

AI-Integrated Computational Framework for Pronunciation Error Analysis Through Native–Learner Speech Comparison

Hastha Sunardi¹, Rendra Gustriansyah², Fery Antony³

^{1,3}Computer Systems, Universitas Indo Global Mandiri, Palembang, Indonesia

²Informatics Department, Universitas Indo Global Mandiri, Palembang, Indonesia

Received:

May 2, 2026

Revised:

July 21, 2026

Accepted:

September 2, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Hastha Sunardi

Email*:

hastha_s@uigm.ac.id

DOI:

10.63158/journalisi.v8i5.1812

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. This study proposes an AI-integrated computational framework for pronunciation error analysis through native–learner speech comparison. The framework integrates Whisper-based Automatic Speech Recognition (ASR), acoustic feature extraction, statistical analysis, and machine learning to analyze learner pronunciation. Unlike previous studies that primarily rely on individual transcription-based or acoustic measures, the proposed framework integrates multiple pronunciation indicators into a composite Pronunciation Accuracy Score (PAS) formulated in this study. The framework was evaluated using a dataset comprising 45 non-native English learners and one native English speaker. PAS combines Pronunciation Error Rate (PER), Word Error Rate (WER), Pitch Similarity, Rhythm Similarity, and Mel-Frequency Cepstral Coefficient (MFCC) Similarity. Pearson correlation analysis showed that Rhythm Similarity ($r = 0.544$) and Pitch Similarity ($r = 0.510$) exhibited the strongest positive correlations with PAS, whereas Word Error Rate showed the strongest negative correlation ($r = -0.495$). K-Means clustering identified three learner groups with distinct pronunciation characteristics within the tested dataset. Although evaluated on a relatively small dataset without validation against expert human ratings, the proposed framework provides a promising foundation for Computer-Assisted Pronunciation Training (CAPT) and future multilingual pronunciation assessment.

Keywords: Automatic Pronunciation Assessment, Pronunciation Accuracy Score, Whisper ASR, Native–Learner Speech Comparison, K-Means Clustering.

1. INTRODUCTION

Recent advances in Artificial Intelligence (AI), particularly in Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Machine Learning (ML), have significantly transformed speech processing and analytics across various domains, including intelligent information systems, human-computer interaction, educational technology, and data-driven decision support systems. These technologies enable modern computational systems to convert speech signals into structured digital representations that can be automatically analyzed to extract linguistic and acoustic information. Consequently, AI-based speech analytics has become an important research area for developing intelligent systems that support automated language evaluation and data-driven decision-making processes [1], [2]

Recent progress in ASR technology has been largely driven by self-supervised and weakly supervised deep learning approaches. State-of-the-art models such as wav2vec 2.0, HuBERT, and Whisper have demonstrated remarkable performance in learning robust speech representations and generating highly accurate transcriptions across diverse acoustic environments and speaker characteristics [3], [4], [5]. Among these models, Whisper has attracted considerable attention because of its strong multilingual transcription capability and robustness across diverse recording conditions, making it suitable for speech analytics and pronunciation assessment tasks. Beyond improving speech recognition accuracy, modern ASR models provide opportunities for advanced speech analytics, including the automatic identification of pronunciation patterns, phonetic deviations, and speech production characteristics. In intelligent educational environments, the integration of ASR and machine learning enables the development of automated pronunciation assessment systems capable of providing objective, scalable, and data-driven evaluation.

One of the most important applications of these technologies is Automatic Pronunciation Assessment (APA) and Computer-Assisted Pronunciation Training (CAPT), which aim to support pronunciation evaluation and learning through computational approaches. Previous studies have shown that ASR-based pronunciation training can improve second-language pronunciation performance while providing faster, more consistent, and scalable feedback than conventional human assessment [1], [6], [7]. A comprehensive

review by Kheir et al. [8] highlighted the rapid evolution of APA through deep learning architectures, transformer-based models, and modern speech representations. Likewise, Gong et al. [9] and Yan et al. [10] demonstrated that transformer-based approaches can effectively capture multiple dimensions of non-native speech and support automated pronunciation assessment.

Pronunciation assessment should also be distinguished from related constructs such as accentedness, intelligibility, and comprehensibility. These constructs are related but not identical. Pronunciation accuracy concerns the degree to which speech production conforms to a target pronunciation reference, whereas accentedness reflects the perceived degree of foreign accent. Intelligibility concerns whether spoken language can be understood by a listener, while comprehensibility refers to the perceived ease or difficulty of understanding speech. Accordingly, the framework proposed in this study does not claim to directly measure communicative intelligibility, comprehensibility, or perceived accentedness. Instead, it focuses on computational pronunciation-related indicators obtained through comparison between learner speech and a designated native-speaker reference.

Despite these advances, previous studies have addressed different aspects of pronunciation assessment, including global pronunciation scoring, intelligibility prediction, speech recognition accuracy, and acoustic analysis, using a variety of computational representations [9], [10], [11], [12]. These approaches provide valuable evidence for pronunciation assessment, but they do not necessarily capture the same construct. Established pronunciation assessment approaches also employ more fine-grained indicators, such as Goodness of Pronunciation (GOP), ASR confidence measures, and phone-level posterior probabilities, while human-rated intelligibility provides a listener-based perspective on speech performance. In particular, transcription accuracy, acoustic similarity, accentedness, intelligibility, and comprehensibility should not be treated as interchangeable measures of pronunciation quality. Furthermore, pronunciation assessment in educational practice continues to rely substantially on expert human raters, whose evaluations may be affected by subjectivity, inter-rater variability, and limited scalability [13], [14], [15]. These considerations motivate the development of computational approaches that can provide reproducible pronunciation-

related indicators while clearly distinguishing similarity-based analysis from validated human-rated pronunciation assessment.

Meanwhile, many of the technological components required for automated pronunciation assessment have matured independently. Modern ASR systems provide highly accurate automatic transcriptions [3], [4], [5], transformer-based architectures have demonstrated strong performance in pronunciation assessment [9], [10], [12], and acoustic representations such as Mel-Frequency Cepstral Coefficients (MFCCs) remain widely used for modeling speech characteristics [16]. Previous studies have further shown that pronunciation quality can be characterized through transcription errors, speech recognition performance, acoustic similarity, and speech intelligibility indicators [11], [12], [14], [17]. However, these complementary sources of information are rarely integrated into a unified AI-based analytical framework capable of providing multidimensional pronunciation assessment through native–learner speech comparison.

Based on the literature reviewed, existing studies have primarily addressed individual aspects of automated pronunciation assessment, including ASR accuracy [3], [4], [5], pronunciation scoring [9], [10], [12], and speech intelligibility evaluation [11], [15]. To the best of our knowledge, limited research has systematically integrated transcription-based error metrics and acoustic similarity measures into a unified computational framework for direct native–learner speech comparison. Furthermore, the use of a native-speaker recording as a designated reference for consistent computational comparison of learner pronunciation has received limited attention in the studies reviewed. Consequently, a research gap remains in developing a comprehensive computational framework capable of performing multidimensional pronunciation-related analysis and learner profiling through native–learner speech comparison.

To address this gap, this study proposes an Integrated AI-Based Framework for Automated Pronunciation Error Analysis Through Native–Learner Speech Comparison. The proposed framework integrates Whisper-based ASR, acoustic feature extraction, statistical analysis, and machine learning techniques within a unified analytical pipeline for computational pronunciation error analysis. Within this framework, the native-speaker recording serves as a designated reference or digital anchor for comparing learner speech using transcription-based error measures and acoustic similarity

indicators. The reference is used to establish a consistent comparison point rather than to represent a universal model of native-like pronunciation. Accordingly, greater similarity to the selected native-speaker reference is interpreted as higher similarity within the tested framework and should not be regarded as a direct measure of general native-like pronunciation, intelligibility, or communicative effectiveness.

