Research on Pronunciation Error Recognition Method for English Conversation Practice Based on Artificial Intelligence Corpus

Rui DANG[1], Zheng ZENG[2]
{drui1@cdtu.edu.cn, zzheng2@cdtu.edu.cn}
[1]School of Humanities and Design, Chengdu Technological University,
No. 1, Section 2, Zhongxin Avenue, Pidu District, Chengdu 611730, Sichuan, China
[2]International Office of Cooperation and Exchange, Chengdu Technological University,
No. 1, Section 2, Zhongxin Avenue, Pidu District, Chengdu 611730, Sichuan, China

Abstract. Methods relying on single-accent samples and simple rule discrimination can only roughly handle common pronunciation deviations, often resulting in low accuracy in pronunciation error recognition. To address this challenge, a novel pronunciation error recognition method for English conversation practice based on an artificial intelligence corpus is proposed. Following the principles of pertinence, representativeness, and scalability, the corpus is meticulously selected considering timeliness, comprehensiveness, typicality, and appropriateness. After initial software setup and preprocessing, a multi-level verbal and non-verbal annotation system is designed, thereby completing the construction of a robust multimodal English pronunciation corpus. The core methodology constructs a mathematical model for the speech recognition system, employs filtering detection on pronunciation signals, and defines a set of highly sensitive pronunciation feature parameters. Feature extraction is completed via specific non-linear formulas derived from these parameters. The extracted features are then scored against corpus patterns, normalized using a specific function, and integrated via a multi-model scoring mechanism. A predetermined threshold is set to determine the correctness of the pronunciation. Experimental results show that the proposed method accurately identifies pronunciation error regions, exhibiting high-position curve characteristics in the Precision-Recall (PR) analysis. The $F1$ score, which increases from 0.87 to 0.99 across a sample size of 100 to 1000, is validated by the robustness of the highly annotated multimodal corpus and the precision of the ensemble scoring approach. Furthermore, the method demonstrates superior classification performance in the confusion matrix, significantly enhancing the accuracy of pronunciation error recognition in dynamic conversational contexts.

Keywords: Artificial Intelligence Corpus, English Conversation Practice, Pronunciation Error Recognition, Feature Extraction, Multi-Model Scoring Mechanism

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365266

1 Introduction

As English maintains its status as a global lingua franca, improving pronunciation and oral expression skills is an urgent need for language learners. While the use of speech recognition technology to assist English pronunciation learning has become a hot topic, existing Computer-Assisted Language Learning (CALL) methods often suffer from insufficient accuracy and struggle to meet personalized needs. For example, Dennis [1] constructed an English pronunciation learning system driven by AI-powered speech recognition technology, utilizing a general speech recognition model for real-time evaluation. However, the performance is limited by the singularity of the model's training data, leading to poor recognition of pronunciation errors specific to non-native speakers. Bashori et al. [2] developed the ASR system "I Can Speak," which detects errors by combining rule matching and statistical models. Nevertheless, the limited nature of its rule base restricts its ability to handle complex phonetic phenomena. Dou [3] proposed a method for capturing pronunciation errors based on speech recognition, extracting traditional acoustic features for classification. However, due to the single feature dimension and failure to consider semantics, accuracy decreases when dealing with homophones. Yu [4] designed an error correction method based on semantic matching for English translation robots, but its lack of coverage of colloquial expressions leads to a high misjudgment rate. Furthermore, the integration of human-machine interaction in learning necessitates more accurate recognition systems. To address the limitations of traditional methods, this paper proposes a refined pronunciation error recognition method for English conversation practice based on a dedicated artificial intelligence corpus, aiming to significantly improve recognition accuracy and provide high-quality feedback.

2 Construction of a Multimodal English Pronunciation Corpus

To provide rich, accurate, and targeted data support for subsequent research on pronunciation error recognition in English conversation practice, this paper constructs a specialized English pronunciation corpus. The construction of this multimodal corpus adheres rigorously to the principles of pertinence, representativeness, and scalability [5]. The key consideration factors during its construction are detailed in Table 1.

The construction steps are as follows.

Software Installation and Conversion: Install ELAN 6.0 and VLC media player. VLC is used to convert videos to the `.wav` audio format, maintaining the original file name and storage location to ensure synchronization [6].

