Precise Guidance for English Prosody Perception Pronunciation Based on Deep Learning

Tingting Liu[1], Qing Li[2]
[1]School of Foreign Languages and Literature, Shandong University, Jinan 250100, Shandong, China
[2]School of Science, Shandong Jianzhu University, Jinan 250101, Shandong, China
Corresponding author: Tingting Liu (e-mail: tengqi2@126.com)

ABSTRACT In English pronunciation teaching, prosody has long received less attention than segmental phoneme training, although stress, rhythm, and intonation strongly affect speech intelligibility and naturalness. Existing technological aids mainly focus on phoneme-level correction and still lack fine-grained diagnosis and adaptive feedback for prosodic waveform patterns. From an engineering perspective, prosody perception can be regarded as a speech-signal analysis problem involving pitch trajectory, energy distribution, temporal modulation, and waveform feature recognition, which is methodologically related to time-frequency analysis in wave-based signal processing. This study constructs an intelligent perception and precise guidance system for English prosody based on deep learning. Speech samples were collected from English learners with different dialect backgrounds, and a specialized corpus containing 4592 valid speech samples was constructed. Acoustic features including fundamental frequency, duration, intensity, energy rising points, and intonation trajectories were extracted. A classifier ensemble model was trained for stress detection, machine learning models were used for rhythm feature analysis, and a convolutional neural network was developed for intonation pattern recognition. The system further integrates diagnostic results into visual feedback, comparative listening, and graded practice modules. The results show that the proposed models can effectively identify learners' prosodic errors across stress, rhythm, and intonation dimensions. The teaching experiment indicates that the system significantly improves learners' prosody perception and pronunciation naturalness, providing a data-driven technical route for intelligent and precise oral pronunciation guidance.

INDEX TERMS deep learning, English prosody, speech signal processing, intonation recognition, waveform feature extraction

I. INTRODUCTION

JUST as the precise alignment of materials determines the tensile strength of industrial materials, the emerging importance of spoken English communication necessitates a focus on pronunciation as the fundamental dimension ensuring the accurate transmission of information. This structural accuracy in speech acts as the essential "warp and weft" of effective global exchange, directly impacting the quality and reliability of international technical discourse within the textile industry. However, in the area of the English pronunciation teaching language, there has been a tendency in the past 50 years or so, to give more importance in segmental phonemes than suprasegmental phonemes. Segmental phonemes place an emphasis on the proper pronunciation of individual phonemes such as vowels and consonants whereas suprasegmental phonemes focus on prosodic features such as stress, rhythm and intonation that exist between individual phonemes. Research has shown that prosodic features determined even more strongly the comprehensibility and naturalness of speech than segmental features did. Even if non-native speakers are able to pronounce every single phoneme accurately, if the stress is incorrect, the rhythm is unbalanced or the intonation is strange, the speech will still be difficult to be understood smoothly by native speakers.

Breakthroughs in deep learning technology have made powerful tools available to prosodic perception and analysis. Convolutional neural networks are good at extracting features from the spectrogram of speech, recurrent neural networks and their variants are good at handling temporal dependencies, and attention mechanisms and Transformer architectures can extract long-range contextual information in speech signals. These technologies make it possible to automatically extract prosodic features, identify prosodic patterns, and quantify prosodic errors from raw speech. Compared to traditional rule-based or shallow model-based prosodic analysis, deep learning approaches possess a much firmer representational ability and generalization ability,

which can cope with complex and diverse natural speech.

Drawing upon the high-precision monitoring used in synthetic material extrusion, the main objective of this study is to formulate a deep learning-based system for perceiving and guiding English prosody. This framework ensures that the flow of spoken rhythm achieves the same level of uniformity and structural integrity required in the production of high-tenacity industrial filaments. The three main dimensions of English prosody to be investigated include stress patterns (word level), rhythmic features (sentence level) and intonation types (communicative function level). A good prosodic annotated corpus is created by gathering speech data from English speakers that have different dialect backgrounds. Deep learning models have been designed in different prosodic dimensions to extract and intelligently diagnose the prosodic dimension of the learners' output. Based on the results of the diagnostic, a fine-tuning guidance module consisting of visual feedback and adaptive practice is created and its effectiveness verified through teaching experiment. This study attempts to answer the following three questions: First, can deep learning models effectively identify English learners' stress, rhythm and intonation errors? Second, the extent to which various acoustic features contribute to prosodic perception? Third, whether precise guidance using deep learning would have a significant enhancement effect on the prosodic output ability of the learners.