The novelty of this study lies not merely in combining existing AI techniques, but in integrating transcription-level error metrics, namely Pronunciation Error Rate (PER) and Word Error Rate (WER), with acoustic similarity measures, including Pitch Similarity, Rhythm Similarity, and MFCC Similarity, into a unified Pronunciation Accuracy Score (PAS) formulated in this study. Furthermore, the proposed framework combines statistical analysis and machine learning-based learner profiling through K-Means clustering within a single analytical pipeline. This integration provides a multidimensional computational representation of learner speech by jointly considering transcription-related errors, acoustic similarity, statistical relationships, and learner profiles through native–learner speech comparison. The resulting PAS is intended as an author-formulated computational indicator and should not be interpreted as an externally validated or human-rated pronunciation proficiency score.

Accordingly, the objectives of this study are: (1) to design an integrated computational framework that combines Whisper-based ASR, acoustic feature extraction, statistical analysis, and machine learning for automated pronunciation error analysis; (2) to develop a multidimensional computational representation of learner pronunciation by integrating transcription-based error metrics and acoustic similarity indicators for native–learner speech comparison; and (3) to evaluate the proposed framework through descriptive statistics, correlation analysis, and K-Means-based learner profiling within the tested dataset.

Based on these objectives, the study addresses the following research questions:

- 1) RQ1: How can Whisper-based ASR, transcription-based error measures, and acoustic similarity indicators be integrated into a computational framework for native–learner speech comparison?
- 2) RQ2: What relationships exist between the transcription-based and acoustic pronunciation indicators and the composite PAS within the tested dataset?

- 3) RQ3: Can K-Means clustering identify distinct learner pronunciation profiles based on the integrated pronunciation indicators within the tested dataset?

2. METHODS

2.1. Research Framework

This study proposes an integrated computational framework for pronunciation error analysis through direct native–learner speech comparison. The framework combines Whisper-based ASR, pronunciation and acoustic feature extraction, statistical analysis, and machine learning techniques to provide a reproducible and data-driven computational analysis of learner pronunciation while generating learner performance profiles within the tested dataset.

Unlike conventional CAPT systems that primarily emphasize a single assessment dimension, the proposed framework integrates transcription-based error metrics and acoustic similarity measures within a unified analytical architecture. As illustrated in Figure 1, the framework consists of six sequential stages: (1) audio preprocessing, (2) speech transcription using Whisper-based ASR, (3) pronunciation and acoustic feature extraction, (4) PAS computation, (5) statistical analysis, and (6) machine learning-based learner profiling. These stages form an end-to-end analytical pipeline that progressively transforms raw speech recordings into multidimensional pronunciation indicators for computational analysis and learner profiling.

During the feature extraction stage, transcription-based pronunciation-related measures, including PER and WER, are combined with acoustic similarity measures comprising Pitch Similarity, Rhythm Similarity, and MFCC Similarity. These complementary indicators are subsequently integrated into the proposed PAS, which serves as the primary composite quantitative indicator used in the computational analysis of learner pronunciation.

Finally, statistical analysis and machine learning techniques are employed to interpret the extracted pronunciation indicators. Descriptive statistics and Pearson correlation analysis are used to characterize the distribution and associations among the pronunciation indicators and the constructed PAS, while K-Means clustering is applied to identify exploratory learner profiles based on the standardized pronunciation feature

space. Consequently, the proposed framework produces multiple analytical outputs, including PAS, learner ranking, correlation analysis, learner clusters, and learner profiles, thereby providing a comprehensive computational approach for pronunciation-related analysis and data-driven learner profiling within the tested dataset.

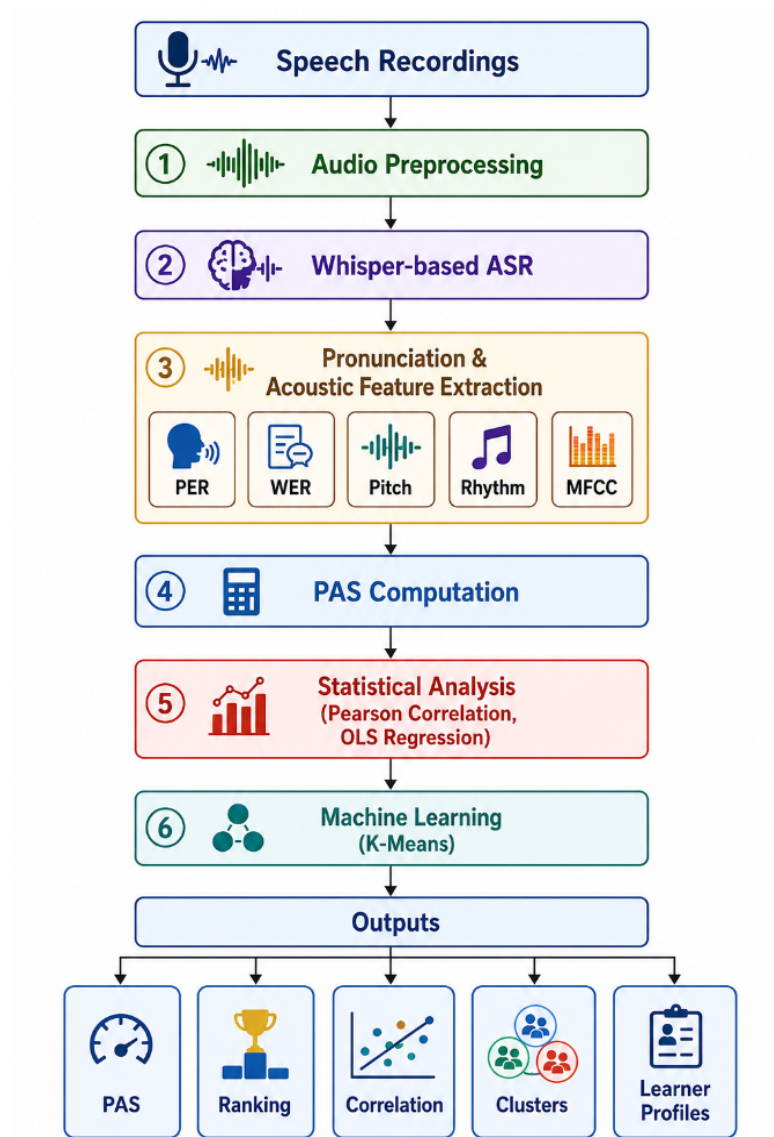


Figure 1. Proposed Integrated Computational Framework for Automated Pronunciation Error Analysis Through Native-Learner Speech Comparison.

2.2. Experimental Speech Dataset

The proposed computational framework was evaluated using a speech dataset consisting of one native English speaker and forty-five non-native English learners. The native-

speaker recording serves as the pronunciation reference (digital anchor), whereas the learner recordings constitute the evaluation samples used throughout the analytical process. The native-speaker recording is used as a designated reference for within-study comparison rather than as a universal representation of native-like pronunciation. This reference-based dataset enables direct comparison between native and learner speech under a consistent experimental design.

