each frame of the speech signal to be compared on the vertical axis of the rectangular coordinate system, a grid diagram can be constructed from the frame numbers in the two sets of data. The intersection element (i,j) on each grid represents the distance d between u_i and v_j, the smaller the distance the higher the similarity, and is generally expressed by the Euclidean distance.

After analysis it can be seen that the path passing through this solid line can be defined as a search path for DTW, denoted by P. The k-th element of P is defined as $p_k = (i,j)_k$, thus defining a mapping of two sequences, $P = p_1, p_3, \dots p_k$, $max(u,v) \le k \le u + v - 1$. This path needs to satisfy the boundary conditions, continuity constraints and monotonicity constraints. A maximum of three possible paths for each point when the monotonicity and continuity constraints are taken into account, and the path with the lowest regularization cost is chosen from those that do, as shown in Eq. (11):

$$DTW(U,V) = min \frac{\sqrt{\sum_{k=1}^{k} pk}}{K}. \quad (11)$$

The compensation for pathways with differing regularity lengths is shown by K in Eq. (11). In order to match the two sequences U and V, establish a cumulative distance starting at $(0,0)$. The distances obtained at each point will be added together to reach the end point (l,k). Eq. (12) states that the ultimate cumulative distance gained equals the path's total distance.

$$h(l,k) = min \begin{cases} h(i-1,j) + d(i,j) \\ h(i-1,j-1) + 2d(i,j). \\ h(i,j-1) + d(i,j) \end{cases} \quad (12)$$

It was discovered that the enhanced DTW search path instead scans the parallelogram region bounded by the two slopes rather than the complete matrix region of the graph. Two points, x_a and x_b, are computed through this parallelogram region, achieving the best results in terms of search speed and

similarity. Fig. 2 depicts the modified DTW algorithm's search path.

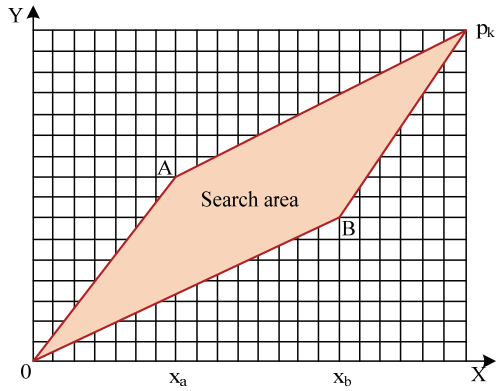

Fig. 2. Search path of improved DTW algorithm.

In Fig. 2, the actual dynamic bending of the path is split into three segments, $(1, x_a)$, $(x_a + 1, x_b)$, and $(x_b + 1, N)$. Both x_a and x_b are considered to be the closest integers. If this requirement is not met, the difference is considered too large for dynamic bend matching. The platform is being developed primarily on the mobile side, using the Android operating system as the development environment, and Asp.net for development on the online side. The platform's operations are divided into numerous categories, including speech diagnosis, listening to phonetic symbols, correct and incorrect words, words and speech instruction.

4. The Effectiveness of the Application of Voice Recognition-based English Assisted Learning Platform

The study employs the phonetic diagnostic function as an illustration and bases its classification on a library of typical speech samples. From the standard speech sample library, 120 speech samples were chosen, of which 90 were chosen as the training set and the remaining 30 as the test set. The test was conducted in a quiet laboratory to reduce the effect of external noise. Fig. 3 shows the results of the platform's recognition rate test.

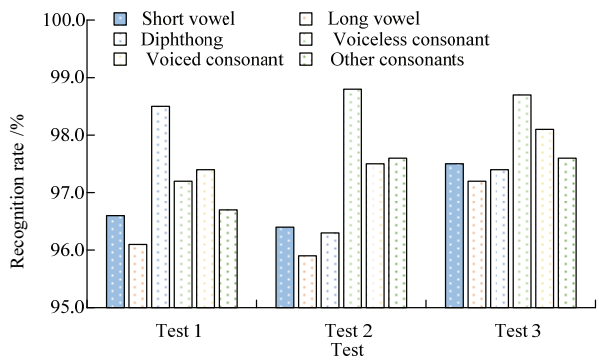

Fig. 3. Identification rate test of the platform.

In Fig. 3, the model has a high recognition rate for all speech symbols, reaching over 95%. In these three tests, the recognition rates reached 98.5%, 98.8%, and 98.7%, respectively. The comparison results of training and testing time are shown in Fig. 4.