II. RELATED WORK

Research on English pronunciation teaching and acquisition presents a multi-dimensional exploration trend, covering multiple key areas such as technology empowerment, mother tongue influence, teaching strategies and teacher cognition. Akhter [1] reviewed and found that continuous AI-assisted practice can consistently improve learners' phoneme pronunciation, word stress, intonation, speech rate, and reduce pauses. Hoang et al. [2] explored the feasibility of using chatbot AI to improve students' English pronunciation in a vocational education environment. Bashori et al. [3] evaluated the effectiveness of language learning systems based on automatic speech recognition (ASR) technology in improving English pronunciation. Octaviani et al. [4] explored the specific impact of learners' first language on their English pronunciation. Utami and Morganna [5] examined the effect of shadow reading technology (i.e., imitating and reading along with audio materials) on improving students' English pronunciation ability.

Traditional pronunciation teaching often focuses on mechanical imitation. Sardegna [6] proposed and provided theoretical support and empirical evidence for a strategy-based English pronunciation teaching model. Iskandar et al. [7] focused on how to reconcile students' resistance to project-based learning and self-study English pronunciation activities through scaffolded task design. Dandee and Pornwiriyakit [8] explored the effect of improving the pronunciation skills of EFL students through English phonetic symbol training. Hsieh et al. [9] explored the impact of an integrated learning system combining robots and physical objects on the English pronunciation and communication willingness of EFL (English as a Foreign Language) learners in primary school. Almusharraf [10] pointed out the gap between ideal pronunciation teaching and real practice in Saudi Arabian University EFL classrooms and emphasized the urgent need to provide teachers with stronger professional training to improve their pronunciation teaching and assessment capabilities. Existing technological aids have not yet been able to achieve in-depth diagnosis and adaptive and accurate feedback on learners' prosodic output. This study constructs a model that can intelligently perceive and analyze learners' English prosodic characteristics, and then provides more targeted and adaptive personalized guidance strategies that go beyond traditional segmental pronunciation correction, in order to make up for the shortcomings of current research in the level of accurate prosodic teaching and promote the development of English oral teaching towards a more refined and intelligent direction.

III. METHODS

A. CORPUS CONSTRUCTION AND DATA PREPROCESSING

To achieve deep learning modeling of English prosodic features, this study first constructed an English phonetic corpus specifically for prosodic analysis. A total of 156 subjects were recruited, including 24 native English speakers as the control group and 132 Chinese English learners, divided into Mandarin dialect area (52 subjects), Wu dialect area (40 subjects), and Yue dialect area (40 subjects) according to dialect background. To mitigate the potential bias introduced by the disparate sample sizes, a Synthetic Minority Oversampling Technique (SMOTE) was applied to the native speaker dataset during the training phase, ensuring balanced class distribution and enhancing the statistical robustness of the comparative prosodic analysis. To ensure a controlled comparative analysis, all non-native participants were verified to be at the B2 level of the Common European Framework of Reference for Languages (CEFR) through a standardized placement test, ensuring that observed prosodic variations were primarily attributable to dialectal transfer rather than disparate general proficiency levels. Furthermore, all subjects had 6–8 years of formal English education and were aged 18–25. Subjects read three types of corpus in a professional recording environment: 200 isolated words, focusing on word stress patterns; 60 sentences, including 20 declarative sentences, 20 general interrogative sentences, and 20 special interrogative sentences, focusing on intonation type; and 15 short passages, focusing on sentence rhythm features [11,12]. This study was conducted in accordance with general ethical standards for linguistic research. All participants were informed of the research purpose and procedure, and signed informed consent forms. All speech data were collected anonymously and used only for academic research. The recording equipment used