Import and Preprocessing: Import the matching audio and video files into ELAN for synchronous playback, and save them as the `.eaf` format. This results in the "raw corpus" for subsequent filtering.

Multimodal Annotation System: The system is macroscopically divided into verbal (pronunciation and text) and non-verbal (paralinguistics and kinesics) categories [7]. Verbal error annotation is performed on transcribed TXT documents, and non-verbal annotation is conducted on video files using ELAN.

Multi-level Annotation: Annotation is performed in "Annotation Mode" for multi-level label-

Table 1: Principles and Consideration Factors for Multimodal Corpus Construction

Serial number	Consider dimensions	Specific requirements
(1)	Epochal character	Using contemporary English as the main material, we strive to restore the true appearance and specific context of the English language in practical dialogue applications, cover commonly used and popular English expressions, and reflect the latest language usage trends in the corpus.
(2)	Comprehensiveness	Reflected in the consideration of various aspects such as corpus type, subject matter, country, and content. The language materials include daily conversations, business exchanges, academic discussions, etc.; themes cover various fields such as life, culture, technology, education, etc.; collects English materials from different English-speaking countries (UK, US, Australia, etc.); ensures rich and diverse content, and the proportion of formal and informal language should be reasonably adjusted.
(3)	Typicality	The selected language must be both practical and authentic, fully considering the language and cultural connotations in cross-cultural communication. The corpus should reflect the usage characteristics of English in different cultural backgrounds, avoid obscure or uncommon expressions, ensuring representativeness.
(4)	Appropriateness	Ensures that the ideological content is positive, the language style is appropriate for the dialogue scenario, and the difficulty level is appropriate for the learner's basic level, avoiding being too simple or too complex.

ing, followed by "Segmentation Mode" and "Transcription Mode" to realize segmentation marking and subtitle transcription.

Corpus Retrieval: The construction is finalized by retrieving the `.eaf` files using keywords.

3 Extraction of Pronunciation Features for English Conversation Practice

This paper constructs a relevant mathematical model to describe the speech recognition system for pronunciation in English conversation practice. The mathematical expression of the speech recognition system for English conversation practice pronunciation can be represented by Formula (1):

$$S = \arg\max_{U \in \Xi} \left\{ Q(U|V) \times f^{\alpha \times \left(1 + \frac{\beta}{|U|}\right)} \right\} \tag{1}$$

In Equation (1), U denotes the text sequence of the English dialogue practice pronunciation, V represents the speech input of the English dialogue practice pronunciation, Ξ denotes the set of all possible text sequences, α represents the regularization parameter used to balance the effects of probability and sequence length, β denotes the parameter dynamically adjusted based on the complexity of the actual speech data for the English dialogue practice pronunciation, $|U|$ denotes the length of the text sequence U for the English dialogue practice pronunciation.

Based on the English dialogue practice pronunciation identified by formula (1), the pronunciation signal of the English dialogue practice pronunciation is then subjected to filtering detection. During the Discrete Wavelet Transform (DWT) processing, l represents the length of the pronunciation signal of the English dialogue practice pronunciation [8]. Its expression is shown in formula (2):

$$Y(l) = \frac{1}{T} \sum_{n=0}^{T-1} S \times \left[\cos\left(\frac{2\pi n l}{T + \varphi} \right) - j \sin\left(\frac{2\pi n l}{T + \varphi} \right) \right] \times \left(1 + \zeta \times \sin\left(\frac{2\pi n}{Q + \psi} \right) \right) \tag{2}$$

Where,$Y(l)$ denotes the complex-valued output of the filtered phonetic signal at discrete points l , n represents the time index,j denotes the imaginary unit, T indicates the duration of the phonetic segment, ζ signifies the adjustment coefficient,Q represents parameters related to the signal characteristics of the English dialogue practice pronunciation, and φ and ψ denote parameters used to adjust frequency calculations in the discrete Fourier transform.