All participants were instructed to read the same phonetically balanced English passage containing approximately 156 words. Using an identical reading passage provides comparable linguistic conditions for the native and learner recordings and enables direct native–learner speech comparison at both the transcription and acoustic levels. Previous studies have demonstrated the usefulness of reference-based approaches for pronunciation analysis and speech intelligibility analysis [9], [10], [17].

Speech transcriptions were automatically generated using Whisper-based ASR as part of the proposed computational framework. In addition, an Indonesian translation of the reading passage was prepared to facilitate linguistic interpretation and qualitative pronunciation feedback. Consequently, the speech dataset supports quantitative pronunciation analysis based on transcription error metrics and acoustic similarity measures, as well as qualitative interpretation of learner pronunciation characteristics.

2.3. Audio Preprocessing

Audio preprocessing constitutes the first stage of the proposed computational framework illustrated in Figure 1. The objective of this stage is to standardize native and learner speech recordings before automatic transcription and acoustic feature extraction. By reducing recording variability while preserving pronunciation-related acoustic information, audio preprocessing establishes consistent input signals for subsequent speech analysis and machine learning processes [18].

Each speech recording undergoes four preprocessing operations. First, all audio recordings are resampled to 16 kHz, corresponding to the input sampling rate used by the Whisper-based ASR model in this study. Second, amplitude normalization is applied to reduce variations in recording intensity according to Equation (1).

$$x_{\text{norm}}[n] = \frac{x[n]}{\max(|x[n]|)}$$
 (1)

where denotes the amplitude of the original speech signal at sample, denotes the corresponding normalized amplitude, and represents the maximum absolute amplitude of the recording. This operation scales each recording to a consistent amplitude range while preserving its waveform characteristics [18].

Next, voice activity detection (VAD)-based silence trimming is performed to remove leading and trailing non-speech segments that do not contribute to pronunciation analysis [19], [20], [21]. Subsequently, background noise reduction is applied to suppress environmental noise while preserving speech components relevant to pronunciation analysis and subsequent automatic speech recognition [22]. Collectively, these preprocessing operations improve signal quality, reduce unwanted variability, and provide more consistent input for speech recognition and acoustic feature extraction [18], [20], [21], [22].

The standardized speech recordings produced during this stage are subsequently processed by the Whisper-based ASR module for transcription generation and by the acoustic analysis module for pronunciation feature extraction. Consequently, audio preprocessing provides a consistent and reproducible input stage for transcription-based error analysis, acoustic similarity measurement, statistical analysis, and machine learning-based learner profiling.

2.4. Whisper-based Automatic Speech Recognition

Following audio preprocessing, the standardized speech recordings are transcribed using the Whisper-based ASR model. In this study, the Whisper base model is used for automatic speech transcription. Whisper is a transformer-based end-to-end speech recognition model trained on large-scale multilingual speech data, enabling robust transcription across diverse recording conditions and speaker characteristics [23]. Recent studies have also demonstrated the effectiveness of Whisper-based architectures and their variants in improving multilingual speech recognition performance and model adaptability [24], [25].

In this study, the same phonetically balanced English reading passage is recorded by one native English speaker and multiple non-native English learners. Each audio recording is independently transcribed by the Whisper ASR model to produce word-level transcription outputs without manual intervention. The use of a common reading text enables direct comparison between native and learner transcriptions while minimizing linguistic variability unrelated to pronunciation.

The native-speaker transcription serves as the reference transcript, whereas each learner transcription is aligned with the native transcript using sequence matching techniques. The alignment process identifies insertion, deletion, and substitution errors occurring during automatic speech recognition. These transcription differences are subsequently used to compute transcription-based and pronunciation-related indicators, including WER and PER, which provide computational measures of differences between the native and learner speech transcripts [8], [11], [18].

By employing Whisper-based ASR as the transcription engine, the proposed framework provides an automated and reproducible mechanism for generating transcription-based pronunciation indicators without manual transcription. The resulting transcription outputs constitute the primary input for subsequent pronunciation feature extraction, statistical analysis, and machine learning-based learner profiling.

2.5. Pronunciation Feature Extraction

Following automatic transcription, multiple pronunciation-related indicators are extracted to characterize pronunciation performance from complementary linguistic and acoustic perspectives. Rather than relying on a single evaluation metric, the proposed framework integrates transcription-based error measures and acoustic similarity measures to provide a more comprehensive representation of pronunciation characteristics. Previous studies have demonstrated that combining multiple pronunciation indicators improves the robustness of automatic pronunciation analysis by capturing different aspects of speech production [8], [11], [16], [18].

2.5.1. Transcription-based Error Measures

Two transcription-based error measures are used to characterize learner speech by comparing the native-speaker reference transcript with each learner transcript

generated by the Whisper ASR model. These measures provide complementary information on transcription-related differences observed between the reference and learner speech.

The first measure is the Pronunciation Error Rate (PER), which is used in this study as an ASR-derived approximation of pronunciation-related error. In the implemented computational pipeline, PER is obtained from the alignment result generated during transcript comparison and uses the alignment-based error rate returned by the text analysis module. Specifically, the implemented alignment procedure calculates this error rate from substitutions, deletions, and insertions relative to the reference transcript length. Therefore, PER in this study should not be interpreted as an independently computed phoneme-level error rate or as a directly validated pronunciation measure. This operational definition is important because the proposed Pronunciation Accuracy Score (PAS) is a composite score formulated by the authors and has not yet been validated against expert human pronunciation ratings.

The second measure is the Word Error Rate (WER), which quantifies word-level transcription errors resulting from substitutions, deletions, and insertions. WER is computed from the alignment between the reference and learner transcripts as:

$$WER = \frac{S + D + I}{N} \quad (2)$$

where S , D , and I denote the numbers of substitutions, deletions, and insertions, respectively, and represents the total number of words in the reference transcript [18].

2.5.2. Acoustic Similarity Measures

To complement transcription-based evaluation, three acoustic similarity measures are extracted directly from the speech signals. The first measure is Pitch Similarity, which evaluates the similarity of fundamental frequency contours between the learner and the native speaker. Pitch contours are extracted using the `librosa.piptrack` function, and the maximum pitch value from each frame is used to form the corresponding pitch sequence. The two sequences are truncated to the same length, after which Pearson correlation is calculated. The resulting correlation coefficient, which ranges from -1 to 1 , is

transformed into a 0–1 similarity scale using $similarity = (c+1)/2$. Higher values therefore indicate greater similarity in the extracted pitch patterns between the learner and native-speaker reference.

The second measure is Rhythm Similarity, which measures temporal correspondence between learner and native speech using onset-strength representations. The onset-strength sequences are extracted using the `librosa.onset.onset_strength` function and truncated to the same length before Pearson correlation is calculated. The resulting correlation coefficient is transformed into a normalized 0–1 similarity value using $similarity = (c+1)/2$, where higher values indicate greater similarity in the temporal patterns represented by the onset-strength sequences.

The third measure is MFCC Similarity, which evaluates spectral similarity between native and learner speech. Thirteen MFCC coefficients are extracted from each recording using the `librosa.feature.mfcc` function and averaged across time to obtain a fixed-length MFCC representation. Pearson correlation is then calculated between the native and learner MFCC representations. The resulting correlation coefficient is transformed into a normalized 0–1 similarity value using $similarity = (c+1)/2$, where higher values indicate greater similarity in the extracted spectral characteristics. MFCCs provide a compact representation of speech spectral characteristics and have become one of the most widely adopted acoustic representations in automatic speech processing and pronunciation analysis [16], [18]. Collectively, the extracted transcription-based and acoustic-based indicators provide complementary information describing learner pronunciation characteristics from lexical, temporal, prosodic, and spectral perspectives. These indicators subsequently serve as the input variables for computing the proposed PAS and for subsequent statistical and machine learning analyses.

