SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(4), 652.
https://doi.org/10.62617/mcb652

Article

Optimizing English pronunciation teaching through motion analysis and intelligent speech feedback systems

Jieru Wang

Department of Economic Management, Weihai Ocean Vocational College, Weihai 264200, China; JieruWang1001@outlook.com

CITATION

Wang, J. Optimizing English pronunciation teaching through motion analysis and intelligent speech feedback systems. Molecular & Cellular Biomechanics. 2024; x(x): 652.
https://doi.org/10.62617/mcb652

ARTICLE INFO

Received: 28 October 2024
Accepted: 4 November 2024
Available online: 20 December 2024

Abstract: This study investigates the effectiveness of integrating Motion Analysis (MA) and Intelligent Speech Feedback Systems (ISFS) to enhance English Pronunciation (EP) accuracy among Chinese learners. Leveraging the OptiTrack Prime 13 Motion Capture System (MCS) and SpeechAce Pronunciation API, the study aims to address challenges non-native English speakers face, particularly in producing accurate articulatory movements and reducing Pronunciation Errors. Forty-three participants were divided into Experimental Groups (EG) and Control Groups (CG), with the EG receiving real-time feedback on articulation and phoneme accuracy. Key metrics, including Pronunciation Accuracy Score (PAS), Articulatory Movement Score (AMS), and Pronunciation Error Rate (PER), were measured alongside engagement indicators, such as session duration and self-corrections. The results show that the EG experienced a significant improvement in pronunciation accuracy, with a 31.2% increase in PAS and a 57.1% reduction in PER. Enhanced AMS scores also indicated refined articulatory precision across various articulatory points, including lip rounding and tongue positioning. Engagement metrics demonstrated higher consistency and task completion rates in the EG, suggesting increased motivation and sustained participation due to the real-time feedback provided. These findings suggest that combining MA with ISFS can provide targeted, adaptive support, enabling learners to make precise corrections and accelerate their progress in achieving native-like pronunciation. This study contributes valuable insights into the potential of advanced feedback-driven approaches in language learning and pronunciation training.

Keywords: Motion Analysis; Intelligent Speech Feedback; articulatory precision; Pronunciation Error reduction; English Pronunciation; non-native speakers

1. Introduction

Effective pronunciation is fundamental to successful communication in English, yet achieving native-like accuracy poses challenges for non-native speakers, particularly those whose first languages differ significantly in phonetic structure [1,2]. For Chinese learners, mastering English Pronunciation (EP) frequently involves overcoming unique articulatory and phonological challenges, such as accurately producing certain vowel sounds, consonant clusters, and stress patterns [3–5]. Traditional pronunciation teaching methods, which often rely on auditory feedback alone, may not fully address these complexities or offer the targeted support learners require to make precise articulatory adjustments [6–8].

In recent years, advances in technology have enabled new methods for enhancing Pronunciation Learning (PL), primarily through Motion Analysis (MA) and Intelligent Speech Feedback Systems (ISFS) [9–12]. MA tools, such as high-precision Motion Capture Systems (MCS), provide detailed insights into facial articulation by tracking movements of key articulation points like the lips, jaw, and tongue [13,14]. When processed with Machine Learning (ML) algorithms, this data can suggest learners'

real-time feedback on their articulation, allowing for immediate adjustments and focused practice [15,16]. Similarly, ISFS can analyze phoneme-level details of learners' pronunciation and identify areas where they deviate from native English patterns [17]. These systems can support more efficient and effective PL by highlighting specific areas for improvement and suggesting targeted corrections [18].

Despite the potential of these technologies, there has been limited research on their combined use in structured pronunciation training for non-native English speakers, particularly in the context of Chinese learners [19–21]. This study aims to address this gap by evaluating the effectiveness of a hybrid approach that integrates MA and ISFS. Using the OptiTrack Prime 13 MCS and SpeechAce Pronunciation API, this study investigates how real-time articulatory and phonetic feedback can impact pronunciation accuracy, error reduction, and learner engagement.

The study is structured to explore three primary research questions:

1) To what extent does real-time articulatory feedback improve the precision of articulatory movements?
2) How does phoneme-level feedback from an ISFS affect error reduction in pronunciation?
3) How does the combined feedback approach impact learner engagement and consistency in pronunciation practice?

The remainder of this paper is organized as follows. Section 2 presents the methodology, detailing the participant selection, data collection process, and tools used for MA and ISFS. Section 3 provides a detailed analysis of the results, with subsections on Articulatory Movement Scores, Pronunciation Error Rates, and engagement metrics. Section 4 concludes the paper, presenting visions for future directions for research and applications in PL technology.

2. Methodology

2.1. Population

This study's participants comprised an initial pool of 60 English learners from various regions of China. After a screening process based on eligibility criteria—such as their willingness to participate in all phases of the study and their availability for the required training sessions—43 participants were selected as the final cohort. The selection process involved filtering out individuals who did not meet the proficiency level requirements (i.e., beginner to intermediate English speakers) or those who could not commit to the study timeline [22–25].

The 43 participants represented a diverse range of linguistic and educational backgrounds. All participants were native speakers of Chinese, with 37 being Mandarin speakers, while the remaining 6 spoke other dialects, such as Cantonese and Hokkien. The age range of the participants spanned from 18 to 40 years, with the majority (approximately 65%) between 18 and 30 years old, reflecting a demographic typical of university students and young professionals.

Of the 43 participants, 26 were male, and 17 were female. Their English proficiency levels were assessed through a preliminary language test, ensuring a range of abilities from lower beginner (A1) to upper intermediate (B2) on the Common European Framework of Reference for Languages (CEFR) scale. This distribution

comprehensively evaluated how learners with different proficiency levels responded to the Motion Analysis and intelligent Speech Feedback Systems.

Four participants withdrew During the study due to personal scheduling conflicts, leaving a final cohort of 39 learners. This group completed the entire training and testing regimen, providing sufficient data for the study's analysis. No participants were excluded due to language or technological issues during the study, as the systems used were designed to accommodate varying levels of familiarity with digital tools. This final cohort served as the basis for analyzing the effectiveness of the pronunciation teaching systems, ensuring that the sample was sufficiently diverse to capture relevant insights into the broader population of English learners in China.

2.2. MA techniques

The MA in this study was performed using the OptiTrack Prime 13-MCS, a state-of-the-art tool designed for detailed facial tracking and articulation analysis. This system is widely used in linguistic and phonetic research due to its high accuracy and ability to capture minute facial movements critical for speech production [26–30].

The OptiTrack Prime 13-MCS comprises a network of high-resolution infrared cameras and reflective facial markers specifically designed for speech and articulation studies. The cameras capture real-time facial dynamics with sub-millimeter precision, focusing on key articulation points like lips, jaw, and cheeks. When tracked by the cameras, these markers produce highly accurate data that can be analyzed to assess pronunciation.

Additionally, the system incorporates depth-sensing technology, which enables the creation of a detailed 3D model of the participant's facial movements. This technology allows for in-depth analysis of articulation from multiple perspectives, including side and vertical views, which are essential for capturing the full range of motions involved in producing specific English phonemes. The OptiTrack system is supported by DigiFace Software, which integrates ML algorithms that are pre-trained to recognize and evaluate standard articulatory patterns of EP [31–34].

