frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 22 January 2025
DOI 10.3389/fpsyg.2024.1484630

OPEN ACCESS

EDITED BY
Adrian Pasquarella,
University of Delaware, United States

REVIEWED BY
Sharif Alghazo,
The University of Jordan, Jordan
Zdena Kralova,
Constantine the Philosopher University, Slovakia

*CORRESPONDENCE
Li Ping
✉ mnxpxib@163.com
Ning Tao
✉ ningtao@guet.edu.cn

RECEIVED 22 August 2024
ACCEPTED 12 November 2024
PUBLISHED 22 January 2025

CITATION
Ping L and Tao N (2025) Innovative approaches to English pronunciation instruction in ESL contexts: integration of multi-sensor detection and advanced algorithmic feedback.
Front. Psychol. 15:1484630.
doi: 10.3389/fpsyg.2024.1484630

Innovative approaches to English pronunciation instruction in ESL contexts: integration of multi-sensor detection and advanced algorithmic feedback

Li Ping[1]* and Ning Tao[2]*

[1]School of Foreign Languages, Jiangsu Ocean University, Lianyungang, China, [2]School of Computer Engineering, Guilin University of Electronic Technology, Beihai, China

Introduction: Teaching English pronunciation in an English as a Second Language (ESL) context involves tailored strategies to help learners accurately produce sounds, intonation, and rhythm.

Methods: This study presents an innovative method utilizing advanced technology and algorithms to enhance pronunciation accuracy, fluency, and completeness. The approach employs multi-sensor detection methods for precise data collection, preprocessing techniques such as pre-emphasis, normalization, framing, windowing, and endpoint detection to ensure high-quality speech signals. Feature extraction focuses on key attributes of pronunciation, which are then fused through a feedback neural network for comprehensive evaluation. The experiment involved 100 college students, including 50 male and 50 female students, to test their English pronunciation.

Results: Empirical results demonstrate significant improvements over existing methods. The proposed method achieved a teaching evaluation accuracy of 99.3%, compared to 68.9% and 77.8% for other referenced methods. Additionally, students showed higher levels of fluency, with most achieving a level of 4 or above, whereas traditional methods resulted in lower fluency levels. Spectral feature analysis indicated that the amplitude of speech signals obtained using the proposed method closely matched the original signals, unlike the discrepancies found in previous methods.

Discussion: These findings highlight the effectiveness of the proposed method, showcasing its ability to improve pronunciation accuracy and fluency. The integration of multi-sensor detection and neural network evaluation provides precise results, outperforming traditional approaches.

KEYWORDS

accuracy, English as a second language, English pronunciation, feedback neural network, speech signal processing, teaching evaluation

1 Introduction

English pronunciation instruction has consistently been a significant focus in the ESL setting. This emphasis is driven by the understanding that accurate pronunciation is crucial for effective communication and integration into English-speaking contexts (Azimova and Solidjonov, 2023). Conventional methods of teaching English pronunciation typically prioritize phonetic symbols and pronunciation norms. These traditional approaches often involve the use of phonetic charts (Alghazo et al., 2023), repetitive drills, and the memorization of pronunciation rules, which aim

to familiarize learners with the standard sounds of the English language. Because learners' mother tongue background, language context, learning habits and motivation are different, traditional pronunciation teaching methods are often difficult to meet the needs of all learners (Clymer et al., 2020). The theory of second language acquisition provides us with important theoretical support. This theory holds that language acquisition is a complex process involving multiple factors such as the learner's age, cognitive ability, learning motivation, and the quality and quantity of language input. Among them, the quality and quantity of language input have a significant impact on the accuracy of pronunciation (Alghazo et al., 2019). Extensive exposure and use of the target language, especially through interaction with native speakers, can help learners better perceive and imitate the pronunciation characteristics of English. However, in ESL contexts, learners often lack sufficient language input and practical opportunities. They do not provide tailored assistance for individual variations and accent replication (Yaccob et al., 2023). Every learner comes with a unique set of phonetic challenges influenced by their native language, individual speech habits, and personal learning pace. For instance, a Spanish speaker might struggle with English vowel sounds, while a Chinese speaker might find it difficult to pronounce consonant clusters. Traditional methods often fail to address these specific issues, leading to persistent pronunciation errors and frustration among learners. Furthermore, conventional instruction does not adequately account for accent replication. Learners might be able to produce isolated sounds correctly but still struggle with the fluid, natural intonation patterns, stress, and rhythm that characterize native speech (Adelekea and Onyebuchib, 2023). This gap in teaching can result in speech that, while technically correct, still sounds unnatural or stilted to native speakers. Therefore, there is a pressing need for innovative instructional methods that go beyond the one-size-fits-all approach. These methods should incorporate personalized guidance, allowing educators to address the specific phonetic challenges and learning styles of individual students. Moreover, they should facilitate accent replication by integrating practices that focus on the suprasegmental aspects of pronunciation, such as intonation, stress, and rhythm, ensuring that learners can achieve a more natural and fluent English speech (Gao et al., 2023). Many learners face difficulties with English pronunciation due to differences in their native language contexts and congenital phonetic conditions. Traditional teaching methods often do not address these unique challenges. Therefore, exploring innovative approaches, such as advanced technology and targeted pronunciation techniques, is crucial for developing effective, personalized instruction and improving learners' pronunciation skills (Priya and Kumar, 2020). The innovative teaching method that combines modern technology, related algorithms, personalized guidance, and real-time feedback offers a promising solution for overcoming pronunciation barriers. By using tools like speech recognition systems and interactive phonetic training software, educators can provide immediate, detailed feedback, allowing learners to quickly correct errors and improve their pronunciation. This approach tailor's instruction to individual phonetic challenges and linguistic backgrounds, enhancing pronunciation accuracy and boosting confidence. As learners master accurate sound production, intonation, and stress patterns, they can integrate more effectively into English-speaking contexts, leading to improved fluency and overall communication effectiveness. Currently, English pronunciation teaching methods primarily include recording and playback, visual aids, and electronic learning tools. However, these approaches have notable limitations. While recording and playback enable students to compare their pronunciation with models, they lack the immediacy of real-time feedback and corrective guidance. Visual aids can illustrate oral morphology, but non-native speakers often struggle with imitation despite these visual representations. The advancement of electronic learning tools has significantly enhanced personalized pronunciation instruction; however, the accuracy of some systems still requires refinement. This indicates that current English pronunciation teaching methods have their limitations (Liu, 2021). In order to improve students' pronunciation accuracy and fluency, these methods should incorporate advanced technological solutions to offer personalized guidance and real-time feedback, thereby addressing the specific needs of learners and improving their English pronunciation skills in the ESL context.