2.6. Pronunciation Accuracy Score (PAS)

To provide a unified computational representation of learner pronunciation characteristics, the proposed framework integrates the five extracted pronunciation indicators into a single composite metric referred to as the PAS. Rather than relying on an individual pronunciation indicator, PAS combines transcription-based error measures and acoustic similarity measures to capture complementary aspects of the analyzed speech characteristics. The use of composite pronunciation indicators has been

recognized as an effective approach for improving the interpretability and comprehensiveness of automatic pronunciation analysis [8], [26].

Since lower values of PER and WER indicate higher transcription-related accuracy, these error measures are first converted into complementary accuracy terms by subtracting their values from one. The resulting accuracy terms are then combined with Pitch Similarity, Rhythm Similarity, and MFCC Similarity, all of which are represented on a normalized 0–1 scale. The proposed PAS is therefore computed as:

$$PAS = \frac{(1 - PER) + (1 - WER) + Pitch + Rhythm + MFCC}{5} \quad (3)$$

Where *PER* and *WER* denote the implemented transcription-based error indicators, while *Pitch*, *Rhythm* and *MFCC*, denote the normalized Pitch Similarity, Rhythm Similarity, and MFCC Similarity values, respectively. The denominator 5 represents the five equally weighted components included in the composite score. All five indicators contribute equally to the composite score. Consequently, PAS ranges between 0 and 1, with higher values indicating a higher composite similarity-oriented score relative to the native-speaker reference within the proposed computational framework.

The equal-weight integration adopted in this study was selected to provide an interpretable baseline for multidimensional pronunciation analysis without introducing task-specific weighting assumptions. Because the objective of this study is to establish a computational framework rather than to optimize a prediction model, equal weighting allows each pronunciation indicator to contribute transparently to the overall pronunciation score [18], [26]. This weighting scheme is treated as a baseline formulation and should not be interpreted as an empirically validated or psychometrically optimized weighting scheme. The computed PAS values are subsequently used as the principal composite indicator for descriptive statistical analysis, correlation analysis, and machine learning-based learner profiling presented in the subsequent sections.

2.7. Statistical Analysis

To quantitatively evaluate the extracted pronunciation indicators, descriptive statistics and Pearson correlation analysis were performed. These analyses were designed to characterize the distribution of the pronunciation indicators and examine their associations with the proposed PAS within the tested dataset. First, descriptive statistics, including the mean, standard deviation, minimum value, and maximum value, were calculated for each pronunciation indicator to summarize the overall distribution of learner pronunciation performance. These statistics provide an overview of the variability of the extracted pronunciation features across all learner recordings.

Second, Pearson correlation analysis was conducted to examine the linear associations between each pronunciation indicator and the proposed PAS. Pearson's correlation coefficient was selected because the pronunciation indicators were represented as continuous numerical variables. The resulting correlation coefficients were interpreted descriptively to characterize the strength and direction of the associations between the individual indicators and the constructed PAS. Two-tailed p -values and 95% confidence intervals were also considered to characterize the statistical uncertainty of the observed associations.

Importantly, no inferential regression model was applied to estimate the independent effects of the pronunciation indicators on PAS. Because PAS is mathematically constructed from PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity, regressing PAS on these same component variables would produce a tautological relationship and would not provide independent evidence of predictor importance. Therefore, the statistical analysis was limited to descriptive statistics and Pearson correlation analysis, while machine learning-based learner profiling was subsequently performed using the standardized pronunciation feature vector. The outputs obtained from the statistical analysis subsequently provide descriptive evidence for interpreting the machine learning-based learner profiling presented in the following subsection.

2.8. Machine Learning-based Learner Profiling

To further investigate pronunciation characteristics beyond individual performance scores, the proposed framework incorporates an unsupervised machine learning approach to identify groups of learners with similar pronunciation patterns. Unlike

conventional pronunciation analysis that evaluates learners individually, learner profiling enables the identification of common pronunciation characteristics shared among learners and facilitates the interpretation of pronunciation variability within the dataset [27]. Prior to clustering, the five pronunciation indicators, namely PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity, were standardized using Z-score normalization to eliminate differences in measurement scales. For each feature, the standardized value was calculated as:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (4)$$

where X_i represents the original value of a feature for learner i , μ denotes the mean of that feature across all learners, and σ represents its standard deviation. Standardization ensures that each feature contributes comparably during distance computation and prevents variables with larger numerical ranges from dominating the clustering process.

The optimal number of learner groups was determined using the Elbow Method, which evaluates the Within-Cluster Sum of Squares (WCSS) for different values of K . WCSS is calculated as:

$$WCSS = \sum_{k=1}^K \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \quad (5)$$

where K denotes the number of clusters, C_k represents cluster k , x_i is the standardized feature vector of learner i , and μ_k denotes the centroid of cluster k . The value of K corresponding to the elbow point was selected as the optimal number of learner clusters because it provides a suitable balance between cluster compactness and model simplicity. After determining the optimal number of clusters, the K-Means algorithm was applied to partition learners into groups exhibiting similar pronunciation characteristics.

Because the extracted pronunciation indicators form a five-dimensional feature space, Principal Component Analysis (PCA) was subsequently employed to project the standardized features into a two-dimensional space for visualization purposes. It should be noted that PCA was used solely for visual interpretation of learner clusters, whereas the K-Means algorithm was applied to partition learners into exploratory groups exhibiting similar pronunciation characteristics within the tested dataset.

The resulting learner profiles provide complementary information to the statistical analyses by revealing grouping patterns in the multidimensional pronunciation feature space that are not directly observable from individual pronunciation scores. Consequently, the machine learning-based learner profiling module extends the proposed computational framework beyond quantitative pronunciation measurement toward data-driven characterization of learner pronunciation profiles, which may support personalized pronunciation analysis and future computer-assisted pronunciation learning applications [27], [28], [29].

3. RESULTS AND DISCUSSION

3.1. Results

The results of the proposed AI-integrated computational framework are presented based on the pronunciation features extracted from the learner speech recordings. The analysis includes descriptive statistics of the extracted features, pronunciation assessment results, acoustic similarity analysis, and learner clustering based on the selected pronunciation indicators and pronunciation performance.

3.1.1. Experimental Overview

The experimental dataset consisted of 46 English speech recordings, including one native English speaker and 45 non-native English learners. All participants read the same phonetically balanced English passage under identical experimental conditions to ensure that pronunciation differences were primarily attributable to speaker characteristics rather than linguistic content. The native speaker recording served as the pronunciation reference throughout the analysis.

As described in Figure 1, each speech recording was processed through the proposed computational pipeline consisting of audio preprocessing, Whisper-based automatic speech recognition, pronunciation feature extraction, PAS computation, statistical analysis, and machine learning-based learner profiling. Five pronunciation indicators were extracted, namely PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity, which were subsequently integrated into PAS for quantitative evaluation.

The experimental evaluation was designed to address three objectives. First, descriptive statistics were used to summarize learner pronunciation characteristics. Second, correlation analysis was performed to investigate relationships between individual pronunciation indicators and PAS. Finally, K-Means clustering combined with PCA was employed to identify learner pronunciation profiles exhibiting similar pronunciation characteristics. Collectively, these analyses provide a comprehensive evaluation of the proposed computational framework from statistical and machine learning perspectives.