The MCS begins by attaching small reflective markers to key points on the participant's face, such as the lips, jawline, and the area surrounding the mouth. The OptiTrack Prime 13 cameras then track these markers as the participant reads aloud a set of English words and phrases. The MCS movements at a high frame rate, collecting detailed data on how each speech sound is physically articulated.

The data from OptiTrack Prime 13 are processed using DigiFace Software, which converts the recorded facial movements into a series of numerical coordinates representing each articulatory gesture. These coordinates are mapped to a phonetic model of correct EP, allowing real-time analysis of discrepancies between the participant's speech patterns and native EP standards. The software's ML capabilities help identify common pronunciation challenges specific to native Chinese speakers, such as difficulties with vowel production or consonant clusters.

The processed data are then used to generate real-time feedback during practice sessions. This feedback may include visual cues illustrating correct mouth positioning and detailed recommendations for articulation adjustments. The system also proposes comprehensive post-session analysis, allowing participants to review their progress

over time. By utilizing the OptiTrack Prime 13-MCS and DigiFace Software, the study ensures precise, real-time tracking of facial movements, providing learners with targeted, actionable feedback on their pronunciation efforts.

2.3. Speech Feedback System

The IDFS employed in this study was powered by SpeechAce Pronunciation API, a widely used solution for assessing and improving pronunciation in language learning applications. This system leverages advanced speech recognition algorithms and ML designed to evaluate non-native pronunciation. It provides learners immediate feedback on their spoken English, identifying pronunciation issues at the phoneme level.

The SpeechAce Pronunciation API integrates seamlessly with audio recording hardware to capture and analyze speech in real time. When a participant pronounces a word or phrase, the system processes the audio to compare the phonetic characteristics of the participant's pronunciation against a model of native EP. The system analyzes multiple aspects of speech, including sound accuracy, intonation, and stress patterns, providing a detailed breakdown of pronunciation quality for each phoneme.

The feedback generated by SpeechAce is displayed to learners through a visual interface highlighting specific sounds needing improvement. For instance, if a participant mispronounces a vowel, the system visually marks the sound and proposes ideas to adjust the vocal tract position, such as raising the tongue or modifying lip shape. This targeted feedback is instrumental in helping learners understand precisely how and where to improve their pronunciation.

Additionally, SpeechAce incorporates an error-detection feature that flags common pronunciation issues experienced by native Chinese speakers, such as challenges with certain English vowel sounds or consonant clusters. The system recognizes and adapts to each learner's specific pronunciation patterns through its ML algorithms, enabling personalized guidance that evolves over time. This adaptability ensures that the feedback remains relevant and practical as the learner progresses.

2.4. Measurements and variables

In this study, various measurements were conducted to evaluate the effectiveness of MA and ISFS in enhancing EP. These measurements focused on pronunciation accuracy and user engagement to comprehensively assess the system's impact.

The Pronunciation Accuracy Score (PAS) served as a primary metric, representing the accuracy of participants' pronunciation compared to a native English standard. This score, ranging from 0 to 100, was generated for each phoneme, word, and phrase to track improvements across different linguistic levels. Scores were collected at three intervals: Baseline (before the introduction of the system), midway through the study, and at the conclusion. This allowed for a detailed understanding of progress over time.

The Articulatory Movement Score (AMS) was also introduced to evaluate the precision and consistency of facial movements associated with correct pronunciation. Data captured by the OptiTrack Prime 13-MCS and analyzed with DigiFace Software produced scores that reflected alignment with an English articulation standard.

Specific articulation points were monitored, such as lip rounding, jaw movement, and tongue positioning. Scores were assessed during key pronunciation exercises to track improvements in participants' articulation accuracy and consistency.

Pronunciation Error Rate (PER) was another essential measurement, recording the frequency of mispronunciations at the phoneme level. The SpeechAce Pronunciation API flagged and categorized errors, helping to identify specific patterns, such as common vowel or consonant mispronunciations. This rate was assessed continuously, with a summary calculated at the end of each session, allowing researchers to observe reduced errors over time.

User engagement was also monitored through metrics such as session duration, the number of pronunciation attempts, and self-correction frequency. A survey conducted at the end of the study provided qualitative feedback on the participants' experiences, gathering insights into the perceived ease of use, effectiveness, and overall satisfaction with the system. This feedback complemented the quantitative data, proving a subjective perspective on the system's impact on user engagement and motivation.

From **Table 1** the independent variables in this study included the feedback system type, comparing real-time feedback from the SpeechAce Pronunciation API with a baseline condition of no feedback, and the pronunciation task type, which varied between isolated words, phrases, and sentences. The dependent variables, encompassing PAS, AMS, PER, and user engagement, provided objective and subjective insights into pronunciation improvement and learner satisfaction with the combined MA and ISFS. These measurements and variables contributed to a detailed evaluation of the system's role in supporting EP learning.

Table 1. Measurements, units, and variables.

Measurement	Description	Unit	Variable Type
Pronunciation Accuracy Score (PAS)	Measures accuracy of pronunciation compared to native standard, assessed at phoneme, word, and phrase levels.	Score (0–100)	Dependent Variable
Articulatory Movement Score (AMS)	Evaluates precision and consistency of facial movements during articulation.	Score (0–100)	Dependent Variable
Pronunciation Error Rate (PER)	It counts the frequency of mispronunciations and categorizes them by type (e.g., vowel or consonant).	Error count	Dependent Variable
User Engagement	Tracks participant interaction with the system (session duration, attempts, self-corrections).	Various (time, count)	Dependent Variable
Feedback System Type	Type of pronunciation feedback system used (e.g., real-time vs. baseline with no feedback).	Type (categorical)	Independent Variable
Pronunciation Task Type	Context of pronunciation tasks, such as isolated words, phrases, or sentences.	Type (categorical)	Independent Variable

2.5. Experiment design and data collection

The experiment was designed to assess the impact of combining MA and ISFA on improving EP among Chinese learners. The study utilized a pre-post design with repeated measures, allowing for an in-depth comparison of participants' pronunciation performance across different stages. Participants were divided into two groups: One receiving real-time feedback from the SpeechAce Pronunciation API and OptiTrack Prime 13-MCS and a baseline Control Group (CG) with no feedback. Each group

underwent similar pronunciation tasks, allowing for a controlled comparison of outcomes.

Data collection occurred in three main phases: Pre-test, training sessions, and post-test. In the pre-test phase, baseline data on pronunciation accuracy, articulatory movement, and error rate were collected. During this phase, participants completed a series of pronunciation tasks, including isolated words, phrases, and sentences, which were recorded and analyzed to establish a baseline for each participant's current proficiency. The pre-test provided initial Pronunciation Accuracy Scores (PAS), Articulatory Movement Scores (AMS), and Pronunciation Error Rates (PER), along with measures of user engagement in terms of task completion time and self-corrections.

The training sessions took place over four weeks, with participants engaging in bi-weekly sessions designed to practice and refine their pronunciation. Each session began with warm-up exercises to familiarize participants with the system and the pronunciation tasks. In the Experimental Group (EG), real-time feedback was provided by the SpeechAce Pronunciation API, offering phoneme-level insights into Pronunciation Errors and specific articulatory adjustments. The OptiTrack Prime 13-MCS facial movements provided additional visual feedback, enabling participants to view their articulatory positions compared to the ideal model. The baseline group completed the same tasks without feedback, allowing researchers to assess the impact of the feedback systems on learning outcomes.