1.1 Literature review

Wen (2020) introduced a method for automatically correcting English pronunciation errors. The system is based on the DTW algorithm and involves optimizing voice recognition sensors and improving speech recognition processors to construct the hardware. Develop and extract parameters that indicate errors in English pronunciation using the English pronunciation collecting tool, and finalize the software design of the system. An automatic correction solution for English pronunciation issues has been developed using the DTW algorithm. However, this study lacks evaluation and analysis of English pronunciation quality, which hinders the effective improvement of pronunciation teaching. Tuba et al. (2018) explored data mining and clustering, focusing on web intelligence, a field facing challenges due to complex, dynamic, unstructured, and large data. Traditional clustering methods were insufficient, prompting the authors to propose a novel approach combining the bare bones fireworks algorithm and k-means clustering. This hybrid method aimed to improve web intelligence data clustering. Li (2020) analyzed the drawbacks of traditional multimedia teaching systems and highlighted the benefits of network-based systems. He focused on designing an intelligent central control multimedia teaching system and discussed using advanced technologies like remote control and streaming media in teaching practices, demonstrating their potential to transform educational contexts efficiently and effectively. Niu and Wei (2023) aimed to improve speech denoising accuracy and robustness by proposing an adaptive noise reduction method based on noise classification. They created a new acoustic feature matrix combining LogFbank features and perceptual linear prediction coefficients and designed a noise classifier using a support vector machine. To address "music noise" issues in traditional methods, they adaptively updated the voice activity detection threshold based on background noise types and optimized parameters for a noise estimation algorithm. Li and Wu (2023) explored the impact of emerging technologies like cloud computing, mobile computing, and AI on higher education, emphasizing the need for digital transformation. As China's higher education system reached popularization, they highlighted the importance of digitalization to meet the demands for diverse quality, lifelong learning, personalized training, and modern governance. To address these needs, Li and Wu developed an embedded voice teaching system using cloud computing and a deep learning model. This system aimed to improve existing university teaching methods and enable ubiquitous learning. They integrated Hidden Markov Models (HMM) and Long Short-Term Memory (LSTM) networks to enhance voice recognition performance, improving recognition rates,

anti-interference, and noise robustness. Experimental tests showed high recognition accuracy and noise immunity, confirming the system's stability and reliability. Feedback from trials indicated that the new system significantly enhanced the intelligence and adaptability of college teaching methods, promoting improvements in students' cultural literacy and innovation abilities. Jiang and Chen (2024) explored the practical implications of machine-assisted translation, addressing challenges in parameter acquisition and model architecture. Their study used Python data analysis to simulate a translation model, introducing a novel approach that integrated the source syntax tree into the encoder-decoder paradigm. They designed an encoder with a bidirectional GRU RNN, processing syntax tree information from the root node downward. Simulation results showed that for sentences longer than 20 words, their model significantly improved performance, especially for sentences over 25 words. The retrained model quickly restored baseline accuracy with minimal precision loss, demonstrating effective model compression. Bardol (2023) presented a reflective analysis of a pedagogical innovation in applied linguistics through an action research project aimed at improving English pronunciation teaching in a blended learning system at the University of the Antilles in Martinique. Conducted over 3 years (2017–2020), this experiment targeted language students in professional disciplines. The chapter detailed the contextual conditions, key actors, implementation strategies, and obstacles encountered. Bardol also provided reflection tools and recommendations for researchers looking to adapt this innovation to other contexts. Nguyen (2024) explored the rising interest in pronunciation research in English language teaching, focusing on textbook content, teachers' beliefs, and learners' attitudes. Despite evidence supporting communicative pronunciation teaching for improving L2 learners' intelligibility, recent studies showed it remained unsystematic, often relying on recasts and prompts. This inconsistency likely stemmed from a lack of guidance for teachers. Nguyen's chapter critically reflected on an innovation in Vietnamese tertiary EFL classrooms, shifting from error correction to a communicative approach. The chapter detailed the motivation, context, and implementation by six Vietnamese EFL teachers, discussing successes and challenges. It concluded with discussion questions to link with other chapters and understand the innovation drivers. Al-Asi (2024) examined the difficulties and approaches involved in instructing non-native speakers in English pronunciation. This included addressing concerns such as disparities in speech sounds, intonation, patterns of emphasis, and individual variations. The study presented efficacious techniques, including clear instruction, visual and aural assistance, and communicative exercises. Al-Asi highlighted the significance of error correction and feedback in assisting learners in recognizing and rectifying pronunciation faults. The research also addressed the incorporation of pronunciation instruction into a wider English teaching framework, so improving learners' understanding, precision, fluency, and self-assurance. In summary, Al-Asi emphasized the significance of teaching pronunciation in enhancing one's ability to communicate effectively, as well as developing listening and speaking abilities, and overall fluency in the English language. Ozodova (2024) explored the methodologies of teaching pronunciation in ESL, focusing on the challenges learners faced, particularly the transfer of phonetic features from their native languages. She highlighted that pronunciation included both segmental (individual sounds) and suprasegmental (stress, rhythm, intonation) features. Historically, instruction focused on individual sounds, but with the rise of communicative language teaching, the emphasis shifted to suprasegmental elements. Contemporary teaching now recognizes the importance of both for achieving intelligibility. Озодова presented exercises like reverse dictation and sound-focused activities, and analyzed common errors such as the substitution of interdental consonants and mispronunciation of silent letters. She underscored the need for a balanced approach to teaching pronunciation to improve students' overall communicative competence in English. Thi-Nhu Ngo et al. (2024) examined the effectiveness of automatic speech recognition (ASR) on ESL/EFL student pronunciation, analyzing data from 15 studies conducted between 2008 and 2021. They found that ASR had a medium overall effect size ($g = 0.69$). Key findings included that ASR with explicit corrective feedback was highly effective, while indirect feedback showed moderate effectiveness. ASR significantly improved segmental pronunciation but had minimal impact on suprasegmental features. Medium to long treatment durations with ASR yielded better outcomes, whereas short durations were no more effective than non-ASR conditions. Practicing with peers in an ASR setting produced substantial benefits, while practicing alone had smaller effects. ASR was particularly effective for adult learners (18+) and those at an intermediate proficiency level. Ngo et al. concluded that ASR was a valuable tool for L2 pronunciation development and recommended its integration into language learning programs. Yang (2023) underscored the importance of phonetics in language acquisition, asserting that improved pronunciation enhances listening and speaking skills. Despite this, non-English majors often struggled with pronunciation and motivation due to environmental constraints and traditional teaching methods. Yang emphasized the need for innovative phonetic instruction, presenting strategies based on literature reviews and personal experience. These included using real-life scenarios and technology, promoting cooperative learning, and gamifying phonetic instruction. Yang argued that these methods could significantly improve phonetic sensitivity, pronunciation accuracy, listening skills, and oral fluency, thereby creating a more effective learning context for non-English majors.

1.2 Novelty and contributions

The novelty and contributions of this work lie in the development of an innovative method for teaching English pronunciation in an ESL context. The study introduces a comprehensive approach that integrates modern technology and advanced algorithms to provide personalized guidance and real-time feedback. Key contributions include the use of multi-sensor detection methods for accurate data collection and scale decomposition of speech signals, advanced signal processing techniques such as pre-emphasis, normalization, framing, windowing, and endpoint detection to optimize speech signal quality, and the extraction and fusion of features related to accuracy, fluency, and completeness using a feedback neural network. Additionally, the development of a computer-assisted evaluation system that leverages multi-feature fusion results to provide detailed feedback enhances the effectiveness of pronunciation teaching. Empirical validation through experimental results demonstrates that the proposed method surpasses existing techniques in terms of accuracy and application effectiveness, significantly improving students' pronunciation fluency and overall teaching quality. These innovations provide a scientific basis for optimizing and personalizing English pronunciation teaching

methods, making them more effective and adaptable to individual learner needs.

The continuation of the article is as follows:

Section two: Optimization of English Pronunciation Teaching Methods, Section Three: Experimental verification analysis, and Section four: Conclusion.

2 Optimization of English pronunciation teaching methods

In the field of English pronunciation teaching methods (Alghazo, 2015), advanced research techniques have significantly improved the ability to assess and enhance learners' pronunciation skills. One critical advancement is the detection of English pronunciation speech signals, which allows for the precise capture of waveform data from learners' spoken output. This technology enables real-time monitoring and recording of pronunciation behaviors, providing valuable insights into how students produce and articulate sounds (Ruan, 2023). Speech signal preprocessing plays a crucial role in this process, as it ensures the accuracy and stability of the data for subsequent analysis. By cleaning and normalizing the raw speech signals, preprocessing eliminates noise and irregularities, which helps maintain the integrity of the data and supports more reliable analyses. The next step involves feature extraction, where key acoustic properties of the speech signals are analyzed. This stage focuses on identifying and extracting essential features such as pitch, volume, and spectral characteristics (Li and Huang, 2024). These features are critical for understanding the nuances of pronunciation and provide the necessary data for in-depth analysis and evaluation. The extracted speech features are then evaluated to assess various aspects of pronunciation, including accuracy, fluency, and overall quality. This evaluation helps determine the effectiveness of pronunciation teaching methods and offers a detailed assessment of students' pronunciation levels. By analyzing these features, educators can identify specific areas where students may need additional support or improvement (Nagle and Hiver, 2023). Ultimately, this comprehensive process converts complex speech signals into quantifiable feature information, offering a scientific basis for optimizing and personalizing English pronunciation instruction. This data-driven approach supports the development of tailored teaching strategies, ensuring that instruction meets the individual needs of learners and enhances the overall effectiveness of pronunciation teaching methods (Li, 2024).