professional-grade condenser microphones with a sampling rate of 44.1kHz and quantization precision of 16 bits. Each subject's recording time was approximately 45 minutes, resulting in 7020 original speech samples. After removing samples with substandard recording quality, 6548 valid speech samples were obtained, including 892 samples from native speakers and 5656 samples from learners. The annotation of speech data is divided into three levels: segment level annotation is generated by an automatic speech recognition system to generate phoneme boundaries; stress level annotation is determined by three phonetics professionals who independently judge whether each syllable is stressed, with a consistency coefficient of 0.91; intonation level annotation is used to annotate the intonation type of each sentence; rhythm level annotation uses a semi-automatic method to extract syllable boundaries and energy curves. Data preprocessing includes steps such as pre-emphasis, framing, and windowing. Framing uses a frame length of 25 milliseconds and a frame shift of 10 milliseconds. For each speech frame, the following acoustic features are extracted: Mel-frequency cepstral coefficients, fundamental frequency trajectory, intensity envelope, and formant frequencies. To capture rhythmic features, the maximum energy rise point parameter was specifically calculated.

B. EMULATION LEARNING-BASED STRESS DETECTION MODEL

The essence of word stress detection is to classify syllables and determine whether each syllable carries the main stress in a word. This study constructs an ensemble classifier model based on multi-feature fusion. Three sets of features are extracted for each syllable: intrasyllable features, including syllable duration, average fundamental frequency, fundamental frequency range, average intensity, and peak intensity; relative features, including the fundamental frequency ratio, intensity ratio, and duration ratio of the syllable to the preceding and following syllables; and positional features, including the syllable's position in the word and whether it is a word-ending syllable.

This study employs a voting ensemble strategy, fusing the prediction results of four base classifiers: Support Vector Machine (SVM), Neural Network (NN), k-Nearest Neighbors (kNN), and Random Forest. This heterogeneous ensemble was selected to leverage the complementary strengths of different decision boundaries—SVM for high-dimensional margin maximization, kNN for local feature similarity, and Random Forest for non-linear interaction capture—providing a more stable decision framework for the extracted acoustic features than single-stream transformer models, which often require significantly larger datasets to converge on low-level prosodic parameters. The SVM uses a radial basis function kernel; the NN has a single hidden layer structure; the k-value of the k-Nearest Neighbors algorithm is set to 5; and the Random Forest contains 100 decision trees. The ensemble strategy uses a soft voting method.

The model was trained using five-fold cross-validation. The training set contained 15,420 syllable samples extracted from the speech of all subjects, including 6,734 stressed syllables and 8,686 unstressed syllables.

Model evaluation used four metrics: accuracy, precision, recall, and F1 score. Feature importance analysis was used to identify the acoustic features that contributed most to stress detection.

C. MACHINE LEARNING-BASED RHYTHM FEATURE ANALYSIS MODEL

The goal of speech rhythm analysis is to identify systematic differences in rhythmic patterns between native and non-native speakers. This study draws on cutting-edge methods in automatic speech rhythm analysis, focusing on the role of the maximum energy rise point and related features [13,14] .

For each reading text, syllable-level temporal features are extracted, and a series of rhythm metrics are calculated: syllable duration variation coefficient, vowel duration ratio, mean difference between adjacent syllable durations, fundamental frequency dynamic range, intensity dynamic range, mean and standard deviation of the maximum energy rise point, and alignment error between the maximum energy rise point and the syllable boundary.

This study employed various machine learning models for comparative experiments, including Support Vector Machines, Random Forests, Gradient Boosting Machines, and Logistic Regression. The training set contained 654 sentence samples. To enhance the interpretability of the models, the Shapley additive interpretation method was introduced to perform attribution analysis on the prediction results of the optimal model, identifying the rhythmic features that most effectively distinguish between native and non-native speakers.

D. INTONATION PATTERN RECOGNITION MODEL BASED ON CONVOLUTIONAL NEURAL NETWORK

The task of intonation pattern recognition is to automatically classify sentences into declarative sentences, general interrogative sentences, and special interrogative sentences. For each speech sentence, the fundamental frequency trajectory and intensity trajectory are extracted. After outlier removal, interpolation padding, and smoothing, to accommodate varying sentence durations without losing temporal resolution, the raw trajectories were first normalized for duration. Subsequently, a piecewise cubic spline interpolation was applied to generate a 200-point fixed-length representation for the neural network input. To compensate for any loss of temporal information during resampling, the original sentence duration and syllable-rate metadata were appended as auxiliary feature vectors to the final fully connected layer.