Throughout the training sessions, data were continuously recorded, capturing each participant's Pronunciation Accuracy Scores, Articulatory Movement Scores, and Pronunciation Error Rates. Real-time measurements from the MA were used to monitor improvements in articulation precision, while SpeechAce recorded phoneme-level pronunciation scores. Session duration, the number of attempts, and self-corrections were also logged to assess engagement and persistence, which were later analyzed for correlations with pronunciation improvement.

In the final post-test phase, participants from both groups completed the same pronunciation tasks they had performed in the pre-test phase. This phase was designed to evaluate progress made over the study period. The post-test data provided updated Pronunciation Accuracy Scores, Articulatory Movement Scores, and Pronunciation Error Rates, which were compared against pre-test scores to measure improvement. The real-time engagement data collected throughout the training sessions were reviewed to identify patterns in learning behavior, particularly how immediate feedback influenced participants' ability to self-correct and refine their pronunciation over time.

3. Results

3.1. AMS analysis

The AMS analysis (**Table 2** and **Figure 1**) examines improvements in articulation precision by comparing pre-test and post-test scores for specific articulation points. The results indicate significant enhancements in AMS for the EG, which received real-time MA feedback compared to the CG. The EG's mean AMS increased from 55.8 to 76.5, an improvement of 37.1%, while the CG showed a more

minor increase from 54.6 to 64.9, an 18.9% improvement. This indicates that real-time feedback had a substantial positive effect on the EG's overall articulation precision, with nearly double the improvement seen in the CG. The EG showed a remarkable improvement for lip rounding, increasing from a mean score of 51.2 to 74.1, representing a 44.7% enhancement. The CG's improvement was less pronounced, from 52.7 to 62.3, achieving only an 18.2% increase. Lip rounding, essential for accurate vowel and consonant production, particularly benefited from the motion feedback, allowing participants to make targeted adjustments.

The EG's jaw movement score rose from 57.4 to 79.2, marking a 37.9% improvement. The CG's score improved from 56.1 to 67.5, a 20.3% increase. Jaw movement is critical in controlling the vertical positioning of the mouth, affecting both vowel and consonant pronunciation, and the EG's access to feedback helped them make more precise adjustments. The EG saw a 43.9% improvement in tongue height, increasing from a pre-test mean of 49.6 to a post-test mean of 71.4. The CG improved by 19.2%, from 50.5 to 60.2. Tongue height is essential for producing distinct vowel sounds, and the EG's higher improvement suggests that motion feedback was beneficial in helping participants refine their tongue positioning.

For tongue advancement, the EG's score increased from 52.3 to 72.8, resulting in a 39.2% improvement. The CG's score rose from 51.8 to 61.1, an 18.0% increase. Accurate tongue advancement, or front-to-back movement, is essential for differentiating between front, central, and back vowels, and real-time motion feedback appeared to assist the EG in making precise adjustments. The EG improved from a pre-test mean of 53.9 to a post-test mean of 74.5, achieving a 38.2% increase in lip-spreading accuracy. In contrast, the CG's mean score improved by 19.6%, from 52.6 to 63.0. Lip spreading contributes to the clarity of various sounds, especially in differentiating rounded and unrounded vowels, and the EG's access to feedback facilitated a greater degree of precision. Velum position, which affects nasality in speech sounds, saw a 39.2% improvement in the EG, rising from a mean of 50.7 to 70.6. The CG improved from 51.3 to 61.4, an increase of 19.7%. Real-time feedback helped the EG better control velum positioning, an articulation point that can be challenging for non-native speakers.

Table 2. Results for articulatory movement score.

Articulation Point	Pre-Test AMS Mean (EG)	Post-Test AMS Mean (EG)	Improvement (%) (EGl)	Pre-Test AMS Mean (CG)	Post-Test AMS Mean (CG)	Improvement (%) (CG)
Overall AMS	55.8	76.5	37.1	54.6	64.9	18.9
Lip Rounding	51.2	74.1	44.7	52.7	62.3	18.2
Jaw Movement	57.4	79.2	37.9	56.1	67.5	20.3
Tongue Height	49.6	71.4	43.9	50.5	60.2	19.2
Tongue Advancement	52.3	72.8	39.2	51.8	61.1	18.0
Lip Spreading	53.9	74.5	38.2	52.6	63.0	19.6
Velum Position	50.7	70.6	39.2	51.3	61.4	19.7

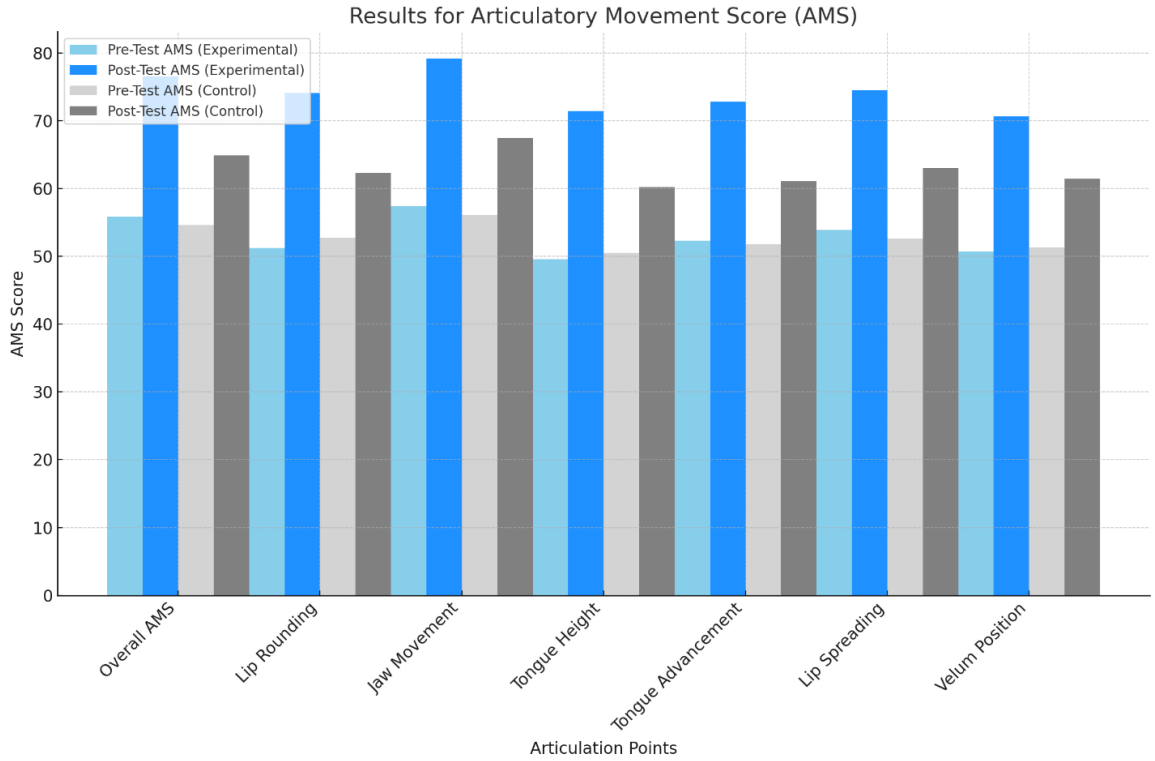


Figure 1. Articulatory movement score analysis.