2.1 English pronunciation speech signal detection

Accurate detection of learners' pronunciation, identification of specific errors, and targeted feedback are crucial for improving English-speaking proficiency. This process starts with constructing a model to detect and analyze English pronunciation signals using advanced technology. Multi-sensor techniques are then used to collect the raw data (Duan and He, 2023a). These techniques employ sensors to capture detailed aspects of spoken English, such as pitch, volume, and articulation. The collected data undergoes scale decomposition to analyze different frequency ranges and temporal aspects of pronunciation. This process isolates specific features of the speech signal, revealing pronunciation patterns and errors. Detailed analysis then identifies and quantifies pronunciation errors, including incorrect articulation, stress patterns, and intonation issues (Lounis et al., 2024). The results offer valuable insights into the accuracy of learners' pronunciation and highlight areas that require improvement. By integrating these steps—model construction, multi-sensor data collection, scale decomposition, and detailed analysis—educators can obtain comprehensive and precise feedback on learners' pronunciation. This feedback is crucial for correcting errors and guiding learners toward more accurate and effective pronunciation. Ultimately, this approach enhances learners' English-speaking skills by providing targeted interventions that address their specific pronunciation challenges (Duan and He, 2023).

The mathematical model expression for the pronunciation and speech signal of spoken English is given in Equation (1) (Weng et al., 2021):

$$h(t) = d(t) + f(t) \otimes u(t) \tag{1}$$

Among them, $d(t)$ represents the amplitude of the received signal; $f(t)$ represents the step transfer function of speech signals; $u(t)$ represents the resonance peak.

Using this model for English spoken pronunciation speech signal detection and recognition, the distribution of the sampled elements of the speech information is z_n, and the echo pulse of the English spoken pronunciation speech signal is represented as Equation (2) (Ma et al., 2021):

$$k(t) = \int s(\phi,\varepsilon)\exp[\pi\phi t] \times z_n y(t-\phi)\,dt \tag{2}$$

Among them, $s(\phi,\varepsilon)$ represents the output extension function of English spoken pronunciation speech signals; $y(t-\phi)$ represents the complex envelope of the frequency components of the English spoken pronunciation speech signal; ε represents the bandwidth of signal acquisition feature expansion; ϕ represents the frequency shift feature of English spoken pronunciation speech signals.

When receiving a linear frequency modulation signal, the result of separating English spoken pronunciation speech features is as Equation (3) (Ma et al., 2021):

$$k'(t) = \int \mu(a,b) \times y(t) k(t) \times \frac{1}{\sqrt{|a|}}\,dt \tag{3}$$

Among them, $y(t)$ represents the instantaneous frequency estimation value of the received English spoken pronunciation speech signal; $\mu(a,b)$ represents the delay component of broadband signal incident on the array element; a represents the high-order statistical feature information of the signal; b represents the frequency shift distribution. At the new cluster head node, the feature components of English spoken pronunciation information obtained are as Equation (4) (Ma et al., 2021):

$$Y_k(t) = \int_{-\infty}^{+\infty} R_k(t) k'(t)\,dt \tag{4}$$

Among them, $R_k(t)$ represents the phase of speech detection.

Update the fusion weights to obtain the output signal component $Y_k^{'}(t)$, which is represented as Equation (5) (Ma et al., 2021):

$$Y_k^{'}(t)=\sqrt{\frac{1-h(t)+Y_k(t)}{2\pi}} \tag{5}$$

Thus, statistical information modeling of English spoken pronunciation speech signals is achieved, and the detection results of English pronunciation speech signals are obtained, providing a foundation for speech signal processing and feature extraction.

2.2 English pronunciation speech signal preprocessing

Preprocessing is a crucial initial step and foundational element in speech comparison and recognition, playing a vital role in the effective extraction of speech signal features. During the preprocessing stage of English pronunciation speech signal detection, as detailed in Section 2.1, it is essential to extract feature parameters that accurately capture the core aspects of English pronunciation. This preprocessing phase encompasses several key processes, including pre-emphasis to enhance signal clarity, normalization to standardize the amplitude, framing to segment the speech signal into manageable pieces, windowing to reduce signal distortion, and endpoint detection to identify the precise start and end points of speech segments. Together, these processes ensure that the speech signal is optimally prepared for accurate and meaningful analysis.

2.2.1 Pre emphasis and normalization of speech signals

The purpose of pre-emphasis is to amplify high-frequency components of a speech signal and achieve a flatter frequency spectrum. This is necessary because the power spectrum of speech signals decreases with increasing frequency, leading to most of the energy being concentrated in the lower frequency range. As a result, high-frequency components often have a much lower signal-to-noise ratio. Various external factors, including the effects of the mouth, nose, lips, and airflow on the vocal cords, can further attenuate high-frequency signals during speech production. Pre-emphasis addresses this issue by boosting these high-frequency components to improve their clarity. This process can be implemented using either analog or digital methods. The method of pre-emphasis using a first-order digital filter is implemented, and the formula is as Equation (6) (Le-Qing, 2011):

$$X(z)=1-\alpha z^{-1} \tag{6}$$

Among them, α represents the pre-emphasis coefficient. If $x[n]$ is a speech signal, the formula for pre-emphasis is presented in Equation (7) (Yegnanarayana and Gangashetty, 2011):

$$y[n]=x[n]-\alpha x[n-1] \tag{7}$$

The purpose of normalization is to avoid negative effects on speech signal processing caused by the strength of the tester's pronunciation or the distance between the tester and the recording device. It fixes speech signals with different amplitudes within a similar range for subsequent processing of speech signals. The normalization formula is presented using Equation (8) (Yegnanarayana and Gangashetty, 2011):

$$x[n]=x[n]/\max\left(x[1,2,3,..,N]\right) \tag{8}$$

Among them, $x[n]$ represents the signal of the n -th sample point of the speech sample; max represents the maximum value of speech sample $x[n]$.

2.2.2 Framing and windowing processing of speech signals

Speech signals are typically continuous and vary over time, making them effectively infinite in length. However, over brief intervals—such as 0 ms to 25 ms—the variations in the speech signal are minimal, allowing these short periods to be approximated as steady-state signals. To analyze the speech signal effectively, it is essential to divide it into frames, which involves segmenting the signal into equal time intervals. This framing ensures that each segment is treated consistently along the timeline. To maintain smooth transitions and continuity between adjacent frames, an overlap is introduced, known as the frame shift. The duration of data within each frame is referred to as the frame length, and the frame shift is typically set to a ratio between 0 and 0.5 of the frame length. The frame rate $f(n)$ can be calculated by the Equation (9) (Quatieri, 2002):

$$f(n)=\frac{|n-s|}{|l-s|} \tag{9}$$

Among them, l represents the frame length; s represents frame shift; n represents the number of voice points.