This paper further extracts pronunciation features by defining a set of parameters. Let m denote the fluctuation extremum of the oral pronunciation frequency vibration during feature extraction for English dialogue practice pronunciation, and let m satisfy $m \in [m_{\min}, m_{\max}]$,where $m_{\min}$ and $m_{\max}$ represent the minimum and maximum values of the frequency vibration fluctuation extremum statistically derived from actual English dialogue practice pronunciation speech data [9].Let Q denote the extreme value of the frequency vibration trough for English conversation practice pronunciation, with a value range of $Q \in [Q_{\min}, Q_{\max}]$. $Q_{\min}$ and $Q_{\max}$ are the trough extreme value boundaries

determined by analyzing a large number of speech samples from English conversation practice pronunciation. Let R denote the correct period of the frequency audio for English conversation pronunciation practice, and $R > 0$. Let D denote the amplitude of the meson transmission frequency for English conversation, and $D \in [D_{\min}, D_{\max}]$.$D_{\min}$ and $D_{\max}$ are determined by the physical characteristics of the speech signal for English conversation pronunciation practice. Let BI denote the standard amplitude for spoken pronunciation in English conversation practice, with BI being a preset reference value. Let d denote the frequency parameter for English conversation pronunciation practice, and $d \in [d_{\text{start}}, d_{\text{end}}]$. d_{start} and d_{end} respectively represent the start and end points of the feature extraction process.

Based on these predefined parameters, the computational process for extracting pronunciation features from English dialogue practice can be expressed by formula (3):

$$G = \left(\frac{1}{d_{\text{end}} - d_{\text{start}} + 1} \sum_{d=d_{\text{start}}}^{d_{\text{end}}} \left(\frac{Q \times (D-1)}{R + \chi} \times \frac{1}{1 + \delta BI} \right) + v \times \left(1 - \cos\left(\frac{2\pi d}{O + \varepsilon} \right) \right) \right) \times Y(l) \quad (3)$$

In Equation (3), v denotes the adjustment coefficient, O represents a parameter related to the periodicity of the speech signal in English pronunciation practice, and χ , δ and ε denote adjustment factors.

The introduction of $v \times \left(1 - \cos\left(\frac{2\pi d}{O+\varepsilon}\right)\right)$ accounts for potential periodic fluctuations in speech signals from English conversation practice pronunciation, while new adjustment factors enhance the comprehensiveness of feature extraction.

4 Automatic Identification of Pronunciation Errors in English Conversation Practice

After completing feature extraction for English conversation practice pronunciation, the process proceeds to the automatic error recognition stage. To normalize feature scores within the $[0,1]$ range for more intuitive pronunciation quality assessment, this paper employs a cubic polynomial function combined with a logarithmic function, as shown in Formula (4):

$$F' = r_1 \cdot F^3 \cdot G_1 + r_2 \cdot F^2 \cdot G_2 + r_3 \cdot F \cdot G_3 + r_4 + l \cdot \ln\left(\sum_{i=1}^{n} \frac{\partial^i t}{\partial^i G} + \vartheta \right) \quad (4)$$

In the equation, F denotes the raw score of the pronunciation feature for the i th English conversation practice.r_1,r_2,r_3 and r_4 represent polynomial coefficients. l and ϑ denote small positive constants. $\sum_{i=1}^{n} \frac{\partial^i t}{\partial^i G}$ represents the sum of the first, second, and third derivatives of the English conversation practice pronunciation score feature G with respect to time t.

To achieve a more comprehensive and holistic evaluation of pronunciation quality in English conversation exercises, this paper introduces a multi-model scoring mechanism. The weighted average of scores from multiple models is regarded as the score of the test sample relative to the average

model constructed from these models [10]. Assuming there are N models H_1,H_2,.....,H_N with corresponding scores $F_1', F_2', \ldots, F_N'$ the average model score F'_{avg} can be expressed via formula (5):

$$F'_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\left(\rho \cdot F' + \tau \cdot Std(F') + \Phi \cdot Median(F')\right) \tag{5}$$

In formula (5), F'_{avg} denotes the average model score, $Std(F')$ represents the standard deviation of the i th English dialogue practice pronunciation model score, $Median(F')$ indicates the median of the i th English dialogue practice pronunciation model score, and ρ , τ and Φ denote adjustment coefficients.

The English dialogue pronunciation error recognition system constructed in this paper uses a diverse English dialogue pronunciation corpus as the comparison benchmark, setting a threshold of 0.6. Scores below this threshold are judged as pronunciation errors. Thus, this paper completes the research on English dialogue pronunciation error recognition methods based on this artificial intelligence corpus.