3.2. PER analysis

The PER analysis (**Table 3** and **Figure 2**) reveals a substantial reduction in errors across multiple categories for the EG compared to the CG, indicating the effectiveness of real-time feedback and MA. The EG achieved a significant reduction in overall PER, from an initial 24.7% to 10.6%, marking a 57.1% reduction. In contrast, the CG reduced overall PER from 25.1% to 18.3%, a 27.1% reduction. This substantial difference underscores the impact of the ISFS in helping learners correct their Pronunciation Errors more effectively.

Vowel errors saw the most significant improvement, with the EG reducing errors from 13.5% to 5.3%, a 60.7% reduction. The CG achieved only a 24.1% reduction, from 14.1% to 10.7%. Given the complexity of vowel sounds for non-native speakers, this result highlights the system's effectiveness in guiding users to refine vowel pronunciation accurately. The EG reduced consonant errors from 11.2% to 5.2%, a 53.6% decrease, compared to a 30.9% reduction in the CG, which dropped from 11.0% to 7.6%. This improvement shows that real-time articulatory feedback played a significant role in assisting learners to correct consonant mispronunciations.

Diphthongs involving gliding vowel sounds were reduced by 55.9% in the EG, from an initial 9.3% to 4.1%. The CG achieved a 28.9% reduction, from 9.7% to 6.9%. This improvement suggests that feedback on mouth and tongue positioning effectively addressed the complexity of diphthong pronunciation. The EG reduced stress errors from 10.8% to 4.6%, a 57.4% reduction, whereas the CG reduced these errors from 10.5% to 7.7%, a 26.7% decrease. Stress accuracy is essential for fluency and natural intonation, and the ISFS appears to have enabled the EG to improve their stress application significantly.

The EG reduced intonation errors by 48.6%, from 7.4% to 3.8%, while the CG refers to a 22.2% reduction, from 7.2% to 5.6%. Intonation, which involves pitch variation, benefited from real-time guidance, helping the EG achieve more native-like intonation. For consonant clusters, the EG's errors decreased from 8.6% to 4.3%, a 50.0% reduction, while the CG reduced cluster errors by 28.1%, from 8.9% to 6.4%. Clusters can be challenging for non-native speakers, and the higher reduction in the EG indicates that the feedback system effectively supported the accurate production of these sounds. Voicing errors, which occur when sounds are mispronounced as voiced or unvoiced, were reduced by 53.3% in the EG, from 9.0% to 4.2%. The CG achieved a 29.3% reduction, from 9.2% to 6.5%. This improvement suggests that real-time feedback helped participants distinguish precisely between voiced and unvoiced sounds.

Table 3. Results for Pronunciation Error Rate.

Error Type	Initial PER (EG)	Final PER (EG)	Reduction (%) (EG)	Initial PER (CG)	Final PER (CG)	Reduction (%) (CG)
Overall PER	24.7	10.6	57.1	25.1	18.3	27.1
Vowel Errors	13.5	5.3	60.7	14.1	10.7	24.1
Consonant Errors	11.2	5.2	53.6	11.0	7.6	30.9
Diphthong Errors	9.3	4.1	55.9	9.7	6.9	28.9
Stress Errors	10.8	4.6	57.4	10.5	7.7	26.7
Intonation Errors	7.4	3.8	48.6	7.2	5.6	22.2
Cluster Errors	8.6	4.3	50.0	8.9	6.4	28.1
Voicing Errors	9.0	4.2	53.3	9.2	6.5	29.3

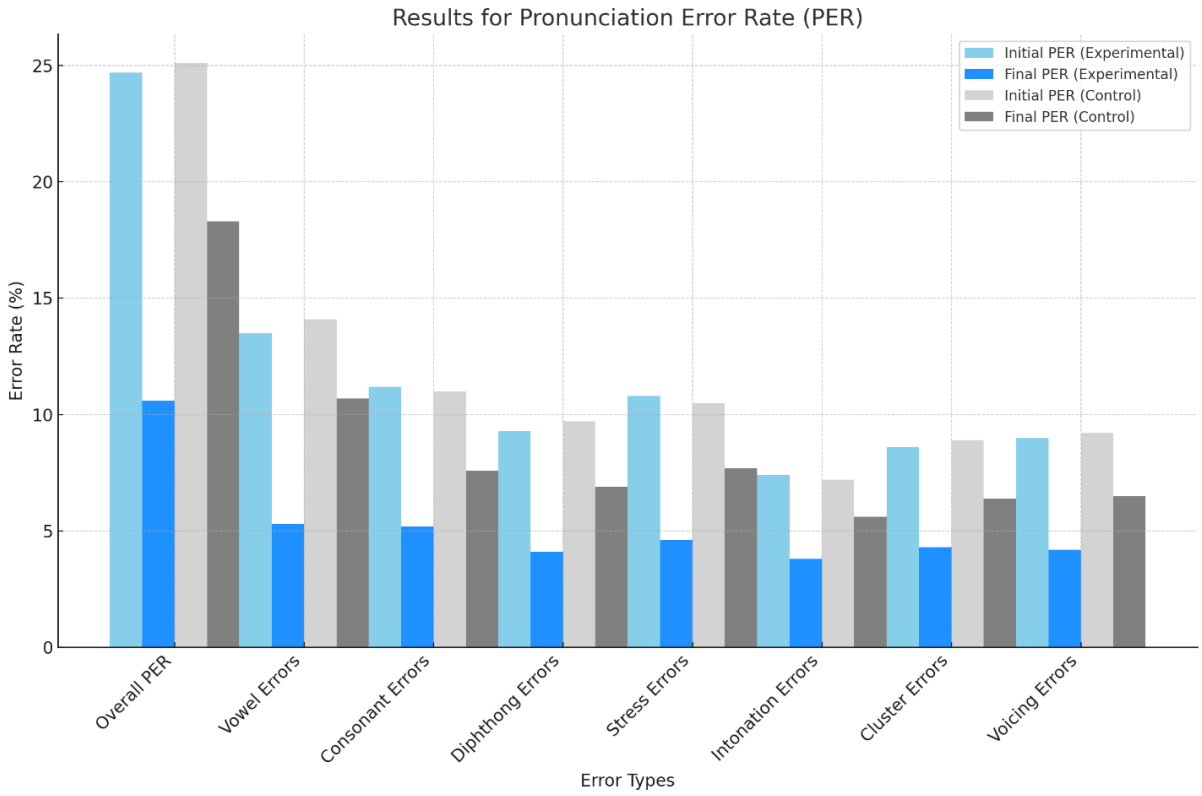


Figure 2. Pronunciation Error Rate analysis.

3.3. Engagement and Effort Metrics analysis

The Engagement and Effort Metrics analysis (**Table 4** and **Figure 3a,b**) examines the participants' interaction with the IFFS and their overall effort, comparing the EG (which received real-time feedback) with the CG. More robust Engagement and Effort Metrics in the EG highlight the positive impact of the ISFS on motivation and sustained participation. The EG spent an average of 45.3 min per session, compared to 38.6 min in the CG. The correlation between session duration and PAS improvement was moderate in the EG (0.68) but lower in the CG (0.42). A similar pattern was observed for AMS improvement, with a correlation of 0.71 in the EG compared to 0.45 in the CG. These results suggest that extended session duration, likely encouraged by real-time feedback, positively influenced pronunciation and articulation improvements.