3.1.2. Descriptive Statistics of Pronunciation Indicators

Descriptive statistics were computed to summarize the distribution of the five pronunciation indicators extracted from the proposed computational framework, namely PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity. These indicators collectively describe learner pronunciation performance from both transcription-based and acoustic perspectives. An overview of these descriptive statistics is provided below, with the complete results summarized in Table 1.

Table 1. Descriptive statistics of pronunciation indicators (N = 45).

Indicator	Mean	Std. Dev.	Min	Max
PER	0.302	0.089	0.110	0.470
WER	0.308	0.106	0.130	0.490
Pitch	0.691	0.127	0.500	0.930
Rhythm	0.734	0.132	0.520	0.940
MFCC	0.751	0.131	0.500	0.940
PAS	0.713	0.055	0.580	0.840

The acoustic similarity indicators generally exhibited higher values than the transcription-based error measures. Pitch Similarity achieved a mean of 0.691 (SD = 0.127), Rhythm Similarity averaged 0.734 (SD = 0.132), and MFCC Similarity showed the highest mean value among the acoustic indicators at 0.751 (SD = 0.131). In contrast, PER and WER had mean error rates of 0.302 and 0.308, respectively, indicating that learners exhibited measurable transcription-level deviations. Nevertheless, the acoustic similarity values suggest that many learners retained prosodic and spectral characteristics that were broadly similar to those of the native-speaker reference.

Overall, the descriptive statistics indicate measurable variability across the pronunciation indicators, supporting their use for subsequent correlation analysis and exploratory machine learning-based learner profiling. For PAS specifically, the observed range of 0.580–0.840 indicates that the composite score captures differences among learners in the tested dataset, despite its relatively small standard deviation ($SD = 0.055$). However, this observed variation should not be interpreted as evidence of established measurement sensitivity. Because PAS is an author-formulated composite score that has not yet been validated against expert human ratings or external proficiency measures, its ability to discriminate pronunciation performance meaningfully beyond the tested dataset remains to be established. The observed variability across the individual indicators, together with their generally limited inter-correlations, suggests that the indicators capture partially distinct aspects of learner pronunciation performance, supporting the multidimensional evaluation strategy adopted in the proposed computational framework [8], [11], [16].

To examine the stability of the K-Means solution, clustering was repeated using multiple random seeds, and the resulting assignments were compared with the reference solution using the Adjusted Rand Index (ARI). Across the tested seeds (0, 10, 21, 42, 100, 123, and 2026), the ARI values ranged from 0.5857 to 1.0000, with a mean ARI of 0.7992. These results indicate moderate-to-high agreement across several initializations, although the lower ARI values show that the clustering solution is not completely invariant to initialization. Accordingly, the three-cluster solution should be interpreted as a stable exploratory grouping within the tested dataset rather than as a universally fixed learner classification. To further illustrate the variation represented by the PAS, three representative learner profiles were selected from the tested dataset, representing high, medium-, and low-PAS cases. Their corresponding pronunciation indicators are presented in Table 2.

Tabel 2. Representative learner profiles across high-, medium-, and low-PAS cases.

No	Learners	PAS	Correct	Substitution	Insertion	Deletion
1	learner01	74,50	136	18	2	2
2	learner02	39,69	2	96	0	58
3	learner03	70,10	121	32	4	3
4	learner14	67,86	118	36	6	2

The high-PAS learner exhibited relatively low transcription-based error rates (PER = 0.17; WER = 0.19) together with high acoustic similarity values, particularly MFCC Similarity (0.94) and Pitch Similarity (0.89). In contrast, the medium-PAS learner showed a mixed computational profile, with relatively high Rhythm Similarity (0.78) and MFCC Similarity (0.93), but lower Pitch Similarity (0.58) and higher transcription-based error rates (PER = 0.39; WER = 0.36). The low-PAS learner exhibited the highest WER among the three cases (0.48) and comparatively lower Rhythm Similarity (0.54) and MFCC Similarity (0.52), resulting in the lowest composite PAS (0.580).

To provide a qualitative interpretation of the PAS distribution, three representative learner profiles were examined, comprising a high-PAS case (learner25, PAS = 0.840), a medium-PAS case (learner23, PAS = 0.708), and a low-PAS case (learner27, PAS = 0.580), as presented in the representative profiles table. The high-PAS learner exhibited relatively low transcription-based error rates (PER = 0.17; WER = 0.19) together with high acoustic similarity values, particularly MFCC Similarity (0.94) and Pitch Similarity (0.89). In contrast, the medium-PAS learner showed a mixed computational profile, with relatively high Rhythm Similarity (0.78) and MFCC Similarity (0.93), but lower Pitch Similarity (0.58) and higher transcription-based error rates (PER = 0.39; WER = 0.36). The low-PAS learner exhibited the highest WER among the three cases (0.48) and comparatively lower Rhythm Similarity (0.54) and MFCC Similarity (0.52), resulting in the lowest composite PAS (0.580). These examples illustrate how the proposed PAS reflects the combined contribution of transcription-based and acoustic indicators rather than relying on a single pronunciation measure. Importantly, the three cases should be interpreted as computational profiles within the tested dataset rather than validated proficiency or pronunciation-quality categories. In particular, the medium-PAS example demonstrates that a learner may exhibit relatively strong performance on some acoustic dimensions while simultaneously showing higher transcription-based error rates. Such variation supports the use of a multidimensional representation for exploratory pronunciation analysis.

To complement the quantitative learner profiles, qualitative inspection of learner-reference differences provide illustrative examples of word-level pronunciation-related deviations observed in the computational analysis. As shown in Table Y, the learner-reference transcript alignment identifies substitution, deletion, and insertion patterns across the selected learner cases. These examples are interpreted cautiously because

differences in ASR output may reflect either learner pronunciation deviations or ASR recognition errors.

Table 3. Illustrative word-level pronunciation-related deviations

No	Learner	Reference form	Learner ASR output	Observed deviation
1.	learner01	historic	history	Word-level substitution
2.	learner02	shift	—	Word-level deletion
3.	learner03	—	on	Word-level insertion
4.	learner14	verge	first	Word-level substitution

These examples illustrate how learner–reference transcript alignment can identify word-level pronunciation-related deviations in learner speech. The observed differences represent discrepancies between the learner and reference transcripts and should not be interpreted as definitive phonetic error labels. In particular, transcript alignment alone cannot determine whether a discrepancy results from an actual learner pronunciation deviation or from an ASR recognition error. Therefore, the examples in Table 3 are presented as illustrative computational indicators rather than definitive phonetic diagnoses.

These examples illustrate how learner–reference transcript alignment can identify word-level pronunciation-related deviations, including substitution, deletion, and insertion patterns. The observed differences represent discrepancies between the learner and reference transcripts and may reflect pronunciation-related variation as captured by the ASR output. However, transcript alignment alone cannot determine whether a given discrepancy results from an actual learner pronunciation error or from an ASR recognition error. Therefore, the examples are presented as illustrative computational indicators rather than definitive phonetic diagnoses.

3.1.3. Correlation Analysis

To examine the relationships among the extracted pronunciation indicators and the proposed PAS, Pearson correlation analysis was performed. The resulting correlation matrix is presented in Table 4, which summarizes the pairwise linear relationships among PER, WER, Pitch Similarity, Rhythm Similarity, MFCC Similarity, and PAS.

Table 4. Pearson correlation coefficients between pronunciation indicators and PAS (N = 45).