For the framing operation of speech signals, a moving window function is generally required. Assuming that the speech signal is $X(n)$, this value is infinitely long in practical applications. Therefore, it is necessary to perform framing operation on it. The formula is as Equation (10) (Yegnanarayana and Gangashetty, 2011):

$$X_c(n)=X(n)\times c(n-m) \tag{10}$$

Among them, $X(n)$ represents the speech signal; $c(n-m)$ represents the window function. Generally speaking, choosing different window functions will have different effects on the analysis of speech signals. For example, choosing a wider window function makes the signal smoother, and the window function generally has low-pass characteristics. Choosing different window functions for speech signals will result in different broadband and spectrum leaks. For window functions, rectangular windows are prone to losing the high-frequency part of the signal and the details of the speech signal. However, the advantage is that they have good smoothness, and rectangular windows are generally suitable for time-domain analysis. The smoothness of the Hamming window is better than that of the rectangular window. The Hamming window attenuates the speech

signal at the edge of the window, effectively solving the problem of rectangular windows losing details. Usually, for frequency domain processing of speech signals, the Hamming window is used as the window function. Therefore, when analyzing speech signals, this paper chooses the Hamming window, which can be obtained by stacking the spectra of three rectangular windows, commonly known as the raised cosine window. Its main lobe width reduces frequency resolution. The expression for The Hanning Window (2024) is as Equation (11):

$$w(n)=\begin{cases}0.5\left[1-\cos\left(\dfrac{2\pi}{N-1}\right)\right], & 0\le n\le N-1\\ 0, & other\end{cases} \tag{11}$$

After windowing, the voice signal forms many frames.

2.2.3 Speech signal endpoint detection

Endpoint detection is a fundamental component in speech processing, playing a vital role in both speech recognition and speech comparison tasks. Its importance cannot be overstated, as it serves two primary functions that significantly impact the efficiency and accuracy of the processing workflow. First, endpoint detection is crucial for distinguishing between meaningful speech segments and periods of silence or background noise. By accurately identifying the start and end points of speech within a signal, endpoint detection helps filter out irrelevant data. This process involves removing non-speech elements such as noise and pauses, which often clutter the signal and complicate subsequent analysis (Ahmed and Lawaye, 2023). This filtering is essential for focusing on the actual spoken content, thereby enhancing the quality of the data being analyzed and improving the reliability of the results. Second, effective endpoint detection accelerates the feature extraction process. By clearly defining the boundaries of speech segments, it allows for the extraction of features from only the relevant portions of the signal. This targeted approach not only streamlines the feature extraction process but also reduces computational load and processing time (Liang et al., 2023). As a result, the overall efficiency of the speech processing system is improved, enabling faster and more accurate analysis. In this paper, an endpoint detection algorithm with enhanced noise resistance is utilized. This algorithm integrates two key techniques: short-term energy analysis and spectral subtraction. Short-term energy analysis evaluates the energy levels of the speech signal over short intervals, helping to identify active speech regions. Spectral subtraction, on the other hand, helps to reduce the influence of noise by subtracting an estimate of the noise spectrum from the speech spectrum. The combination of these techniques enhances the algorithm's ability to accurately detect speech endpoints even in noisy contexts, leading to more precise and effective speech processing outcomes.

Divide the speech signal into frames, perform square operations on each frame's signal, and sum up to obtain short-term energy. The formula is as Equation (12) (Huang et al., 2001):

$$E(n)=\Sigma(x_i)^2 \tag{12}$$

Among them, x_i represents the i -th sample point.

Set an energy threshold based on the actual situation to determine whether the current frame is a speech frame. In a short-term energy sequence, find the first point that continuously exceeds the energy threshold as the starting point. In a short-term energy sequence, starting from the starting point and looking back, find the first point that is continuously below the energy threshold as the ending point. To ensure more accurate endpoint position, trace forward or move back a certain number of frames.

Perform Fast Fourier Transform (FFT) on each frame and calculate the amplitude spectrum using Equation (13) (Shen and Wai, 2022):

$$M(n,k)=\left|X(n,k)\right| \tag{13}$$

Among them, $X(n,k)$ represents the complex representation of the k -th frequency point in frame n.

Select frames from non-speech segments, calculate the average value of their spectrum as the noise spectrum, and the formula is as Equation (14) (Shen and Wai, 2022):

$$N(k)=\frac{1}{M\times\sum M(n,k)} \tag{14}$$

Perform spectral subtraction on each frame, subtract the amplitude spectrum of the current frame from the noise spectrum, and take the absolute value of the result to obtain the amplitude spectrum after spectral subtraction. Determine whether the current frame is a speech frame by comparing the average energy of the amplitude spectrum values processed by spectral subtraction with the threshold. In the frames identified as speech frames, find the first and last frames as the starting and ending points, adjust the endpoint position to improve the accuracy of the endpoints.

From the above analysis, it can be seen that pre emphasis can weaken the attenuation of high-frequency signals during transmission, improve the energy of the high-frequency part of speech signals, and help improve the clarity and accuracy of sound; Normalization can eliminate audio amplitude differences between different sentences or utterances, making each segment of audio have the same energy level, making it easier for feature extraction and model training; Framing and windowing can segment continuous speech signals into short frames, while windowing helps eliminate boundary effects and facilitates subsequent feature extraction and analysis; Endpoint detection can accurately determine the starting and ending points of speech signals, avoiding the interference of noise and silence on speech recognition and analysis, thereby improving teaching effectiveness and learning experience. The integration of these processing steps makes English pronunciation speech signals more standardized and stable, providing a better foundation for speech features and signal quality assurance for ESL teaching.

2.3 Feature extraction of speech signals

Building upon the preprocessing of English pronunciation speech signals, this approach involves extracting and analyzing key features—accuracy, fluency, and completeness—to evaluate students'

pronunciation quality. Accuracy assesses how closely students' pronunciation aligns with target phonetic standards, focusing on the correctness of individual sounds and phoneme production. Voice fluency verification reflecting how effortlessly students speak without undue hesitation. Completeness evaluates whether students effectively convey the full intended meaning of their speech, including correct stress, intonation, and rhythm. These features are then fused to provide a comprehensive evaluation of pronunciation performance, offering a nuanced view that goes beyond isolated feature assessments. This integrated evaluation serves as a valuable reference for educators, enabling them to refine instructional strategies, implement targeted practice exercises, and enhance the overall effectiveness of English pronunciation teaching. To comprehensively measure the quality of students' pronunciation, rating features are extracted from three aspects: (1) accuracy features (logarithmic posterior probability, GOP); (2) Fluency characteristics (speech speed, segment duration, and pause duration); (3) Integrity feature (word matching degree).

2.3.1 Accuracy characteristics

Firstly, starting from the levels of logarithmic posterior probability and GOP, the accuracy features of student pronunciation are extracted.

For phoneme q_i, its corresponding observation vector for each frame is O_t, and the frame level posterior probability is defined as Equation (15) (Rabiner and Juang, 1993; Jelinek, 1998):

$$P(q_i|O_t) = \frac{P(O_t|q_i)P(q_i)}{\sum_{i=1}^{M} P(O_t|q_i)P(q_i)} \tag{15}$$

Among them, $P(q_i)$ represents the prior probability of q_i; $P(O_t|q_i)$ represents the likelihood of the observation vector O_t given q_i; M represents the total number of factors within the acoustic space.