Participants in the EG attempted pronunciation tasks 28 times on average per session, compared to 20 attempts in the CG. The number of attempts correlated highly with PAS improvement in the EG (0.75) and AMS improvement (0.78), while lower correlations were observed in the CG (0.47 and 0.51, respectively). This indicates that real-time feedback may have motivated participants to attempt more repetitions, leading to more significant improvement. The EG had a self-correction frequency of 12 per session, while the CG averaged 8. The correlation with PAS improvement was moderate for the EG (0.64) but lower for the CG (0.39). The trend was similar for AMS improvement, with correlations of 0.66 in the EG and 0.43 in the CG. These findings suggest that access to feedback encouraged participants to actively identify and correct their mistakes, promoting more rapid learning.

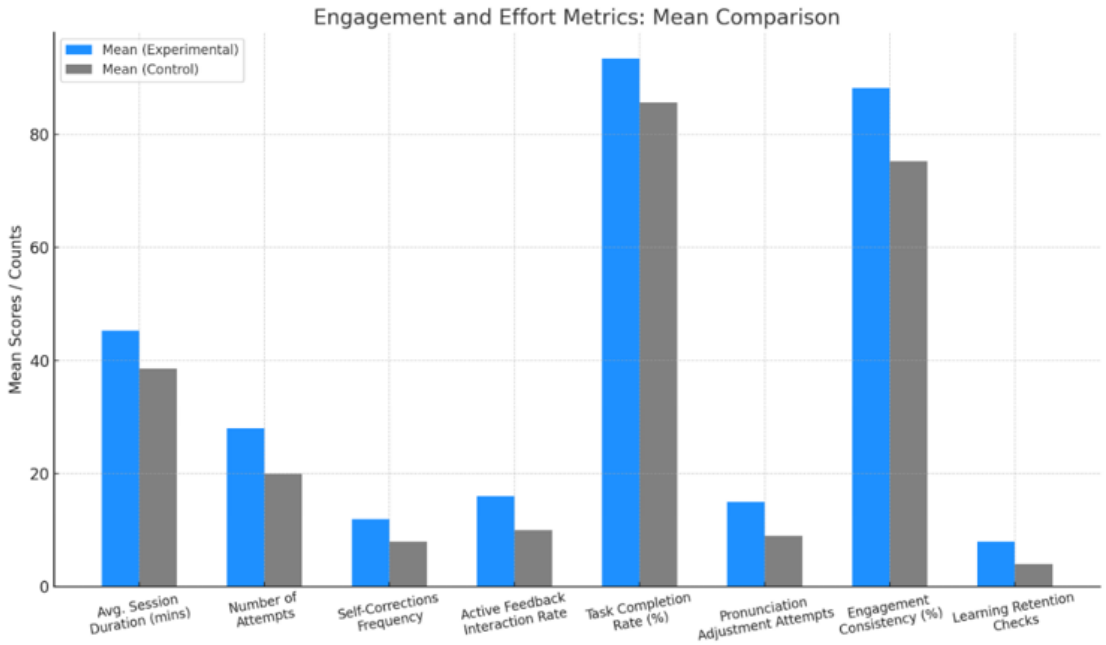
The EG actively interacted with the ISFS (e.g., pausing, replaying, or adjusting) 16 times per session, compared to 10 times in the CG. The correlation with PAS improvement was 0.70 for the EG and 0.45 for the CG. For AMS improvement, the correlations were 0.69 and 0.46, respectively. This shows that higher interaction with the ISFS was associated with more significant learning gains, especially in the EG, where feedback was available. The task completion rate was higher in the EG (93.4%) compared to the CG (85.6%). The correlation between task completion rate and PAS improvement was 0.76 in the EG and 0.50 in the CG, with a similar pattern for AMS improvement (0.74 for EG, 0.52 for CG). This difference indicates that real-time feedback may enhance task completion rates, which correlates with improved pronunciation accuracy and articulation.

Participants in the EG made an average of 15 pronunciation adjustment attempts per session, compared to 9 in the CG. The correlation between pronunciation adjustments and PAS improvement was 0.73 in the EG and 0.49 in the CG, while the correlation with AMS improvement was 0.77 in the EG and 0.53 in the CG. This suggests that the ISFS encouraged participants to make more targeted adjustments, facilitating more accurate pronunciation. Engagement consistency, which reflects the regularity of participation across sessions, was 88.2% in the EG and 75.3% in the CG. The correlation with PAS improvement was 0.67 in the EG and 0.41 in the CG, while for AMS improvement, it was 0.69 for the EG and 0.42 for the CG. Higher consistency among the EG highlights the motivational role of feedback in sustaining engagement. The EG performed an average of 8 learning retention checks per session, double that

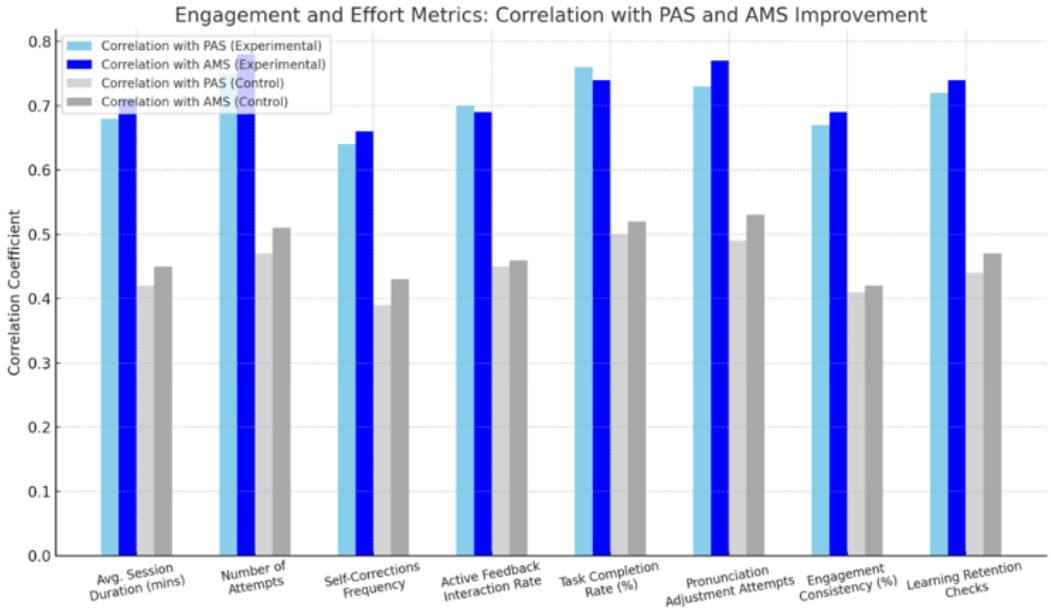
of the CG, which averaged 4. The correlation between retention checks and PAS improvement was 0.72 for the EG, while the CG correlated 0.44. For AMS improvement, the correlations were 0.74 in the EG and 0.47 in the CG. This finding suggests that the feedback system effectively supported retention, helping participants to apply learned corrections in subsequent sessions.

Table 4. Engagement and Effort Metrics results.