Variable	PER	WER	Pitch	Rhythm	MFCC	PAS
PER	1.000	-0.048	0.083	-0.187	0.068	-0.325
WER	-0.048	1.000	-0.254	0.034	-0.050	-0.495
Pitch Similarity	0.083	-0.254	1.000	0.038	-0.089	0.510
Rhythm Similarity	-0.187	0.034	0.038	1.000	-0.004	0.544
MFCC Similarity	0.068	-0.050	-0.089	-0.004	1.000	0.430
PAS	-0.325	-0.495	0.510	0.544	0.430	1.000

Note: Pearson correlation coefficients are reported with two-tailed p -values and 95% confidence intervals for the correlations between the five pronunciation indicators and PAS. PER showed $r = -0.325$, $p = 0.029$, 95% CI $[-0.565, -0.035]$; WER showed $r = -0.495$, $p < 0.001$, 95% CI $[-0.689, -0.236]$; Pitch Similarity showed $r = 0.510$, $p < 0.001$, 95% CI $[0.254, 0.699]$; Rhythm Similarity showed $r = 0.544$, $p < 0.001$, 95% CI $[0.298, 0.722]$; and MFCC Similarity showed $r = 0.430$, $p = 0.003$, 95% CI $[0.157, 0.643]$. Confidence intervals were calculated using Fisher's z transformation.

As shown in Table 4, Rhythm Similarity exhibited the strongest positive correlation with the constructed PAS ($r = 0.544$, $p < 0.001$, 95% CI $[0.298, 0.722]$), followed by Pitch Similarity ($r = 0.510$, $p < 0.001$, 95% CI $[0.254, 0.699]$) and MFCC Similarity ($r = 0.430$, $p = 0.003$, 95% CI $[0.157, 0.643]$). Within the tested dataset, higher acoustic similarity values tended to coincide with higher values of the constructed PAS. Because these acoustic indicators are themselves components of PAS, the observed associations primarily describe the internal behavior of the composite score rather than independent effects on pronunciation quality. The results are consistent with previous studies suggesting that acoustic similarity provides valuable complementary information for automatic pronunciation evaluation [8], [11], [17].

Conversely, the transcription-based error measures exhibited negative correlations with PAS. WER demonstrated the strongest negative association ($r = -0.495$, $p < 0.001$, 95% CI $[-0.689, -0.236]$), whereas PER showed a weaker negative association ($r = -0.325$, $p = 0.029$, 95% CI $[-0.565, -0.035]$). Within the tested dataset, higher transcription error rates tended to coincide with lower values of the constructed PAS. As with the acoustic indicators,

these correlations should be interpreted as describing the internal behavior of the composite score rather than as evidence of independent effects on pronunciation quality.

The correlation matrix also reveals that the relationships among the individual pronunciation indicators themselves are generally weak. For example, PER and WER exhibited only a weak negative correlation ($r=-0.048$), while Pitch Similarity and Rhythm Similarity showed a negligible positive correlation ($r=0.038$). Likewise, MFCC Similarity demonstrated only weak associations with the remaining pronunciation indicators, with correlation coefficients close to zero. These observations suggest that the pronunciation indicators capture different aspects of learner speech rather than measuring exactly the same characteristic.

Overall, the correlation analysis indicates that the transcription-based error measures and acoustic similarity measures capture different aspects of learner speech within the proposed multidimensional framework. However, the correlations between PAS and its component indicators should be interpreted cautiously because PAS is mathematically constructed from PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity. Therefore, these correlations partly reflect the internal mathematical relationship between the composite score and its constituent variables and should not be interpreted as independent evidence that any individual indicator determines overall pronunciation quality. In this context, Rhythm Similarity and Pitch Similarity are appropriately described as the strongest correlates of the constructed PAS in the present dataset rather than as definitive determinants of pronunciation quality.

3.1.4. Machine Learning-based Learner Profiling

Building upon the statistical analyses presented in the previous sections, unsupervised machine learning was employed to investigate whether learners could be automatically grouped according to their multidimensional pronunciation characteristics. While the descriptive statistics demonstrated substantial variability among learners and the correlation analysis confirmed that the selected pronunciation indicators provide complementary information, machine learning-based learner profiling was performed to identify meaningful pronunciation patterns without requiring predefined proficiency labels [2], [27].

The clustering analysis utilized the five pronunciation indicators proposed in this study, namely PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity. Prior to clustering, all features were standardized using the StandardScaler transformation to eliminate differences in measurement scales and ensure equal contribution during Euclidean distance computation. The optimal number of learner groups was subsequently determined using the Elbow Method, with the resulting WCSS trend illustrated in Figure 4.

As shown in Figure 4, the WCSS decreases substantially from $K = 1$ to $K = 3$, after which the rate of improvement gradually diminishes, indicating a clear elbow point at $K = 3$. Therefore, three clusters were selected as the optimal configuration for machine learning-based learner profiling.

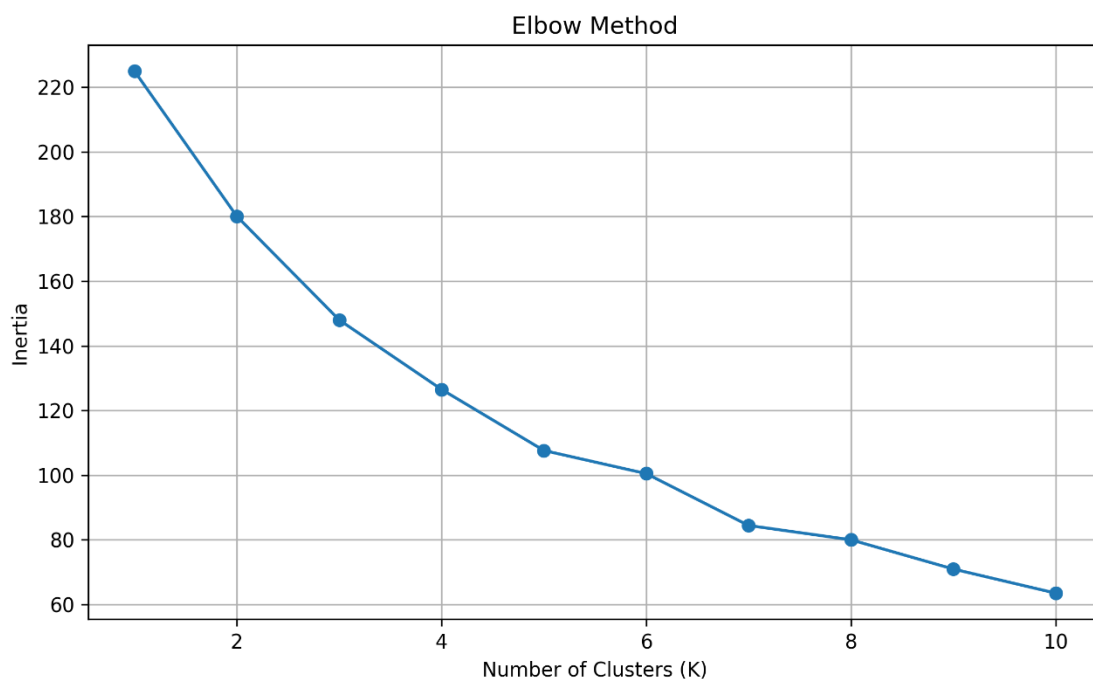


Figure 4. Elbow Method

Figure 4 presents the Elbow Method used to determine the optimal number of clusters. As the number of clusters increases, the WCSS decreases because observations become increasingly homogeneous within each cluster. A pronounced reduction in WCSS is observed between $K = 1$ and $K = 3$, after which the curve gradually flattens, indicating diminishing improvements in cluster compactness. Based on this elbow point, $K = 3$ was selected as the optimal configuration for subsequent learner profiling.