5 Experiments

5.1 Experimental Environment Setup and Corpus Construction

The experiment adopts a dual-operating system configuration. In terms of hardware, a Windows 11 system (equipped with an Intel i9 CPU and NVIDIA RTX 3080 GPU) and an Ubuntu 20.04 system (equipped with an AMD Ryzen 9 5950X CPU and NVIDIA RTX 3090 GPU) are used. The software adopts Python 3.8 and deep learning frameworks. The parameters of the corpus constructed in this paper are shown in Table 2.

Based on the phonetic characteristics of pronunciation and common error patterns, this paper classifies pronunciation errors in detail, and the classification is shown in the following table.

To more intuitively display the characteristics of the audio signals used in the experiment, this paper selects some representative audio signals for visualization.

It can be seen from the waveforms in Figure 1 that audio signals of different pronunciations have obvious differences in amplitude and frequency, providing a reference for pronunciation error recognition.

5.2 Experimental Results and Analysis

Pronunciation errors in actual English conversation practice are identified, and the recognition results are shown in Figure 2.

From the recognition results in Figure 2, it can be seen that the proposed method for pronunciation error recognition in English conversations achieves remarkable effects. In the waveform diagram, the pronunciation error regions are in sharp contrast with the fluctuation characteristics of normal speech, and the model can accurately capture abnormal fluctuations and correctly correspond to pronunciation errors.

Table 2: Parameter Table of English Pronunciation Error Recognition Corpus

Parameter category	Specific parameters	Parameter description
Language source	English learners' daily conversation recordings, english movie clips, english radio programs	Daily conversations reflect the true level of pronunciation; the movie clips contain a variety of accents; radio programs provide standard pronunciation comparison
Corpus size	Total duration of 500 hours (300 hours of learner dialogue, 150 hours of movie clips, 50 hours of radio programs)	Adequate scale ensures the reliability of results, and multiple types of language materials cover comprehensive pronunciation scenarios
Corpus format	Audio format: wav; sampling rate: 44.1khz; bit depth: 16 bits	Lossless format ensures audio quality, and parameter settings comply with speech analysis standards
Annotated content	Text transcription, pronunciation error location annotation, error type classification	Fine annotation supports error localization and pattern analysis
Annotation tool	Praat (speech analysis), elan (multimodal annotation)	Professional tools ensure the spatiotemporal accuracy of annotation (error¡50ms)
Data partitioning	Training set: 70% (350 hours); validation set: 15% (75 hours); test set: 15% (75 hours)	Stratified sampling to avoid data bias
Extended features	Support filtering data by accent (british/american) and level (beginner/intermediate/advanced)	Meet the needs of differentiated research

Table 3: Details of Pronunciation Error Types

Error type	Specific description	Example
Vowel error	There are deviations from standard pronunciation in terms of vowel length, pitch, and sound quality	Pronounced as/ɪ/,such as pronouncing "see" as a sound similar to "si", affects the accurate pronunciation of the word
	The opening degree of vowel pronunciation is inconsistent with the standard pronunciation	Translate/*a*/into /ʌ/, for example, pronounce "car" as a sound similar to "cur" to change the original pronunciation of the word
Consonant error	Errors occur in the aspirated, voiced, and pronounced parts of consonant pronunciation	Pronouncing/θ/as/s/, such as pronouncing "think" as "sink", leads to incorrect pronunciation of the word
	The duration of consonant pronunciation does not match the standard pronunciation	Pronunciation of/t/at the end of the word is too long, such as "cat" being pronounced as a sound similar to "ca-t", which affects the naturalness of pronunciation
Concatenation error	Not correctly connected where it should be	Not linking the final consonant/k/of "look" with the initial vowel/α/of "at" in "lookat" disrupts the fluency of the sentence
	Incorrect linking in inappropriate places	Incorrect linking of 'an's /n/ with 'apple's /æ/ in "anapple", causing confusion in pronunciation
Accent error	Incorrect placement of stress on words or sentences	Place the stress of 'record' (noun, with emphasis on the first syllable) on the second syllable, similar to the pronunciation of 'record' (verb), to change the semantic meaning of the word
	Error in stress distribution in polysyllabic words	Incorrect emphasis on 'photograph' affects the correct understanding and pronunciation of words
Intonation is wrong	The intonation pattern of the sentence does not conform to english expression habits, such as interrogative sentences not using rising tones, etc	The declarative sentence uses the intonation of a interrogative sentence, or the interrogative sentence uses the intonation of a declarative sentence, resulting in inaccurate semantic communication
	The intonation fluctuation of the sentence does not meet the contextual requirements	When expressing strong emotions, the tone is too flat and cannot accurately convey the emotions