Metric	Mean (EG)	Mean (CG)	Correlation with PAS Improvement (EGl)	Correlation with AMS Improvement (EGl)	Correlation with PAS Improvement (CG)	Correlation with AMS Improvement (CG)
Average Session Duration (min)	45.3	38.6	0.68	0.71	0.42	0.45
Number of Attempts	28	20	0.75	0.78	0.47	0.51
Self-Corrections Frequency	12	8	0.64	0.66	0.39	0.43
Active Feedback Interaction Rate	16	10	0.70	0.69	0.45	0.46
Task Completion Rate (%)	93.4	85.6	0.76	0.74	0.50	0.52
Pronunciation Adjustment Attempts	15	9	0.73	0.77	0.49	0.53
Engagement Consistency (%)	88.2	75.3	0.67	0.69	0.41	0.42
Learning Retention Checks	8	4	0.72	0.74	0.44	0.47



(a)



(b)

Figure 3. Engagement and Effort Metrics results, **(a)** mean comparison; **(b)** correlation analysis.

3.4. Results for Between-Group comparison

The Between-Group comparison analysis (**Table 5** and **Figure 4**) highlights the differences in Pronunciation Accuracy Score (PAS), Articulatory Movement Score (AMS), Pronunciation Error Rate (PER) reduction, and key engagement metrics between the EG, which received real-time feedback, and the CG. The EG achieved a 31.2% improvement in PAS compared to only a 12.2% improvement in the CG, resulting in a 19.0% difference. This substantial disparity indicates that real-time feedback significantly boosted pronunciation accuracy, likely by guiding participants toward more precise articulation.

The EG improved their AMS by 37.1%, nearly double the 18.9% improvement observed in the CG, marking an 18.2% difference. This indicates that participants with access to MA feedback were better able to refine their articulatory movements, contributing to more accurate pronunciation. The EG's PER reduction was 57.1%, significantly higher than the 27.1% reduction in the CG, with a 30.0% difference between groups. The pronounced reduction in errors among the EG suggests that real-time feedback effectively supported participants in identifying and correcting their pronunciation mistakes, leading to fewer errors over time.

The EG's average session duration was 45.3 min, while the CG averaged 38.6 min, resulting in a 6.7 min difference. The additional practice time in the EG contributed to the group's higher improvement rates, as feedback encouraged them to refine their pronunciation more. Participants in the EG averaged 28 attempts per session, compared to 20 in the CG, with an 8-attempt difference. The higher number of attempts reflects the EG's increased motivation to practice pronunciation, likely driven by immediate feedback.

The EG self-corrected 12 times on average, compared to 8 times in the CG, a difference of 4 self-corrections. Real-time feedback likely prompted more self-corrections as participants received immediate cues to improve specific articulatory

points, facilitating quicker adjustments. The EG achieved a 93.4% task completion rate, compared to 85.6% in the CG, with a 7.8% difference. This higher completion rate suggests that feedback increased task engagement and encouraged participants to complete more exercises, leading to more significant overall improvement. Engagement consistency, which measures steady participation across sessions, was 88.2% for the EG and 75.3% for the CG, showing a 12.9% difference. The consistency in engagement in the EG reflects the motivating effect of interactive feedback, which may have contributed to sustained improved pronunciation.

Table 5. Between-Group comparison analysis.

Metric	EG	CG	Difference
PAS Improvement (%)	31.2	12.2	19.0
AMS Improvement (%)	37.1	18.9	18.2
PER Reduction (%)	57.1	27.1	30.0
Average Session Duration (min)	45.3	38.6	6.7
Number of Attempts	28	20	8
Self-Corrections Frequency	12	8	4
Task Completion Rate (%)	93.4	85.6	7.8
Engagement Consistency (%)	88.2	75.3	12.9

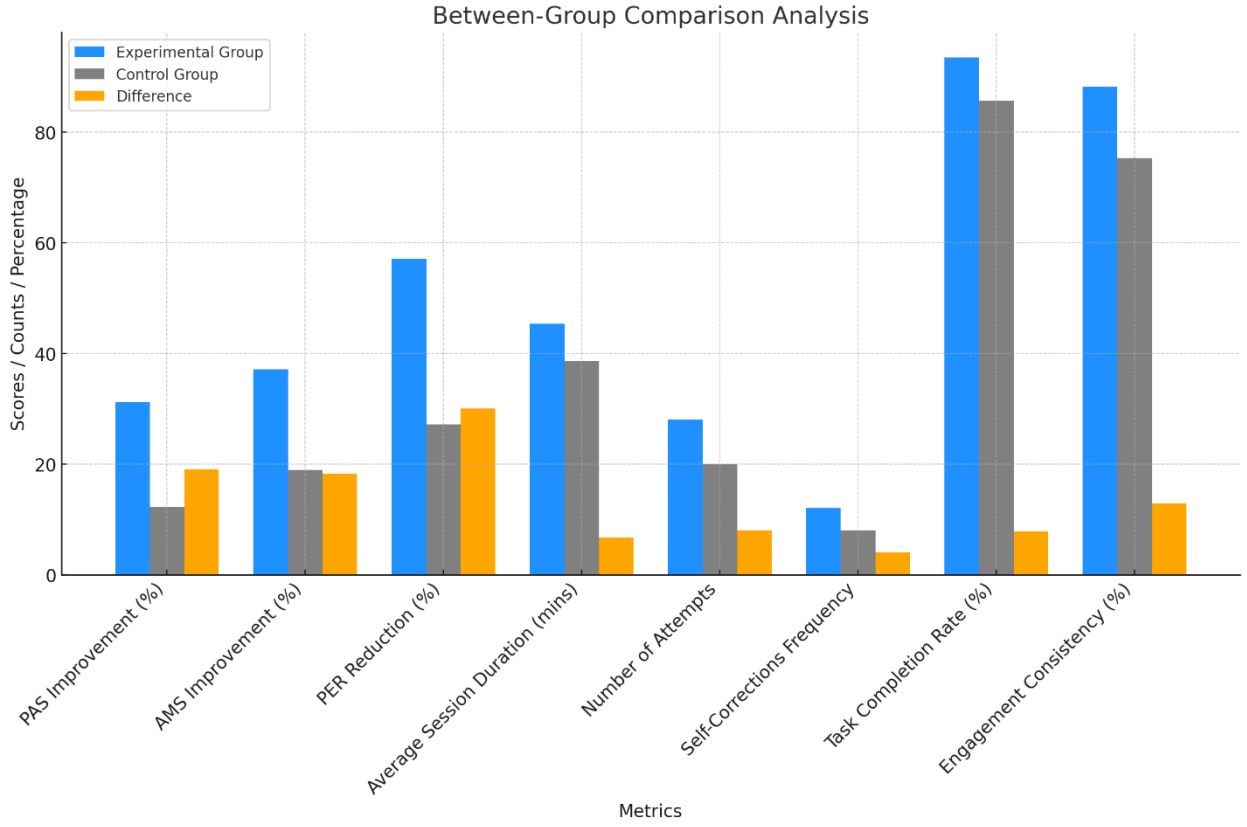


Figure 4. Comparison analysis between groups.