Figure 5 presents the Principal Component Analysis (PCA) visualization of learner pronunciation profiles obtained using K-Means clustering. Each point represents an individual learner projected onto the first two principal components for visualization purposes, while colors indicate the assigned learner cluster. Clustering was performed in the original standardized five-dimensional feature space consisting of PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity.

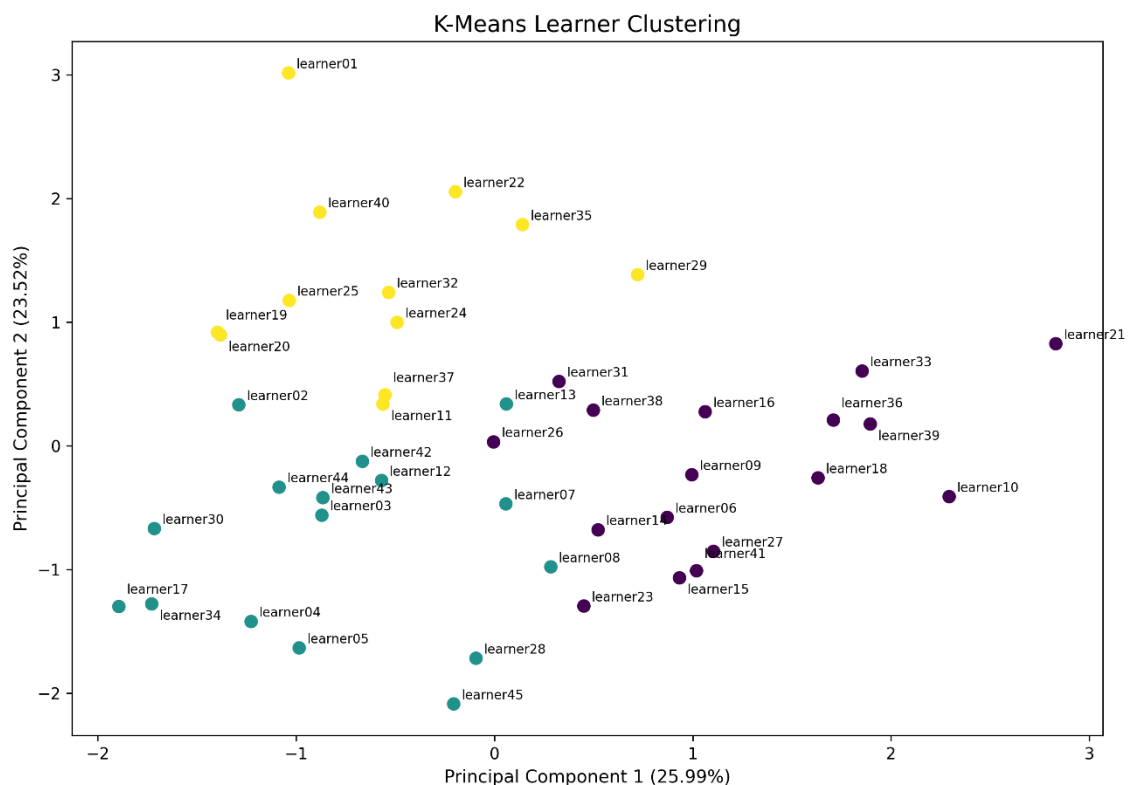


Figure 5. K-Means Learner Clustering

The PCA projection used to visualize the K-Means clustering in Figure 5 shows that PC1 explained 25.99% of the total variance, while PC2 explained 23.52%. Together, the first two principal components accounted for 49.51% of the total variance. After determining the optimal number of clusters, the standardized pronunciation feature vectors were grouped using the K-Means algorithm. To facilitate interpretation of the clustering results, PCA was employed to project the original five-dimensional feature space onto two principal components for visualization. As shown in Figure 5, PC1 explained 25.99% of the total variance, while PC2 explained 23.52%, with the first two components accounting for 49.51% collectively. The projected data reveal three distinguishable learner

clusters with some overlap between neighboring groups. Although PCA provides a two-dimensional representation for visualization, the clustering itself was performed using the original standardized five-dimensional feature space rather than the projected data [2].

The PCA projection therefore provides a visual representation of the multidimensional pronunciation feature space, while the quantitative cluster characteristics are further summarized in Table 5. The table presents the mean values of PAS, PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity for each cluster, enabling comparison of the pronunciation characteristics among the identified learner groups.

Table 5. Characteristics of Learner Clusters Obtained Using K-Means Clustering

Cluster	Number of Learners	Mean PAS	Mean PER	Mean WER	Mean Pitch	Mean Rhythm	Mean MFCC	Interpretation
Lower								
0	17	0.687	0.287	0.408	0.604	0.781	0.745	Pronunciation Performance
Intermediate								
1	16	0.694	0.348	0.272	0.709	0.606	0.776	Pronunciation Performance
Higher								
2	12	0.776	0.261	0.214	0.790	0.838	0.727	Pronunciation Performance

In contrast, Cluster 0, which contains 17 learners, obtained the lowest average PAS (0.687) and the highest WER (0.408), suggesting that transcription errors substantially affected overall pronunciation performance despite maintaining relatively high rhythm similarity. Cluster 1, comprising 16 learners, represents an intermediate pronunciation profile. Interestingly, this cluster achieved the highest MFCC Similarity (0.776) among all learner groups while exhibiting lower Pitch Similarity and Rhythm Similarity than Cluster 2. This observation indicates that spectral similarity alone is insufficient to fully characterize pronunciation quality and that prosodic information together with transcription-based accuracy provides additional complementary information [11], [17].

Taken together, the clustering analysis complements the statistical findings presented in Sections 3.2 and 3.3. While descriptive statistics characterize the overall distribution of learner pronunciation performance and correlation analysis identifies the relationships among pronunciation indicators, machine learning-based learner profiling demonstrates how these indicators jointly distinguish learners with different pronunciation characteristics. The experimental results therefore confirm that integrating transcription-based error metrics with complementary acoustic similarity measures provides an effective basis for objective learner categorization within the proposed computational framework and demonstrates its potential for supporting future data-driven pronunciation learning applications [27], [28], [30].

3.2. Discussion

3.2.1. Effectiveness, Comparison, and Implications

Furthermore, the machine learning-based learner profiling identifies three exploratory learner groups within the tested dataset, providing different pronunciation profiles rather than definitive or generalizable learner categories.. Unlike conventional pronunciation assessment methods that primarily rely on a single transcription or acoustic metric, the proposed framework combines PER, WER, Pitch Similarity, Rhythm Similarity, and MFCC Similarity into the proposed PAS. The statistical analyses and machine learning results consistently indicate that these indicators contribute complementary information for representing learner pronunciation from both linguistic and acoustic perspectives [8], [11], [16].

The effectiveness of the proposed framework is supported by the experimental findings presented in Sections 3.2–3.4. The descriptive statistics reveal measurable variability among learner pronunciation characteristics, while the correlation analysis shows that Rhythm Similarity and Pitch Similarity exhibit the strongest positive correlations with the constructed PAS, whereas Word Error Rate (WER) shows the strongest negative association. Because PAS is constructed from the same pronunciation indicators, these correlations should not be interpreted as evidence of causal determinants of pronunciation quality. Furthermore, the machine learning-based learner profiling identifies three exploratory learner groups with different pronunciation characteristics, suggesting that the multidimensional feature representation provides useful information for exploratory learner profiling.