Assuming τ_i is the starting time of q_i and d_i is the duration of q_i, then the logarithmic posterior probability score $\overline{\omega}$ of factor q_i is the mean logarithmic posterior probability score of all frame series in q_i which is presented in Equation (16) (Rabiner and Juang, 1993):

$$\overline{\omega}_i = \frac{1}{d_i} \sum_{t=\tau_i}^{\tau_i+d_i-1} logP(q_i|O_t) \tag{16}$$

The logarithmic posterior probability score $\overline{\omega}$ is defined as the mean logarithmic posterior probability score of all N phonemes in the sentence based on Equation (17) (Rabiner and Juang, 1993):

$$\overline{\omega} = \frac{1}{N} \sum_{i=1}^{N} \overline{\omega}_i \tag{17}$$

GOP is a simplification of logarithmic posterior probability, and the GOP of phoneme q_i is defined as Equation (18) (Rabiner and Juang, 1993):

$$G_i = \log \frac{P(q_i|O_t)}{\max_{i \in M}^{P(q_i|O_t)}} \tag{18}$$

The GOP score of a sentence can be obtained by taking the average GOP score of all N phonemes as Equation (19) (Rabiner and Juang, 1993):

$$G = \frac{1}{N} \sum_{i=1}^{N} G_i \tag{19}$$

2.3.2 Fluency characteristics

Speech speed is defined as the number of phonemes a student reads per unit of time, and the speed of speech reflects the fluency of the student's reading. The calculation formula for speech speed R_S is presented in Equation (20) (Zhu et al., 2024):

$$R_S = Q / T \tag{20}$$

Among them, T represents the reading time; Q represents the number of phonemes read by students during this period.

The duration of a segment represents the duration of pronunciation for different phonemes in a student's pronunciation, and the segment duration rating is as Equation (21) (Rabiner and Juang, 1993):

$$D = \frac{1}{N} \sum_{i=1}^{N} \log P\left(f(d_i)|q_i\right) \tag{21}$$

Among them, $f(d_i)$ represents the normalization function.

When students read aloud, if they are not clear about how to pronounce a word, there will be pauses between the words, and the proportion of the total pause time in the reading time reflects the fluency of the student's reading. The duration of pause is as Equation (22) (Yuen et al., 2023):

$$D_{PAU} = T_{SIL} / T \tag{22}$$

Among them, T_{SIL} represents the total duration of the silent part in the spoken speech.

2.3.3 Integrity features

When students read aloud, there is a possibility of missing words. The proportion of the number of words read by students in the reading content is used as the evaluation index for completeness, and word matching is defined as Equation (23) (Yuen et al., 2023):

$$R_{MAT} = \omega / W \tag{23}$$

Among them, ω represents the number of words that match the recognition result with the specified reading content; W represents the total number of words read aloud.

2.4 Multi feature fusion of speech signals

A feedback neural network that can fuse multi granularity features is proposed for the accuracy, fluency, and completeness features extracted above, for the fusion of multiple features in speech signals. Feature fusion plays a crucial role in the research of innovative methods for English pronunciation teaching in ESL contexts. By fusing features of different granularities and types such as accuracy, fluency, and completeness, the characteristics of speech signals can be more comprehensively described, thereby improving

the accuracy and sensitivity of evaluation platforms to student pronunciation performance. At the same time, multi feature fusion also helps to comprehensively consider the correctness, fluency, and coherence of pronunciation, making the platform more comprehensive in evaluating students' pronunciation level, providing teachers with more targeted feedback and guidance, and improving students' learning effectiveness and teaching quality.

RNN adds a feedback mechanism on the basis of artificial neural networks. RNN networks include input layers, hidden layers, memory layers, and output layers. The input layer inputs accuracy features, fluency features, and completeness features, respectively. The memory layer is a collection of neurons fed back from the hidden layer, used to record the content of the previous moment in the hidden layer (He, 2021; Gui, 2024). The structure of the RNN network is shown in Figure 1.

Let t be the current moment where the network is located, $f(t)$ represent the current speech frame features, $s(t)$ represent the segment features within the t time period, $E(t)$ represent the cognitive window features, and $x(t)$ and $z(t)$ represent the outputs of the two hidden layers of the network, respectively. W_1 is the weight matrix connecting input layers $f(t)$ and $s(t)$ to hidden layer $x(t)$, W_2 is the weight matrix connecting hidden layer $x(t)$ to hidden layer $z(t)$, W_3 is the weight matrix connecting hidden layer $z(t)$ to output layer $y(t)$, W_4 is the weight matrix connecting memory layer $x_c(t)$ to hidden layer $x(t)$, and W_5 is the weight matrix connecting cognitive window feature input layer $E(t)$ to hidden layer $z(t)$. The output of the hidden layer $x(t)$ is shown in Equation (24):

$$x(t) = f\left(W_1 s(t) + f(t)\right) + W_4 x_c(t) \tag{24}$$

Among them, $f(\cdot)$ takes the sigmoid function. The calculation formula for memory layer $x_c(t)$ is as Equation (25):

$$x_c(t) = x(t-1) \tag{25}$$

The calculation formula for the hidden layer $z(t)$ is as Equation (26):

$$z(t) = f\left(W_2 x(t)\right) + W_5 E(t) \tag{26}$$

The calculation formula for the hidden layer $y(t)$ is as Equation (27):

$$y(t) = f\left(W_3 z(t)\right) \tag{27}$$

The classic Back Propagation Error (BP) algorithm (Mirsadeghi et al., 2021; Pommé and Pelczar, 2021; Fatema et al., 2022) is used to update the weights of each layer node in the RNN network. Using the updated weights to expand RNN network training, the training process is shown in Figure 2.

The process of using a trained RNN to achieve multi feature fusion of speech signals can be described as follows: Firstly, different types of features such as accuracy, fluency, and completeness are input into the pre trained RNN network. Through the multi-layer neural structure of the network, learning and extraction of these features are achieved. RNN networks can effectively capture temporal and contextual information between features, thereby better expressing the features

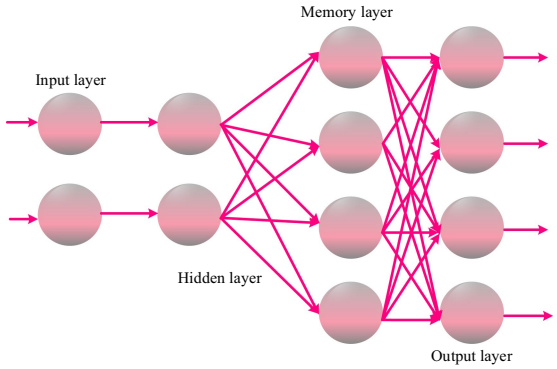

FIGURE 1
RNN network structure diagram.

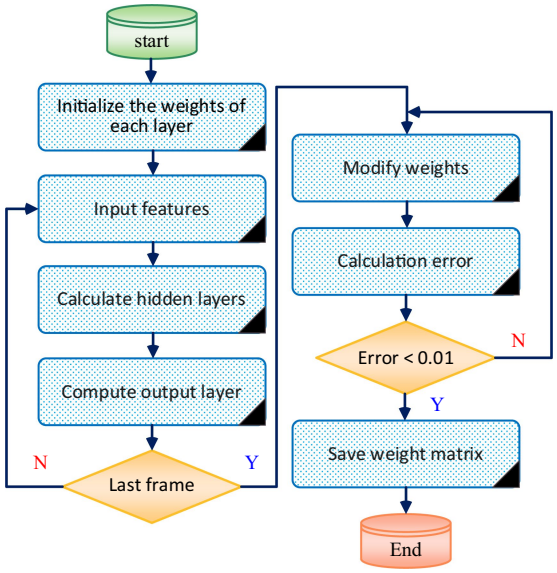

FIGURE 2
RNN training process.

of speech signals. Next, through the output layer or related connection layer of the network, different levels of features are fused, and the automatic learning and weight adjustment mechanism of the neural network is utilized to achieve feature synthesis and fusion. Finally, through the backpropagation and optimization process on the training dataset, the network parameters are continuously optimized, enabling the network to accurately fuse different features and extract more representative speech features, providing an effective method for multi feature fusion of speech signals.

2.5 Computer assisted English pronunciation evaluation

Based on the multi feature fusion results of English speech signals, computer-aided English pronunciation evaluation is carried out to provide reference for English pronunciation teaching. The English pronunciation evaluation process is shown in Figure 3.