The Pre- and Post-Test comparison analysis (**Table 6** and **Figure 5**) evaluates the differences in pronunciation accuracy, articulation precision, error reduction, and

engagement metrics for both the EG and CG from the beginning to the end of the study. The results highlight significant improvements in the EG, which are driven by integrating real-time feedback and MA. The EG's PAS improved from a mean score of 62.3 to 81.7, representing a 19.4-point increase. In contrast, the CG's PAS increased from 63.1 to 70.8, showing a 7.7-point improvement. This substantial difference demonstrates that real-time feedback provided in the EG led to more pronounced gains in pronunciation accuracy.

Table 6. Pre- and Post-Test comparison analysis.

Metric	Pre-Test (EG)	Post-Test (EG)	Improvement (EG)	Pre-Test (CG)	Post-Test (CG)	Improvement (CG)
PAS (Mean Score)	62.3	81.7	19.4	63.1	70.8	7.7
AMS (Mean Score)	55.8	76.5	20.7	54.6	64.9	10.3
PER (Error Rate %)	24.7	10.6	−14.1	25.1	18.3	−6.8
Average Session Duration (min)	39.2	45.3	6.1	38.1	38.6	0.5
Number of Attempts	20	28	8	18	20	2
Self-Corrections Frequency	7	12	5	6	8	2
Task Completion Rate (%)	80.4	93.4	13.0	78.9	85.6	6.7
Engagement Consistency (%)	72.5	88.2	15.7	70.1	75.3	5.2

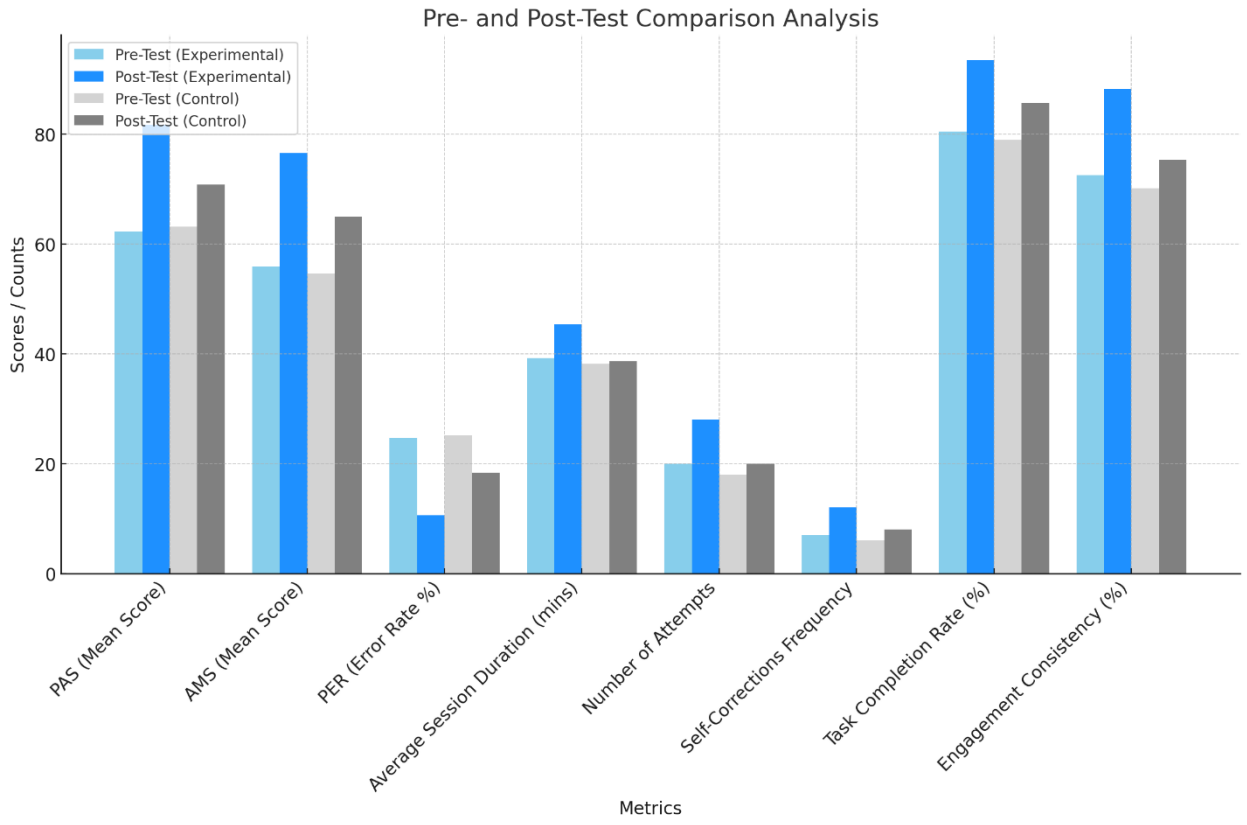


Figure 5. Pre- and post-test comparison analysis.

The AMS for the EG increased from 55.8 to 76.5, a 20.7-point improvement, whereas the CG improved from 54.6 to 64.9, a 10.3-point increase. This indicates that the EG's access to MA feedback significantly refined their articulation, allowing for better accuracy in producing English phonemes. The EG reduced its PER from 24.7% to 10.6%, a reduction of 14.1 percentage points. The CG's PER decreased from 25.1% to 18.3%, a reduction of 6.8 percentage points. The more significant reduction in error rate in the EG suggests that immediate feedback helped participants identify and correct mispronunciations more effectively than the CG.

The EG's session duration increased from 39.2 to 45.3 min, a 6.1 min increase. The CG showed minimal change, increasing from 38.1 to 38.6 min. The longer session duration in the EG indicates that real-time feedback likely motivated participants to spend more time practicing and refining their pronunciation. The number of pronunciation attempts rose from 20 to 28 in the EG, an increase of 8, compared to a minor increase of 2 attempts (from 18 to 20) in the CG. This difference shows that participants in the EG were more engaged in practicing and attempting corrections due to the continuous feedback.

Self-corrections increased from 7 to 12 in the EG, a 5-correction increase, while the CG increased from 6 to 8, a 2-correction increase. The higher frequency of self-corrections in the EG reflects how feedback encouraged active self-monitoring and adjustments, helping participants achieve more precise pronunciation. The EG's task completion rate improved from 80.4% to 93.4%, a 13.0% increase, while the CG's completion rate increased from 78.9% to 85.6%, a 6.7% improvement. The higher completion rate in the EG suggests that feedback maintained participants' motivation and commitment to completing the tasks. Engagement consistency, which measures regular participation and adherence to the study schedule, increased from 72.5% to 88.2% in the EG, a 15.7% improvement. The CG showed a more minor increase from 70.1% to 75.3%, a 5.2% improvement. The consistent engagement observed in the EG suggests that interactive feedback helped maintain participants' dedication to the study.

4. Conclusion and future work

Integrating MA and ISFS provides a promising approach to enhancing EP learning for non-native speakers, particularly those with phonological and articulatory challenges. This study's results demonstrate that real-time, adaptive feedback can significantly improve pronunciation accuracy, reduce error rates, and refine articulation precision among Chinese learners of English. The EG, which received feedback-driven practice, showed markedly higher improvements in PAS, AMS, and PER metrics than the CG. Engagement metrics, such as session duration and task completion rates, were also positively influenced, highlighting the role of feedback in fostering motivation and consistent practice. These findings underscore the potential of combining MCS and intelligent feedback to address specific pronunciation issues common among non-native speakers, such as vowel articulation and stress patterns. The insights gained from this study contribute to a deeper understanding of how technology can support personalized, practical pronunciation training.