The structure of the convolutional neural network is as follows: The input layer receives data from two channels. The first convolutional layer contains 32 convolutional kernels, followed by a max pooling layer; the second convolu-

tional layer contains 64 convolutional kernels, followed by a max pooling layer; the third convolutional layer contains 128 convolutional kernels, followed by a global average pooling layer; then two fully connected layers are connected, and the output layer is a 3-node softmax layer. The model was trained using the Adam optimizer with an initial learning rate of 0.001. The loss function was cross-entropy loss. The training set contained 3600 sentence samples, 1200 sentences per class. Data augmentation techniques were used to add slight background noise and pitch perturbations to the original speech, expanding the training set to three times its original size. Model evaluation was performed using the confusion matrix and classification accuracy for each class.

IV. RESULTS AND DISCUSSION

A. PERFORMANCE AND FEATURE CONTRIBUTION ANALYSIS OF THE ACCENT DETECTION MODEL

Table 1 shows the difference in performance between the stress detection model for the subject group. The observation shows the model works best with the native speakers, with an accuracy of 96.8% and the accuracy rates are 93.5 and 92.7% respectively with Mandarin and Wu speakers, and 91.4% accuracy for Cantonese speakers. This decreasing trend is consistent with the differences in phonological characteristics between different Chinese dialects: Mandarin and English stress patterns are relatively alike, with significant contrasts between stressed and unstressed syllables; Wu preserves some tonal features from Middle Chinese, and therefore the contrast between stressed and unstressed syllables is relatively weaker; Cantonese has a complicated tonal system, with what is probably the greatest negative transfer effect on English stress. Stress detection is the most difficult for Cantonese learners, whose stress output is the most significant different from that of native speakers, so that Cantonese learners result in the lowest accuracy in stress classification.

TABLE 1. Performance of the accent detection model on different subject groups

Subject group	Sample size (number of syllables)	Accuracy (%)	Accuracy (%)	Recall rate (%)	F1 score (%)
native English speakers	3120	96.8	96.5	97.1	96.8
Mandarin learners	4080	93.5	92.9	94.2	93.5
Wu dialect learners	3840	92.7	92.1	93.3	92.7
Cantonese learners	3840	91.4	90.8	92.1	91.4
overall	14880	94.2	93.8	94.5	94.1

Feature importance analysis showed the contribution of different acoustic features to the stress detection. The result given by the feature importance ranking model for the random forest model identified that the top 5 important features were, in order: syllable intensity ratio (compared to the previous syllable), syllable duration, syllable peak intensity, fundamental frequency ratio (compared to the following syllable), and vowel space size in the syllable. This result supports the relative nature of stress perception: the relative perception of a syllable to be stressed is not only a matter of its own acoustic salience but also its contrast to the preceding and following syllables. The significance of intensity ratio and fundamental frequency ratio even outweighs the significance of absolute intensity value, which may indicate that learners are advised to focus on dynamic contrast between syllables. Rather than merely increasing the absolute vocal effort of a single syllable, learners should master the intersyllabic intensity ratio, ensuring the stressed syllable is perceptually prominent through its acoustic relationship with adjacent unstressed units.

B. RHYTHM FEATURE RECOGNITION AND NATIVE/NON-NATIVE SPEAKER CLASSIFICATION RESULTS

Rhythm feature analysis models were quite successful distinguishing between native or non-native speakers. Compared to various models of machine learning algorithms, Random Forest gave the best performances, have accuracy score of 87.6% on validation dataset and have the area under the ROC curve of 0.93 and F1- Score value of 0.86. Gradient Boosting Machine came next with an accuracy of 86.2%. Support Vector Machine (Radial basis Kernel) accuracy scored 84.5% and Logistic Regression accuracy scored 81.3%. The advantage of Random Forest is that it can effectively deal with non-linear interactions between features and rhythm features represent complex interactions with multiple dimensions on acousto-processing.