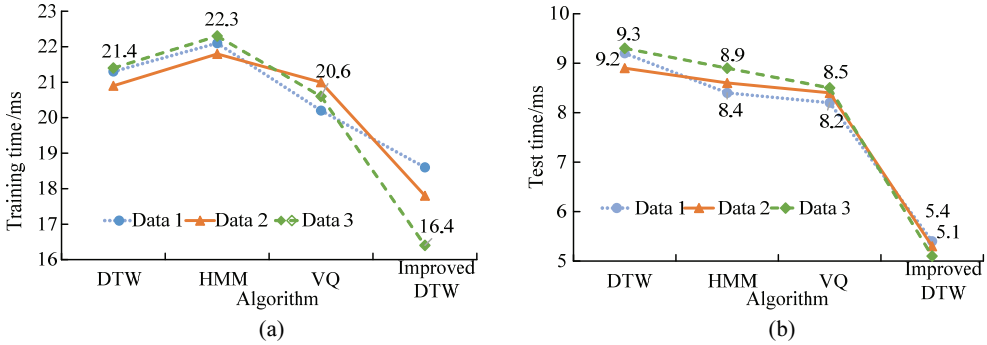

Fig. 4. Comparison results of (a) training time and (b) test time.

In Fig. 4, the training time of the model is shorter, with 18.6 ms, 17.8 ms, and 16.4 ms, respectively. The testing time is 5.4 ms, 5.3 ms, and 5.1 ms, respectively, which is the lowest value in the comparative model. It can be seen that improving the DTW model is more effective. The precision-recall (PR) curve and receiver operating characteristic (ROC) curve are shown in Fig. 5.

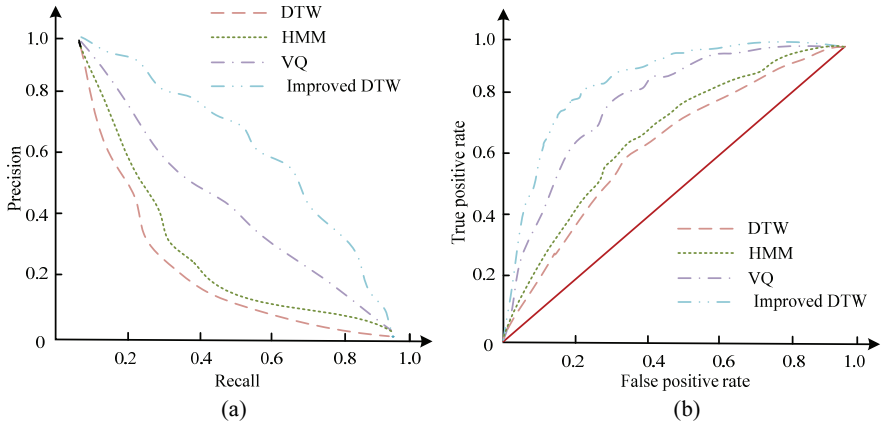

Fig. 5. Comparison results of (a) PR curve and (b) ROC curve.

In Fig. 5, the study compares the improved DTW model with several other models. The PR curve and ROC curve of the improved DTW model are located above the curves of other models, indicating that the improved DTW model has higher confidence and accuracy.

The study compared the convergence of the model under test functions, with Sphere and Schwefel2.22 functions selected for test functions F1 and F2, as shown in Fig. 6.

In Fig. 6, as the number of iterations increases under F1, all four models show a certain downward trend, with the VQ model and the improved DTW model showing a more significant downward trend. The improved DTW model shows a decreasing trend of fast initial speed and gradually entering a gradual easing stage in the later stage, with the overall curve position consistently maintaining the lowest state

and the best convergence effect. Under F2, as the number of iterations increases, the four models show a decreasing trend of faster initial speed and slower later speed, with VQ model and improved DTW model showing a more significant downward trend. The improved DTW model has a faster and larger descent speed, and the overall curve position remains at its lowest state, with the same optimal convergence effect. Overall, the improved DTW model designed through research has the best overall performance and can maintain better speech recognition performance.

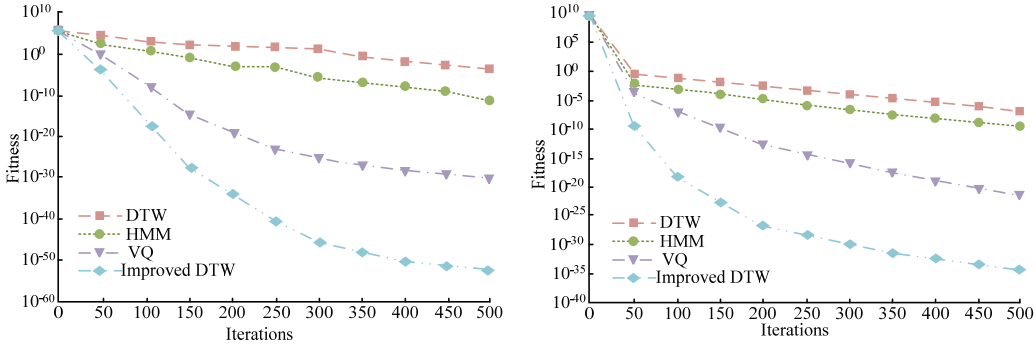

Fig. 6. Model convergence under (a) F1 test and (b) F2 test.