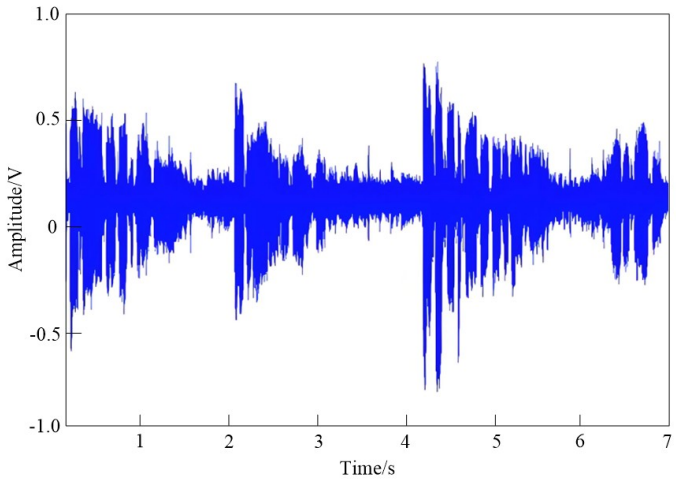

Fig. 1. Waveforms of Some Representative Audio Signals

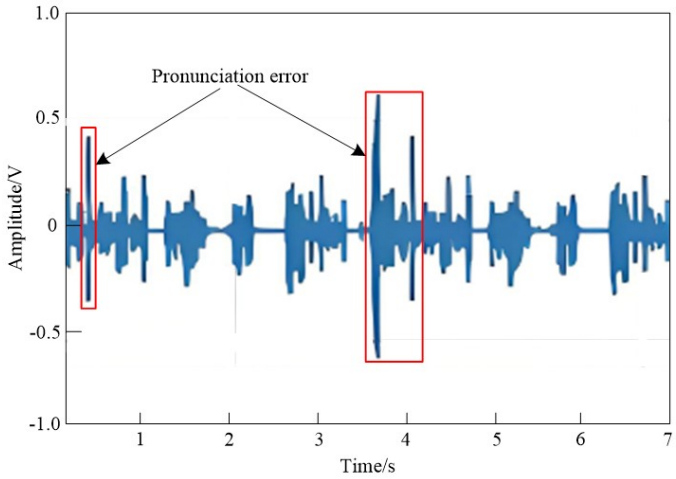

Fig. 2. Pronunciation Error Recognition Results of English Conversation Practice Based on Artificial Intelligence Corpus

The outstanding effect of this method is attributed to the rich and diverse corpus, which covers pronunciation samples of different accents and proficiency levels, enabling the model to learn various pronunciation error feature patterns. At the same time, the model adopts advanced deep learning algorithms to automatically extract high-level abstract features, and combines multimodal

information to further improve the accuracy of pronunciation error recognition.

To evaluate the performance of the pronunciation error recognition model for English conversation practice based on the artificial intelligence corpus, this paper analyzes the key indicator of the PR curve, as shown in Figure 3.