Table 2 presents the ranking of the contributions of different rhythmic features in the classification of the native and non-native languages. The results of analysis conducted with the Shapley additive interpretation method indicate that the top 5 features contributing the most are, in order, the mean of the maximum energy rise point value, the standard deviation of the intensity ratio, the alignment error between the maximum energy rise point to the syllable boundary, the coefficient of variation of the syllable duration, and the proportion of the vowel duration.

TABLE 2. Ranking of the contribution of rhythmic features to native language and non-native language classification

Sort	Feature Name	Average Shapley value	Feature Description
1	Mean value of maxD	0.142	The average of the maximum energy rise points of each syllable
2	Strength ratio to standard deviation	0.118	Standard deviation of the intensity ratio of adjacent syllables
3	maxD alignment error	0.097	The average time difference between the point of maximum energy rise and the syllable boundary
4	Syllable duration variation coefficient	0.085	The ratio of the standard deviation to the mean of syllable duration
5	Vowel duration percentage	0.076	The proportion of vowel duration to total sentence duration
6	Fundamental frequency dynamic range	0.064	The difference between the maximum and minimum fundamental frequencies within the statement
7	The ratio of the duration of adjacent syllables to the mean	0.052	The average of the ratios of the durations of adjacent syllables
8	Intensity dynamic range	0.043	Difference between the maximum and minimum strength values within a statement
9	Standard deviation of maxD value	0.031	Standard deviation of the maximum energy rise point value of each syllable
10	Fundamental frequency change rate	0.022	Mean of the absolute value of the first-order difference of the fundamental frequency

Further analysis of the distributional difference of some

important features between the two groups showed the mean maximum energy rise point value of native speakers was considerably higher than that of non-native speakers (0.62 vs 0.43, p<0.001), and the alignment error was considerably less (23 ms vs 47 ms, p<0.001). This means that the energy increase of native speakers are steeper, and in sync with the exact point where the syllable nucleus is, than in non-native speakers where the energy changes are slower and more scattered. The standard deviation of the intensity ratio is informative of the degree of intensity contrast between syllables, and the standard deviation of native speakers is bigger (0.31 vs. 0.19, p<0.001), meaning that the intensity contrast between stressed and unstressed syllable is more pronounced. With regard to the coefficient of variation of the syllable duration, the coefficient of variation of native speakers is significantly greater than that of non-native speakers (0.42 vs. 0.28, p<0.001) and this reflects the nature of English as a stress-timed language: that is, stressed syllables are significantly longer than unstressed syllables. The distribution of syllable duration in non-native speakers is rather even, with the typical features of syllable-based languages.

C. ACCURACY AND ERROR ANALYSIS OF INTONATION PATTERN RECOGNITION

The classification performance of the four types of sentences was more or less balanced: there were 92.4% accuracy for declarative sentences, 89.8% for general and 91.7% for special questions. Confusion matrix analysis revealed that the most frequent types of confusions were of general questions misclassified as declarative sentences (6.2% of samples of general questions) and special questions misclassified as declarative sentences (4.8% of samples of special questions). Declarative sentences were classified as questions less often (only 2.3%). This pattern of confusion represents typical learner errors: intonation features of questions are not obvious; in particular, the insufficient rise at the end of the sentence is confusing, and their intonation curve is drawn toward that of declarative sentences.

Table 3 shows the accuracy of classification in the intonation pattern recognition model in on different subjects. As can be seen from the table, the best accuracy is obtained for the native speakers (95.6%); with Mandarin speakers, the accuracy rates 91.8%; Wu speakers achieve 90.2%; and Cantonese speakers reach 87.5%. This gradient trend is consistent with the results of stress detection model which further confirms the differentiated effects of different dialect backgrounds on English prosody acquisition. Cantonese speakers have the most difficulty in intonation recognition task, with the most significant difference between their produced intonation and the native speaker.