From Figure 3, it can be seen that first, the preprocessed learner's English pronunciation is subjected to speech segment validation, including vowel segment segmentation, establishment of validation system, and reliability of validation system; Then, the HMM model is used to train acoustic models on a large number of standard pronunciation databases. The Viterbi algorithm is used to segment and decode speech, and this information, including the extraction of evaluation parameters, regularization of evaluation parameters, parameter association process, and evaluation mechanism, is sent to the core of the English pronunciation evaluation system, which is the pronunciation evaluation module. Through this module, the weights of each evaluation parameter in English pronunciation evaluation are calculated, to reflect the opinions of human experts on the quality of English sentences and provide feedback to learners, including comprehensive ratings and corrective suggestions compared through expert knowledge bases.

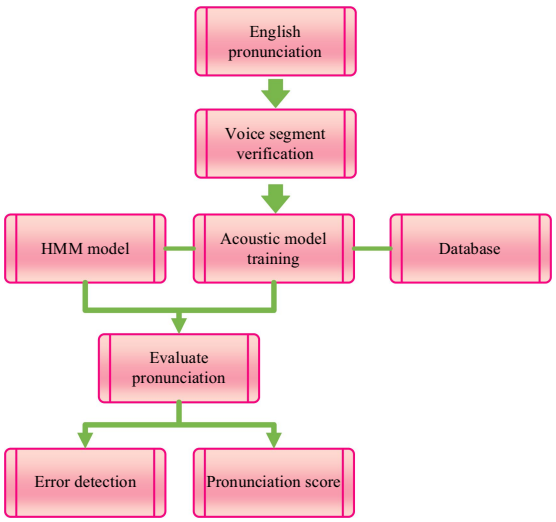

FIGURE 3
English pronunciation evaluation process.

3 Experimental verification analysis

3.1 Experimental data

Use the Sphinx4 speech recognition system released by Carnegie Mellon University as the experimental platform. In order to extract feature parameters that represent the essence of English pronunciation, pre-emphasis, normalization, framing, windowing, and endpoint detection are performed on English speech data, laying the foundation for subsequent feature extraction and pronunciation evaluation. The sample size of this study is the English pronunciation of 100 college students, half male and half female, covering students with different pronunciation levels. They read sentences from the Arctic corpus, each containing 8 to 20 words, totaling 1,000 sets of pronunciation data. The type of English studied is General English.

3.2 Evaluation indicators

- Teaching evaluation accuracy: Teaching evaluation accuracy refers to the degree to which a method's evaluation of student learning outcomes or teaching effectiveness is in line with the actual situation.
- Speech fluency: Set speech fluency to 5 levels, with higher levels indicating higher speech fluency. According to the established fluency assessment criteria, professional speech evaluators or teachers are used as evaluators to rate or grade the spoken language of learners. The evaluation criteria include pronunciation accuracy, speaking speed, intonation, and appropriateness of pauses.Spectral features: Spectral features reveal the characteristics of speech signals in the frequency domain, which are used to verify that the method can effectively extract the essential feature parameters of speech.

3.3 Result analysis

3.3.1 Verification of teaching evaluation accuracy

In order to verify the application effect of the method proposed in this paper, the accuracy of teaching evaluation was taken as the experimental indicator. The teaching evaluation accuracy of the method proposed in this paper, the method Wen (2020), and Tuba et al. (2018) were compared, and the results are shown in Table 1.

According to the analysis of Table 1, the highest accuracy value of the teaching evaluation method in this paper can reach 99.3%, the highest accuracy value of the teaching evaluation method Wen (2020) is 68.9%, and the highest accuracy value of the teaching evaluation method Tuba et al. (2018) is 77.8%. And in multiple tests, the accuracy of the method proposed in this paper is higher than that of the methods Wen (2020) and Tuba et al. (2018), indicating that the method proposed in this paper has high evaluation accuracy and can help teachers understand students' learning situations and optimize teaching strategies.

3.3.2 Voice fluency verification

On the basis of the above results, speech fluency was used as an evaluation indicator to compare and verify the methods Wen (2020) and Tuba et al. (2018), and our method. The results are shown in Table 2.

According to the analysis of Table 2, it can be seen that under the application of the method in this paper, students have a higher level of fluency in English pronunciation, both reaching level 4 or above. However, under the application of the methods Wen (2020) and Tuba et al. (2018), students have a lower level of fluency in English pronunciation. The application effect of the English pronunciation teaching method proposed in this paper is superior to existing methods, and it helps to improve the fluency of students' English pronunciation. Therefore, it has higher application value.

3.3.3 Spectral feature verification

In English pronunciation teaching, amplitude is used to represent the characteristics of speech signals in the frequency domain. By analyzing the spectral features, we can better understand the energy distribution of speech signals in the frequency domain, which is helpful for pronunciation teaching and speech recognition research. The test results of different methods are shown in Figure 4.

TABLE 1 Teaching evaluation accuracy test results.

Test no	Accuracy %		
	Wen, 2020	Tuba et al., 2018	Proposed method
1	68.5	72.3	98.0
2	62.9	76.4	96.1
3	65.4	67.9	99.3
4	66.0	64.2	98.6
5	68.9	69.0	97.2
6	73.1	77.8	99.0
7	67.0	70.8	97.8

TABLE 2 Results of speech fluency test.

Test no	Voice fluency		
	Wen, 2020	Tuba et al., 2018	Proposed method
1	4	2	5
2	3	5	4
3	3	4	4
4	4	3	5
5	3	3	5
6	2	4	5
7	5	5	5

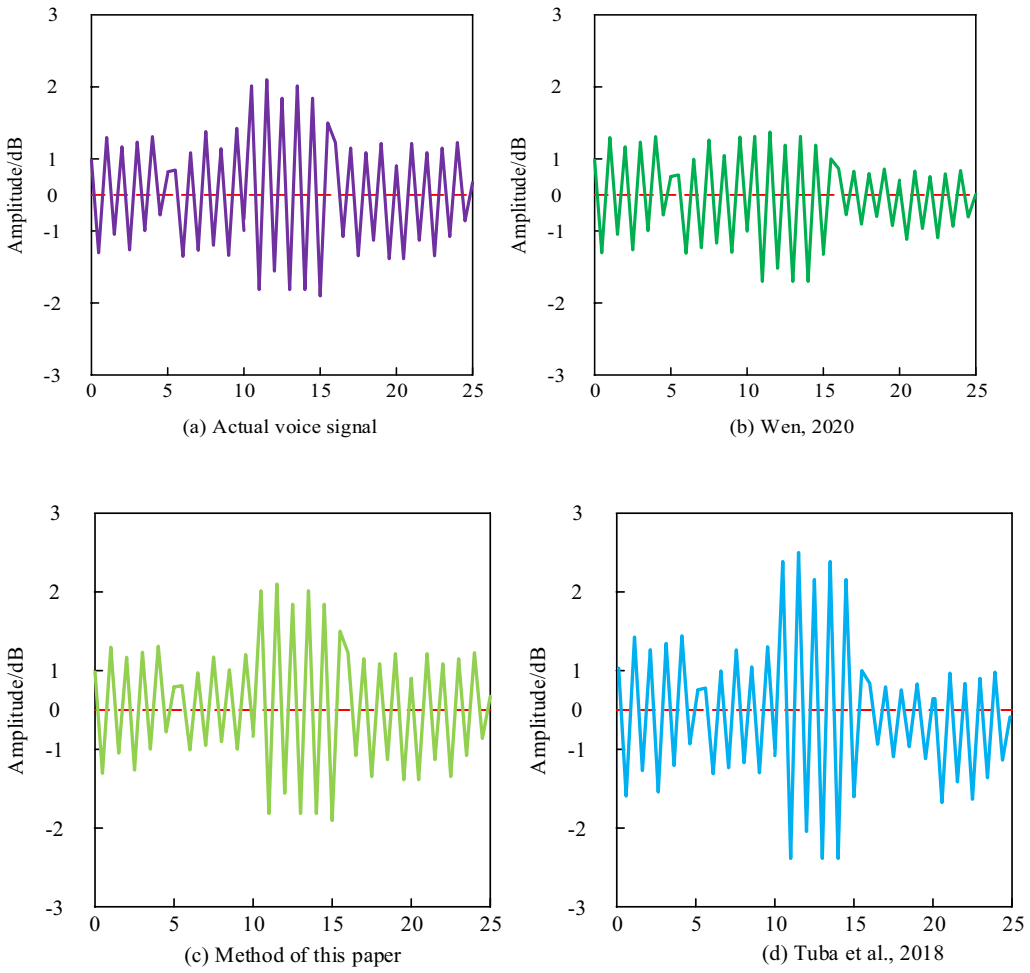

FIGURE 4
Spectral feature test results.