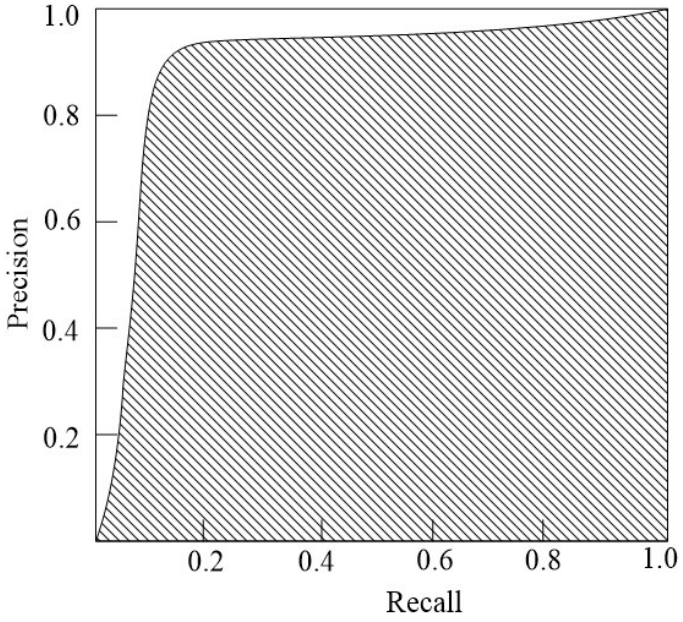

Fig. 3. PR Curve of Pronunciation Error Recognition for English Conversation Practice Based on Artificial Intelligence Corpus

From the PR curve in Figure 3, the performance of the pronunciation error recognition model for English conversations based on the artificial intelligence corpus can be observed: as the recall rate increases, the precision rate gradually decreases but the curve is in a high position, demonstrating the superiority of the method.

The good performance of the proposed method benefits from the carefully constructed corpus. It covers rich scenarios, diverse pronunciation levels and accents, providing comprehensive training data for the model. The model training adopts an efficient algorithm architecture, which can mine the implicit information of speech, accurately distinguish the correctness of pronunciation, and balance precision and recall, enabling it to fully capture pronunciation errors and reduce misjudgments.

The proposed recognition method based on the artificial intelligence corpus is set as the experimental group, and representative methods from references [1], [3], and [4] are selected as comparison objects, denoted as Comparison Method 1, Comparison Method 2, and Comparison Method 3 respectively. Tests are conducted under the same experimental environment and dataset conditions, and the F1 score results corresponding to the four methods are shown in Table 4.

Table 4: Comparison of F1 Scores of Different Methods in Pronunciation Error Recognition for English Conversation Practice

Number of english dialogue samples (pieces)	Proposed method F1 score	Comparison method 1 F1 score	Comparison method 2 F1 score	Comparison method 3 F1 score
100	0.87	0.62	0.58	0.60
200	0.89	0.65	0.61	0.63
300	0.90	0.68	0.64	0.66
400	0.91	0.70	0.66	0.68
500	0.92	0.72	0.68	0.70
600	0.94	0.73	0.69	0.71
700	0.95	0.74	0.70	0.72
800	0.96	0.75	0.71	0.73
900	0.98	0.76	0.72	0.74
1000	0.99	0.77	0.73	0.75

From the F1 score comparison in Table 4, it can be seen that there are differences in the performance of each method. The proposed pronunciation error recognition method for English conversations based on the artificial intelligence corpus has significant advantages under all sample sizes. When the number of samples increases from 100 to 1000, its F1 score continues to rise and is far higher than other methods.

This is due to the strong support of the corpus, which covers diverse pronunciation scenarios, accents and error types, helping the model learn comprehensive pronunciation features. At the same time, the proposed method may adopt more advanced algorithms, which can deeply mine key speech information and accurately distinguish the correctness of pronunciation. The comparison methods perform poorly due to limitations in corpus or algorithms, thus proving the robustness and efficiency of the proposed method.

The confusion matrix can clearly show the classification of each method in recognizing correct and incorrect pronunciations. The confusion matrix results corresponding to the four methods are shown in Figure 4.

It can be seen from Figure 4 that the proposed pronunciation error recognition method for English conversations based on the artificial intelligence corpus has obvious advantages. In terms of true positives, the number of correct pronunciations recognized by this method is far more than that of the comparison methods, which benefits from the rich and diverse correct pronunciation samples in the corpus.