TABLE 3. Classification accuracy of the intonation pattern recognition model on different subject groups

Subject group	Accuracy rate of declarative sentences (%)	Accuracy rate of general questions (%)	Accuracy rate of special interrogative sentences (%)	Overall accuracy (%)
native English speakers	96.7	94.2	95.8	95.6
Mandarin learners	92.8	90.3	92.3	91.8
Wu dialect learners	91.5	88.7	90.4	90.2
Cantonese learners	88.6	86.1	87.9	87.5
overall	92.4	89.8	91.7	91.3

In-depth examination of the misjudged samples showed certain intonation error patterns in the learners. In general questions the error of not having enough final rise, the most common error among native speakers was an average final fundamental frequency rise of 68 (female) and 42 (male) and very little among learners who suffered highest average final fundamental frequency rise of 31 (female) and 19 (male) with. Some learners even showed a last fall (account for 23.5% of wrong estimated general questions), as a complete change of the communicative intention of the question). In special questions, the most typical mistake was of a flat intonation through to the sentence, with no important fall after the core syllable: the average of the final fundamental frequency in native speakers fell by an average of 52 Hz (female) and 33 Hz (male) after the core syllable in special questions, whereas the average of the learners was only 24 Hz (female) and 15 Hz (male). Furthermore, some learners misjudged the intonation pattern of special questions (18.7% misjudged), confusing the intonation pattern of special questions and general questions.

According to the results of the diagnosis, the design of the precision guidance module has the following three aspects: visual feedback, comparative listening, and tiered practice. Visual feedback features the learner's basic frequency trajectory superimposed on a standard model, pointing out error-prone areas (e.g., lack of rise at end of sentences), and yielding explanations of correction points verbally. Comparative listening gives a comparison of the pronunciation of native speakers and learners to help learners perceive the difference of their pronunciation and target pronunciation. Tiered practice utilizes a rule-based decision engine to assign exercises. If the classifier's posterior probability for a specific prosodic error exceeds a threshold of 0.75, the system triggers a targeted module: learners with stress detection errors are routed to contrastive stress drills; those with rhythm alignment errors exceeding 40ms receive metronome-paced rhythmic synchronization tasks; and intonation errors trigger pitch-contour imitation exercises using a Dynamic Time Warping (DTW) score for real-time reinforcement, strong and weak syllable combination exercises are recommended for those with rhythm errors, and different sentence structure transformation exercises are recommended for those with intonation errors. Through an eight-week teaching experiment, accuracy in prosody perception of learners in the experimental group using this system improved by 18.6%, and pronunciation naturalness score improved by 15.3%, which is significantly higher than those of the control

group using traditional teaching method (7.2% and 6.8% improvement in perception and naturalness), which validates the effectiveness of deep learning-based precise guidance in actual teaching.

V. CONCLUSION

Leveraging deep learning technology, this study develops an English prosody perception and guidance system that integrates the precision of structural analysis to synchronize the core dimensions of stress, rhythm, and intonation. By mapping the rhythmic patterns of language to the mechanical consistency of industrial material processing, the system provides a high-fidelity framework for linguistic accuracy. By building a high-quality speech corpus of learners from multiple dialect backgrounds, an ensemble classifier model for stress detection, a machine learning model for rhythm feature analysis, and a convolutional neural network model for intonation pattern recognition were designed. Theoretically, this study verifies the application potential of deep learning technology in English prosody teaching, providing new analytical tools and data support for suprasegmental phoneme acquisition research. Practically, the system prototype developed in this study can provide a reference for the development of intelligent language learning products, promoting the transformation of English oral teaching from experience-based and general to data-driven and precise.

Limitations of the study include: the corpus size is still insufficient, especially the sample size of native speakers; the model's ability to analyze prosodic features in continuous speech needs improvement; and the precise guidance module is not yet fully adaptive, requiring manual setting of practice paths. Future research is planned to proceed in the following directions: expanding the corpus to include more English variants and learner groups; exploring end-to-end prosodic analysis models to reduce reliance on manual features; developing an adaptive feedback mechanism based on reinforcement learning to achieve dynamic optimization of practice paths; and extending the system to prosodic teaching in other languages to verify the cross-linguistic applicability of the method.

AUTHOR CONTRIBUTIONS

Conceptualization –Tingting Liu and Qing Li; methodology – Tingting Liu and Qing Li; investigation – Tingting Liu and Qing Li; writing-original draft preparation – Tingting Liu and Qing Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

HUMAN RESEARCH SUBJECTS

This study was conducted in accordance with general ethical standards for linguistic research. All participants were informed of the research purpose and procedure, and signed informed consent forms. All speech data were collected anonymously and used only for academic research.