From Figure 4, it can be seen that the amplitude of the English speech signal obtained by the method proposed in this paper has a high similarity with the amplitude of the original speech signal, while the amplitude of the English speech signal obtained by the methods proposed Wen (2020) and Tuba et al. (2018) has a certain difference from the amplitude of the original speech signal. This indicates that the method proposed in this paper can accurately extract the characteristics of student speech signals, which helps teachers to obtain information about their learning status, This is because the method proposed in this paper not only extracts features such as accuracy, fluency, and completeness of student English pronunciation, but also integrates these features to achieve effective evaluation of student pronunciation quality, providing reliable reference for improving the effectiveness of English pronunciation teaching.

4 Conclusion

This study presents an innovative approach to teaching English pronunciation in an ESL context, incorporating modern technology and advanced algorithms to enhance the learning experience. The methodology includes multi-sensor detection methods for accurate data collection, preprocessing techniques such as pre-emphasis, normalization, framing, windowing, and endpoint detection to optimize speech signal quality, and the extraction and fusion of features related to pronunciation accuracy, fluency, and completeness using a feedback neural network. Empirical validation of this method shows significant improvements over existing techniques. The teaching evaluation accuracy of the proposed method reached up to 99.3%, compared to 68.9 and 77.8% for methods referenced in previous studies. Furthermore, students demonstrated higher fluency in English pronunciation, with most achieving a fluency level of 4 or above, whereas existing methods resulted in lower fluency levels. The proposed method also excelled in spectral feature verification, accurately extracting speech signal characteristics and aligning closely with the original speech signal amplitudes, unlike the notable discrepancies observed in methods from previous studies. Overall, this innovative approach provides a robust framework for improving English pronunciation teaching effectiveness, offering personalized guidance and real-time feedback, which are critical for addressing individual learner needs and enhancing overall teaching quality.

Future research can further explore the optimization of English oral pronunciation and speech signal detection models, as well as more diverse feature extraction methods, to more comprehensively evaluate students' English pronunciation ability. At the same time, it is possible to consider integrating artificial intelligence and machine learning technologies more deeply into English pronunciation teaching, achieving more personalized and intelligent teaching guidance. In addition, research can also be conducted on how to combine pronunciation teaching with the teaching of other language skills such as listening, reading, and writing to promote the improvement of students' comprehensive English abilities.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LP: Conceptualization, Formal analysis, Project administration, Supervision, Writing – original draft, Writing – review & editing. NT: Data curation, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adelekea, I. J., and Onyebuchib, C. N. (2023). Challenges teachers experience in teaching English second language in secondary schools in the northwest province. *World Lang. Lit. Cult. Stud.* 2, 11–19. doi: 10.26480/wllcs.01.2023.11.19

Ahmed, G., and Lawaye, A. A. (2023). CNN-based speech segments endpoints detection framework using short-time signal energy features. *Int. J. Inf. Technol.* 15, 4179–4191. doi: 10.1007/s41870-023-01466-6

Al-Asi, Z. (2024). Challenges and strategies in teaching English pronunciation to non-native speakers. *Afr. J. Adv. Stud. Hum. Soc. Sci.* 2024, 34–41.

Alghazo, S. (2015). The role of curriculum design and teaching materials in pronunciation learning. *Res. Lang.* 13, 316–333. doi: 10.1515/rela-2015-0028

Alghazo, S., and Zidan, M. (2019). Native-speakerism and professional teacher identity in L2 pronunciation learning. *Indones. J. Appl. Linguist.* 9, 241–251.

Alghazo, S. M., Jarrah, M., and Al Salem, M. N. (2023). The efficacy of the type of instruction on second language pronunciation acquisition. *Front. Educ.* 8:Article 1182285. doi: 10.3389/feduc.2023.1182285

Azimova, D., and Solidjonov, D. (2023). Learning English language as a second language with augmented reality. *Qo'Qon Univ. Xabarnomasi* 6, 112–115. doi: 10.54613/ku.v6i6.264

Bardol, F. (2023). "Teaching English pronunciation in a blended environment: feedback on a pedagogical innovation" in Innovation in language learning and teaching: The case of the southern Caribbean. eds. D. Mideros, N. Roberts, B.-A. Carter and H. Reinders (Cham: Springer International Publishing), 43–55.

Clymer, E., Alghazo, S., Naimi, T., and Zidan, M. (2020). CALL, native-speakerism/culturism, and neoliberalism. *Interchange* 51, 209–237. doi: 10.1007/s10780-019-09379-9

Duan, J., and He, Z. (2023). An English pronunciation and intonation evaluation method based on the DTW algorithm. *Soft. Comput.*, 1–9. doi: 10.1007/s00500-023-08027-w

Fatema, N., Farkoush, S. G., Hasan, M., and Malik, H. (2022). Deterministic and probabilistic occupancy detection with a novel heuristic optimization and Back-propagation (BP) based algorithm. *J. Intell. Fuzzy Syst.* 42, 779–791. doi: 10.3233/JIFS-189748

Gao, Y., Wu, Y., and Qian, J. (2023). Intelligent multimedia network security and Pbl teaching mode in the basic course teaching of college design major. *ACM Trans. Asian Low Resour. Lang. Inf. Proc.* 23, 1–14. doi: 10.1145/3597429

Gui, L. (2024). *Design and application of an English-assisted teaching system based on recurrent neural network-long short-term memory*. In: 2024 Second International Conference on Data Science and Information System (ICDSIS), (IEEE), pp. 1–4.

He, Z. (2021). English grammar error detection using recurrent neural networks. *Sci. Program.* 2021, 1–8. doi: 10.1155/2021/7058723

Huang, X., Acero, A., Hon, H. W., and Reddy, R. (2001). Spoken language processing: A guide to theory, algorithm, and system development. *1st* Edn. United States: Prentice Hall PTR.

Jelinek, F. (1998). Statistical methods for speech recognition. Cambridge, MA: MIT Press.

Jiang, S., and Chen, Z. (2024). *Simulation of machine-assisted translation model based on Python data analysis*. In: International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE), pp. 724–728.

Le-Qing, Z. (2011). *Insect sound recognition based on MFCC and PNN*. In: 2011 International Conference on Multimedia and Signal Processing, 42–46.