Future research could expand upon these findings by exploring long-term retention of improvements, applying these systems across broader linguistic backgrounds, and refining feedback algorithms to address increasingly nuanced pronunciation challenges. By embracing advanced technologies, pronunciation training can become a more precise, engaging, and efficient process, ultimately supporting non-native learners in achieving fluency and confidence in English.

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Kissová O. Contrastive analysis in teaching English pronunciation. SWS Journal of Social Sciences and Art. 2020; 2(1), 39–65.
2. Dao DNA. Critical success factors in learning English pronunciation: A look through the lens of the learner (Doctoral dissertation, University of Nottingham). 2021.
3. AbdAlgane M, Idris SAM. Challenges of pronunciation to EFL learners in spoken English. Multicultural Education. 2020; 6(5).
4. Duyen TMT. Exploring Phonetic Differences and Cross-Linguistic Influences: A Comparative Study of English and Mandarin Chinese Pronunciation Patterns. Open Journal of Applied Sciences. 2024; 14(7), 1807–1822.
5. Lavitskaya Y, Zagorodniuk A. Acquisition of English onset consonant clusters by L1 Chinese speakers. English Pronunciation Instruction: Research-based insights. 2021; 11, 255–278.
6. Mehta S. Effects of Visual Feedback on the Production and Perception of Second Language Speech Sounds: A Comparison of Articulatory and Auditory Instruction. The University of Texas at Dallas. 2020.
7. O'Connell S. Investigating a speech and language therapy-informed approach to pronunciation teaching in the English language teaching classroom (Doctoral dissertation, University of Limerick). 2023.
8. Sconiers MG. Guided Articulation Treatment with Supplemental Visual Feedback (Doctoral dissertation, California State University San Marcos). 2021.
9. Bogach N, Boitsova E, Chernonog S, et al. Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching. Electronics. 2021; 10(3), 235.
10. Rogerson-Revell PM. Computer-assisted pronunciation training (CAPT): Current issues and future directions. Relc Journal. 2021; 52(1), 189–205.
11. Liu Y, Quan Q. AI recognition method of pronunciation errors in oral English speech with the help of big data for personalized learning. Journal of Information & Knowledge Management. 2022; 21(Supp02), 2240028.
12. Kholis A. Elsa speak app: automatic speech recognition (ASR) for supplementing English pronunciation skills. Pedagogy: Journal of English Language Teaching. 2021; 9(1), 1–14.
13. Nemani P, Krishna GS, Supriya K, Kumar S. Speaker independent VSR: A systematic review and futuristic applications. Image and Vision Computing. 2023; 104787.
14. Wang Y, Huang H. Audio–visual deepfake detection using articulatory representation learning. Computer Vision and Image Understanding. 2024; 248, 104133.
15. Indumathi N, Savaram P, Sengan S, et al. Impact of Fireworks Industry Safety Measures and Prevention Management System on Human Error Mitigation Using a Machine Learning Approach, Sensors, 2023, 23 (9), 4365; DOI:10.3390/s23094365.
16. Parkavi K, Satheesh N, Sudha D, et al. Effective Scheduling of Multi-Load Automated Guided Vehicle in Spinning Mill: A Case Study, IEEE Access, 2023, DOI:10.1109/ACCESS.2023.3236843.
17. Ran Q et al., English language teaching based on big data analytics in augmentative and alternative communication system, Springer-International Journal of Speech Technology, 2022, DOI:10.1007/s10772-022-09960-1.

18. Ngangbam PS, Suman S, Ramachandran TP, et al. Investigation on characteristics of Monte Carlo model of single electron transistor using Orthodox Theory, Elsevier, Sustainable Energy Technologies and Assessments, Vol. 48, 2021, 101601, doi: 10.1016/j.seta.2021.101601
19. Huang H, Wang X, Sengan S, et al. Emotional intelligence for board capital on technological innovation performance of high-tech enterprises, Elsevier, Aggression and Violent Behavior, 2021, 101633, doi: 10.1016/j.avb.2021.101633.
20. Sudhakar S, Kumar K, Subramaniyaswamy V, et al., Cost-effective and efficient 3D human model creation and re-identification application for human digital twins, Multimedia Tools and Applications, 2021. DOI:10.1007/s11042-021-10842-y.
21. Prabhakaran N, Sengan S, Marimuthu BP, et al. Novel Collision Detection and Avoidance System for Mid-vehicle Using Offset-Based Curvilinear Motion. Wireless Personal Communication, 2021. DOI:10.1007/s11277-021-08333-2.
22. Balajee A, Rajagopal V, Sengan S, et al., Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest, J Ambient Intell Human Comput, 2021, DOI:10.1007/s12652-020-02869-0.
23. Omnia SN, Setiawan R, Jayanthi P, et al. An educational tool for enhanced mobile e-Learning for technical higher education using mobile devices for augmented reality, Microprocessors and Microsystems, 83, 2021, 104030, doi: 10.1016/j.micpro.2021.104030
24. Firas TA, Ayasrah FT, Alsharafa NS, et al. Strategizing Low-Carbon Urban Planning through Environmental Impact Assessment by Artificial Intelligence-Driven Carbon Foot Print Forecasting, Journal of Machine and Computing, 4(4), 2024, doi: 10.53759/7669/jmc202404105.
25. Shaymaa HN, Sadu VB, Sengan S, et al. Genetic Algorithms for Optimized Selection of Biodegradable Polymers in Sustainable Manufacturing Processes, Journal of Machine and Computing, 4(3), 563–574, https://doi.org/10.53759/7669/jmc202404054.
26. Hayder MAG, Sengan S, Sadu VB et al. An open-source MP + CNN + BiLSTM model-based hybrid model for recognizing sign language on smartphones. Int J Syst Assur Eng Manag. 2024. https://doi.org/10.1007/s13198-024-02376-x
27. Bhavana Raj K, Webber JL, Marimuthu D, et al. Equipment Planning for an Automated Production Line Using a Cloud System, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems, 565, 707–717, Springer, Singapore. DOI:10.1007/978-981-19-7455-7_57.
28. Zhai X, Haudek KC, Shi L, et al. From substitution to redefinition: A framework of machine learning‐based science assessment. Journal of Research in Science Teaching. 2020; 57(9), 1430–1459.
29. Lee HS, Gweon GH, Lord T, et al. Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. Journal of Science Education and Technology. 2021; 30(2), 168–192.
30. Ibragimova S. Creation of An Intelligent System for Uzbek Language Teaching Using Phoneme-Based Speech Recognition. Revue d'Intelligence Artificielle. 2023; 37(6).
31. Liakina N, Liakin D. Speech technologies and pronunciation training: What is the potential for efficient corrective feedback. Second Language Pronunciation: Different Approaches to Teaching and Training, 2023; 287–312.
32. Sun W. The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. Frontiers in Psychology, 2023; 14, 1210187.
33. Zhu J, Zhang X, Li J. Using AR filters in L2 pronunciation training: Practice, perfection, and willingness to share. Computer Assisted Language Learning, 2024; 37(5-6), 1364–1396.
34. Lan EM. A comparative study of computer and mobile-assisted pronunciation training: The case of university students in Taiwan. Education and Information Technologies, 2022; 27(2), 1559–1583