REFERENCES

[1] E. Akhter, "The Impact of Human-Machine Interaction On English Pronunciation And Fluency: Case Studies Using AI Speech Assistants," Review of Applied Science and Technology, vol. 4, no. 02, pp. 473-500, 2025, doi: 10.63125/1wyj3p84.

[2] N. T. Hoang, D. N. Han, and D. H. Le, "Exploring Chatbot AI in improving vocational students English pronunciation," AsiaCALL Online Journal, vol. 14, no. 2, pp. 140-155, 2023, doi: 10.54855/acoj.231429.

[3] M. Bashori, R. van Hout, H. Strik, et al., "I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems," Innovation in Language Learning and Teaching, vol. 18, no. 5, pp. 443-461, 2024, doi: 10.1080/17501229.2024.2315101.

[4] R. P. Octaviani, L. M. Jannah, M. Sebrina, et al., "The impacts of first language on students English pronunciation," IJIET (International Journal of Indonesian Education and Teaching), vol. 8, no. 1, pp. 164-173, 2024, doi: 10.24071/ijiet.v8i1.6758.

[5] H. S. Utami and R. Morganna, "Improving students English pronunciation competence by using shadowing technique," English Franca: Academic Journal of English Language and Education, vol. 6, no. 1, pp. 127-150, 2022, doi: 10.29240/ef.v6i1.3915.

[6] V. G. Sardegna, "Evidence in favor of a strategy-based model for English pronunciation instruction," Language Teaching, vol. 55, no. 3, pp. 363-378, 2022, doi: 10.1017/S0261444821000380.

[7] I. Iskandar, R. Dewanti, S. D. Sulistyaningrum, et al., "Scaffolding Assignments to Conciliate the Disinclination to Employ Project-Based Learning of English Pronunciation and Autodidacticism," International Journal of Language Education, vol. 8, no. 2, pp. 199-227, 2024, doi: 10.26858/ijole.v8i2.64087.

[8] W. Dandee and P. Pornwiriyakit, "Improving English Pronunciation Skills by Using English Phonetic Alphabet Drills in EFL Students," Journal of Educational Issues, vol. 8, no. 1, pp. 611-628, 2022, doi: 10.5296/jei.v8i1.19851.

[9] W. M. Hsieh, H. C. Yeh, and N. S. Chen, "Impact of a robot and tangible object (R&T) integrated learning system on elementary EFL learners English pronunciation and willingness to communicate," Computer Assisted Language Learning, vol. 38, no. 4, pp. 773-798, 2025, doi: 10.1080/09588221.2023.2228357.

[10] A. Almusharraf, "Pronunciation instruction in the context of world English: exploring university EFL instructors perceptions and practices," Humanities and Social Sciences Communications, vol. 11, no. 1, pp. 1-11, 2024, doi: 10.1057/s41599-024-03365-y.

[11] F. M. A. M. Awadh, N. H. P. S. Putro, A. E. Pohan, et al., "Improving English pronunciation through phonetics instruction in Yemeni EFL classrooms," JOLLT Journal of Languages and Language Teaching, vol. 12, no. 2, pp. 930-940, 2024, doi: 10.33394/jollt.v12i2.10720.

[12] A. D. J. Rubio, "The situation of English pronunciation in primary education classrooms in Spain," International Journal of Instruction, vol. 17, no. 3, pp. 529-544, 2024, doi: 10.29333/iji.2024.17329a.

[13] M. D. Ammar, "English pronunciation problems analysis faced by English education students in the second semester at Indo Global Mandiri University," Global Expert: Jurnal Bahasa Dan Sastra, vol. 10, no. 1, pp. 1-7, 2022, doi: 10.36982/jge.v10i1.2166.

[14] X. Xue and R. E. Dunham, "Using a SPOC-based flipped classroom instructional mode to teach English pronunciation," Computer Assisted Language Learning, vol. 36, no. 7, pp. 1309-1337, 2023, doi: 10.1080/09588221.2021.1980404.