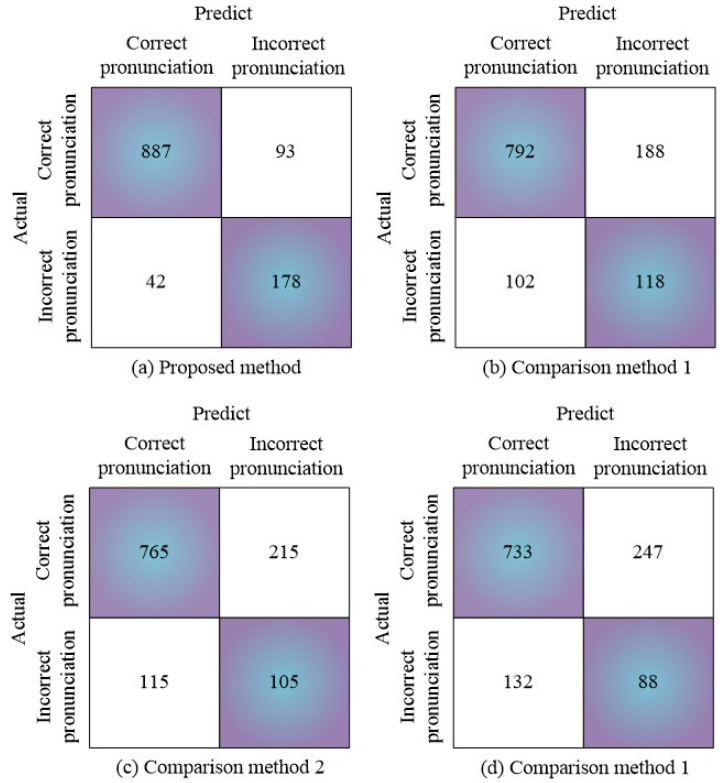

Fig. 4. Confusion Matrix Comparison of Different Methods in Pronunciation Error Recognition for English Conversation Practice

6 Conclusion

This study constructs a multimodal English pronunciation corpus and accurately extracts English conversation pronunciation features by combining mathematical models and algorithms. Experimental results show that the proposed method has a high PR curve position, and the F1 score significantly increases from 0.87 to 0.99 under different sample sizes, with better classification performance in the confusion matrix. This method addresses the limitations of traditional methods, improves recognition accuracy by introducing multi-model scoring and normalization functions, can provide refined error correction for learners, enhance their pronunciation and oral expression abilities, and also provide new ideas for related applications. This study successfully constructed a multimodal English pronunciation corpus and developed an automated method that accurately extracts English conversation pronunciation features by combining a proposed set of non-linear mathematical models and advanced feature engineering. Experimental results confirmed the method's

effectiveness, showing a high PR curve position, and the F1 score significantly increased from 0.87 to 0.99 under different sample sizes, along with superior classification performance in the confusion matrix. This method effectively addresses the limitations of traditional approaches, significantly improves recognition accuracy by introducing multi-model scoring and a sophisticated normalization function, and can provide refined error correction for learners, thereby enhancing their pronunciation and oral expression abilities.

Acknowledgments

The authors acknowledge the Second Batch of Modern Industry School in Sichuan Province: "Industry School of Artificial Intelligence Media and Software" (project No. [2023]263).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Dennis NK. Using AI-Powered Speech Recognition Technology to Improve English Pronunciation and Speaking Skills. IAFOR Journal of Education. 2024;12(2).

[2] Bashori M, Hout RV, Strik H, Cucchiarini C. I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems. Innovation in Language Learning and Teaching. 2024;18(5):443-61.

[3] Dou W. A Method for Capturing English Oral Pronunciation Errors based on Speech Recognition. International Journal of Computing Science and Mathematics. 2024;1(1).

[4] Yu X. English translation robot pronunciation error correction method based on semantic matching. International Journal of Biometrics. 2025;17(1-2):151-69.

[5] Sardegna VG. Evidence in favor of a strategy-based model for English pronunciation instruction. Language Teaching. 2022;55(3):16.

[6] Iskandar I, Dewanti R, Sulistyaningrum SD, Santosa I. Scaffolding Assignments to Conciliate the Disinclination to Employ Project-Based Learning of English Pronunciation and Autodidacticism. International Journal of Language Education. 2024;8(2).

[7] Utami HS, Morganna R. Improving Students' English Pronunciation Competence by Using Shadowing Technique. ENGLISH FRANCA : Academic Journal of English Language and Education. 2022;6(1):127.

[8] Calvano M, Curci A, Pagano A, Piccinno A. Speech Therapy Supported byAI andSmart Assistants. Lecture Notes in Computer Science. 2024:97-104.

[9] Hoang NT, Han DN, Le DH. Exploring Chatbot AI in improving vocational students' English pronunciation. AsiaCALL Online Journal. 2023;14(2):140-55.

[10] Prashant P. Importance Of Pronunciation In English Language Communication. Working papers. 2018.