Li, M. (2020). Realization of the intelligent long-distance multimedia teaching system. *Front Comput* 8, 1526–1531. doi: 10.1007/978-981-15-3250-4_199

Li, K. (2024). Research on optimization of English teaching in universities under the guidance of applied talent training. *Adult Higher Educ* 2024:6. doi: 10.23977/aduhe.2024.060209

Li, X., and Huang, X. (2024). Improvement and optimization method of college English teaching level based on convolutional neural network model in an embedded systems context. *Comput. Aided Des. Appl* 21, 212–227. doi: 10.14733/cadaps.2024.S8.212-227

Li, Y., and Wu, F. (2023). Design and application research of embedded voice teaching system based on cloud computing. *Wirel. Commun. Mob. Comput.* 2023, 1–10. doi: 10.1155/2023/7873715

Liang, D., Su, H., Singh, T., Mahadeokar, J., Puri, S., Zhu, J., et al. (2023). *Dynamic speech endpoint detection with regression targets*. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (IEEE), 1–5.

Liu, X. (2021). A study of English phonetic teaching strategies from the perspective of embodied cognition. *Theory Pract. Lang. Stud.* 11, 556–560. doi: 10.17507/tpls.1105.14

Lounis, M., Dendani, B., and Bahi, H. (2024). Mispronunciation detection and diagnosis using deep neural networks: a systematic review. *Multimed. Tools Appl.* 83, 62793–62827. doi: 10.1007/s11042-023-17899-x

Ma, L., Yang, S., Gong, Y., Wang, X., and Wu, Z. (2021). Echofilter: end-to-end neural network for acoustic echo cancellation. *arXiv* 2105:14666. doi: 10.48550/arXiv.2105.14666

Mirsadeghi, M., Shalchian, M., Kheradpisheh, S. R., and Masquelier, T. (2021). STiDi-BP: spike time displacement based error backpropagation in multilayer spiking neural networks. *Neurocomputing* 427, 131–140. doi: 10.1016/j.neucom.2020.11.052

Nagle, C., and Hiver, P. (2023). Optimizing second language pronunciation instruction: replications of Martin and Sippel (2021), Olson and Offerman (2021), and Thomson (2012). *Lang. Teach.* 57, 1–14. doi: 10.1017/S0261444823000083

Nguyen, L. T. (2024). "Pronunciation teaching innovation in the English as a foreign language classroom" in Innovation in language learning and teaching: The case of Vietnam and Cambodia. eds. L. Phung, H. Reinders and V. P. H. Pham (Cham: Springer Nature Switzerland), 115–133.

Niu, B., and Wei, Y. (2023). *An adaptive speech noise reduction method based on noise classification*. In: 2023 IEEE 13th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 252–258.

Ozodova, M. C. (2024). Methodology of teaching pronunciation in ESL. *Int. J. Sci. Res.* 5, 31–36.

Pommé, S., and Pelczar, K. (2021). Empirical decomposition and error propagation of medium-term instabilities in half-life determinations. *Metrologia* 58:035012. doi: 10.1088/1681-7575/abf7df

Priya, M. L. S., and Kumar, P. (2020). Teaching phonetics to enhance pronunciation in an ESL classroom. *J. Crit. Rev.* 7, 669–672. doi: 10.31838/jcr.07.02.121

Quatieri, T. F. (2002). Discrete-time speech signal processing: Principles and practice. India: Pearson Education India.

Rabiner, L., and Juang, B. H. (1993). Fundamentals of speech recognition. Hoboken, NJ: Prentice-Hall, Inc.

Ruan, G. (2023). A practical study of applying optimization theory in English teaching work in colleges and universities. *Appl. Math. Nonlinear Sci.* 9:11.

Shen, Y.-L., and Wai, R.-J. (2022). Fast-Fourier-transform enhanced progressive singular-value-decomposition algorithm in double diagnostic window frame for weak arc fault detection. *IEEE Access* 10, 39752–39768. doi: 10.1109/ACCESS.2022.3165793

The Hanning Window. (2024). Available at: https://www.sciencedirect.com/topics/engineering/hanning-window (Accessed July 23, 2024).

Thi-Nhu Ngo, T., Hao-Jan Chen, H., and Kuo-Wei Lai, K. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: a meta-analysis. *ReCALL* 36, 4–21. doi: 10.1017/S0958344023000113

Tuba, E., Jovanovic, R., Hrosik, R. C., Alihodzic, A., and Tuba, M. (2018). Web intelligence data clustering by bare bone fireworks algorithm combined with k-means., in Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, 1–8.

Wen, Y. (2020). Design of DTW algorithm based automatic correction system for English pronunciation mistakes. *Mod. Electro. Tech.* 43, 124–126.

Weng, Z., Qin, Z., and Li, G. Y. (2021). *Semantic communications for speech signals*. In: ICC 2021-IEEE International Conference on Communications, (IEEE), pp. 1–6.

Yaccob, N. S., Yunus, M. M., and John, D. S. (2023). Global education movement: English as a second language teachers' perceptions of integrating volatility, uncertainty, complexity, and ambiguity elements in lessons. *Front. Psychol.* 14:1007970. doi: 10.3389/fpsyg.2023.1007970

Yang, H. (2023). An analysis of innovative teaching of English phonetics for non-English major students. *J. Contemp. Educ. Res.* 7, 86–92. doi: 10.26689/jcer.v7i8.5259

Yegnanarayana, B., and Gangashetty, S. V. (2011). Epoch-based analysis of speech signals. *Sadhana* 36, 651–697. doi: 10.1007/s12046-011-0046-0

Yuen, I., Ibrahim, O., Andreeva, B., and Möbius, B. (2023). *Non-uniform cue-trading: differential effects of surprisal on pause usage and pause duration in German*. In: Proceedings of the 20th international congress of phonetic sciences. Prague: Guarant International, pp. 619–623.

Zhu, J., Chen, H., Wen, X., Huang, Z., and Zhao, L. (2024). *An adaptive speech speed algorithm for improving continuous speech recognition*. In: Proceedings of the 2023 4th international conference on machine learning and computer application. New York, USA: Association for Computing Machinery, pp. 606–610.

Glossary

ASR	Automatic speech recognition	$P(q_i)$	The prior probability of q_i
BP	Back Propagation	$P(O_t\|q_i)$	The likelihood of the observation vector
CNN	Convolutional Neural Network	$s(\phi,\varepsilon)$	The output extension of English pronunciation signals
ESL	English as a Second Language	$s(t)$	The segment features in the t time period
FEA	Finite Element Analysis	$u(t)$	The resonance peak
FFT	Fast Fourier Transform	$X(n)$	The speech signal
GOP	Goodness of Pronunciation	$X(n,k)$	The complex representation of the k -th frequency point in frame n
HMM	Hidden Markov Models	$x(t)$	The output of the hidden layer of the network
IT	Information Technology	$x[n]$	The speech signal
LSTM	Long Short-Term Memory	$y(t)$	The output layer
MOLP	Multi-Objective Linear Programming	$y(t-\phi)$	The complex envelope of English pronunciation frequency components
RNN	Recurrent Neural Network	$z(t)$	The output of the hidden layer of the network
a	The advanced statistical features of the signal	d_i	The normalization function
b	The frequency shift distribution	$R_k(t)$	The phase of speech detection
l	The frame length	T_{SIL}	The total duration of speech silence
M	The total number of factors within the acoustic space	W_i	The weight matrix
n	The number of voice point	$x_c(t)$	The connecting memory layer
N	The phonemes in the sentence	x_i	The i-th sample point
Q	The number of phonemes read by students	$Y_k^{'}(t)$	The output signal component
s	The frame shift	z_n	The distribution of sample speech elements
T	The reading time	α	The pre-emphasis coefficient
w	The total number of words read aloud	ε	The bandwidth of signal acquisition feature expansion
$d(t)$	The amplitude of the received signal	$\varnothing$	The frequency shift in English pronunciation signals
$f(0)$	The sigmoid function	ϖ	The mean logarithmic posterior probability score
$f(n)$	The frame rate	τ_i	The starting time
$f(t)$	The step transfer function of speech signals	$\mu(a,b)$	The delay component of broadband signal
$M(n,k)$	The amplitude spectrum		