5. Conclusion

A better DTW voice recognition model is suggested in the study in response to the low input quality of spoken instructional materials used in online English teaching. In order to enhance the model's voice recognition quality and efficiency, the model's function and performance are also optimized from two angles, namely pre-processing and computing approaches. The outcomes demonstrated that the enhanced DTW voice recognition model had a high recognition rate for all phonetic symbols, with the greatest recognition rates in three tests being 98.5%, 98.8%, and 98.7%. The training times for the enhanced DTW voice recognition model were shorter for all three datasets, 18.6 ms, 17.8 ms and 16.4 ms, respectively, while the test times were 5.4 ms, 5.3 ms and 5.1 ms, respectively, which were also the lowest values among the comparison models. The area under the curve of the modified DTW voice recognition model was larger and the model shown increased confidence and accuracy in the PR curve and ROC curve analysis. The revised DTW model has the highest KS value of 0.63 among the comparison models, according to the KS value examination. The modified DTW model also has the lowest convergence curve position, the fastest rate of decline, and the greatest decline for both F1 and F2 test functions. From a voice recognition point of view, the modified DTW model in the study has the best overall performance and can be a sufficient aid for online English teaching.

Conflict of Interest

The author declare that they have no competing interests.

Funding

None.

References

[1] W. Hu, "A study on the scoring method of oral English test in college English online computer test," in Application of Big Data, Blockchain, and Internet of Things for Education Informatization *(BigIoT-EDU).* Cham, Switzerland: Springer, 2022, pp. 25-36. https://doi.org/10.1007/978-3-031-23950-2_4

[2] D. Zhang and P. Perez-Paredes, "Chinese postgraduate EFL learners' self-directed use of mobile English learning resources," *Computer Assisted Language Learning*, vol. 34, no. 8, pp. 1128-1153, 2021. https://doi.org/10.1080/09588221.2019.1662455

[3] R. R. F. Sinaga and R. Pustika, "Exploring STUDENTS'ATTITUDE towards English online learning using Moodle during COVID-19 pandemic at SMK Yadika Bandarlampung," *Journal of English Language Teaching and Learning*, vol. 2, no. 1, pp. 8-15, 2021.

[4] T. N. L. Ly, T. L. Nguyen, and H. N. Nguyen, "Using E-learning platforms in online classes: a survey on tertiary English teachers' perceptions," *AsiaCALL Online Journal*, vol. 12, no. 5, pp. 34-53, 2021.

[5] S. S. Fuentes Hernandez and A. N. S. Florez, "Online teaching during Covid-19: How to maintain students motivated in an EFL class," *Linguistics and Literature Review*, vol. 6, no. 2, pp. 157-171, 2020. https://doi.org/10.32350/llr.v6i2.963

[6] R. W. Todd, "Teachers' perceptions of the shift from the classroom to online teaching," *International Journal of TESOL Studies*, vol. 2, no. 2, pp. 4-16, 2020. https://doi.org/10.46451/ijts.2020.09.02

[7] E. F. Prastikawati, W. Wiyaka, and M. Y. W. Lestari, "Secondary school students' perception on Edmodo as online learning platform in English learning," *Language Circle: Journal of Language and Literature*, vol. 16, no. 2, pp. 296-307, 2022. https://doi.org/10.15294/lc.v16i2.33712

[8] L. D. Nguyen and L. V. Nguyen, "Schoology as an online learning platform to enhance English language ability for undergraduates in Vietnam," *Computer-Assisted Language Learning Electronic Journal*, vol. 23, no. 4, pp. 139-161, 2022.

[9] Z. Sun, M. Anbarasan, and D. Praveen Kumar, "Design of online intelligent English teaching platform based on artificial intelligence techniques," *Computational Intelligence*, vol. 37, no. 3, pp. 1166-1180, 2021. https://doi.org/10.1111/coin.12351

[10] W. Huang, "Simulation of English teaching quality evaluation model based on Gaussian process machine learning," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2373-2383, 2021. https://doi.org/10.3233/JIFS-189233

[11] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: an overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33-66, 2020. https://doi.org/10.1109/OJSP.2020.3045349

[12] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124-148, 2021. https://doi.org/10.1109/JPROC.2020.3018668

Hui Xu **https://orcid.org/0009-0000-6719-2455**

She obtained Master of Education in English from Central China Normal University in 2010. She is currently a lecturer at Zhumadian Preschool education College. She was awarded as the famous teacher of Ideology and Politics in Henan Province and the academic technology leader in Zhumadian City. She once presided over the planning project of Henan Province and won the first prize in the Teaching Skills Competition of Henan Province. She has edited three textbooks and published seven papers in CN journals, including one in the Peking University Core Journal. Her area of interest is English education.