Compared with previous Automatic Pronunciation Assessment (APA) studies, the proposed framework extends conventional pronunciation evaluation by integrating statistical analysis and machine learning-based learner profiling into a single computational framework. Previous studies have primarily emphasized pronunciation scoring or pronunciation error detection using automatic speech recognition and deep learning techniques [8], [10], [13]. In contrast, the proposed framework not only evaluates pronunciation accuracy but also characterizes learner pronunciation patterns through multidimensional acoustic and linguistic analysis. transcription-based error metrics alone [11], [27]. A comparative summary between representative APA studies and the proposed framework is presented in Table 6.

Table 6. Comparison between Representative Automatic Pronunciation Assessment Studies and the Proposed Framework

Study	Primary Approach	Multi-dimensional Pronunciation Indicators	Learner Profiling	Integrated PAS
Kheir et al. [8]	Review of Automatic Pronunciation Assessment	Partial	No	No
Gong et al. [9]	Transformer-based Pronunciation Assessment	Partial	No	No
Yan et al. [10]	Deep Learning-based Pronunciation Assessment	Partial	No	No
Proposed Framework	Whisper-based ASR + Acoustic Analysis + Machine Learning	Yes	Yes	Yes

Note: "Yes" indicates that the study explicitly implements the corresponding capability within the proposed framework; "No" indicates that the capability is not implemented; and "Partial" indicates that the related aspect is addressed to some extent but does not constitute a complete implementation of the capability.

An important contribution of this study is the demonstration that pronunciation quality cannot be adequately represented by a single pronunciation indicator. The clustering analysis shows that learners exhibiting higher spectral similarity do not necessarily achieve the highest overall pronunciation performance, whereas stronger pronunciation quality is consistently associated with lower transcription error rates together with higher prosodic similarity. These findings support the integration of complementary pronunciation indicators into a unified evaluation framework capable of providing a more comprehensive representation of learner pronunciation [11], [17].

The proposed framework also has practical implications for CAPT. By combining automatic speech recognition, acoustic feature analysis, statistical evaluation, and machine learning-based learner profiling, the framework enables objective pronunciation assessment while simultaneously providing interpretable learner profiles. Such information can support personalized pronunciation feedback, adaptive learning strategies, and learner progress monitoring in future CAPT systems. Consequently, the proposed framework provides a promising foundation for developing intelligent pronunciation learning environments that combine objective assessment with data-driven learner profiling [6], [28], [30].

3.2.2. Limitations and Future Research

Despite the encouraging results, several limitations should be acknowledged. First, the proposed framework was evaluated using a relatively small dataset consisting of one native English speaker and forty-five non-native learners reading the same speech passage. Although this controlled experimental setting enabled objective comparison between native and learner speech, it may not fully represent the diversity of pronunciation variations encountered in spontaneous speech or across speakers with different linguistic backgrounds. Furthermore, the proposed PAS has not yet been validated against evaluations performed by expert human raters. Consequently, although the experimental results demonstrate the feasibility of the proposed computational framework, additional validation is required to establish its generalizability, reliability, and practical applicability across broader pronunciation assessment scenarios.

Future research will focus on extending the proposed framework through larger and more diverse multilingual speech corpora involving multiple native speakers and learners

with different first-language backgrounds. Additional studies will investigate the integration of expert human ratings to validate the proposed PAS, as well as the incorporation of more advanced artificial intelligence techniques, such as transformer-based speech representation learning and explainable artificial intelligence (XAI), to improve model transparency and interpretability [3], [4], [10], [29]. Furthermore, integrating the proposed framework into real-time CAPT systems capable of providing adaptive and personalized pronunciation feedback represents a promising direction for future research [6], [28], [30].

4. CONCLUSION

This study proposed a computational framework for automated pronunciation error analysis through native–learner speech comparison by integrating Whisper-based ASR, audio preprocessing, pronunciation and acoustic feature extraction, statistical analysis, and machine learning-based learner profiling within a unified computational workflow. The framework integrates transcription-based pronunciation indicators, including PER and WER, with complementary acoustic similarity measures comprising Pitch Similarity, Rhythm Similarity, and MFCC Similarity to compute the proposed PAS. This integrated approach provides a multidimensional computational representation of learner pronunciation by considering both transcription-based errors and acoustic similarity to a native-speaker reference. The experimental results show the preliminary feasibility of the proposed framework for automated pronunciation analysis within the tested dataset. Correlation analysis indicated that Rhythm Similarity ($r = 0.544$) and Pitch Similarity ($r = 0.510$) exhibited the strongest positive associations with the constructed PAS, whereas WER showed the strongest negative association ($r = -0.495$). These relationships should be interpreted cautiously because PAS is an author-formulated composite score constructed from the same pronunciation indicators. The machine learning-based analysis identified three learner groups within the tested dataset, providing exploratory profiles of learner pronunciation characteristics rather than definitive or generalizable learner categories. The clustering analysis showed moderate stability across random seeds (mean ARI = 0.799), while the internal validation results indicated limited cluster separation (Silhouette Score = 0.1814). These findings therefore support the use of clustering as an exploratory component of the proposed framework.

The primary contribution of this study lies in the development of a computational framework that extends conventional automatic pronunciation assessment beyond single-metric evaluation. Rather than relying solely on transcription accuracy or individual acoustic features, the proposed framework integrates multiple complementary pronunciation indicators within a unified computational architecture while incorporating statistical analysis and machine learning-based learner profiling. The findings indicate the preliminary feasibility of this integrated approach for computational pronunciation analysis; however, the proposed PAS has not yet been validated against expert human pronunciation ratings and should therefore not be interpreted as a validated measure of pronunciation proficiency. Several limitations should be acknowledged. The experimental evaluation was conducted using a relatively small dataset comprising 45 non-native English learners and one native-speaker reference, with all participants using a single reading passage. The study also lacks expert human pronunciation scoring and external validation, and the use of ASR-generated transcripts may introduce potential bias, particularly for non-native speech. In addition, the reference-based design primarily characterizes learner pronunciation similarity relative to the selected native-speaker reference rather than directly measuring communicative intelligibility. Future research will therefore focus on larger and more diverse multilingual speech corpora, multiple native-speaker references, expert human-rater validation, external and psychometric validation of PAS, and further investigation of ASR robustness. Extending the framework toward real-time adaptive CAPT systems with personalized pronunciation feedback also represents a promising direction for future research; however, large-scale pronunciation assessment and educational deployment should be considered only after appropriate validation has been established.

REFERENCES

- [1] M. Farrús, "Automatic speech recognition in L2 learning: A review based on PRISMA methodology," *Languages*, vol. 8, no. 4, art. no. 242, 2023, doi: 10.3390/languages8040242.
- [2] H. Ahlawat, N. Aggarwal, and D. Gupta, "Automatic speech recognition: A survey of deep learning techniques and approaches," *Int. J. Cogn. Comput. Eng.*, vol. 6, pp. 201–237, 2025, doi: 10.1016/j.ijcce.2024.12.007.

- [3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.
- [4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, vol. 33, 2020. [Online]. Available: <http://arxiv.org/abs/2006.11477>
- [5] V. Roger, J. Farinas, and J. Pinquier, "Deep neural networks for automatic speech processing: A survey from large corpora to limited data," *EURASIP J. Audio Speech Music Process.*, vol. 2022, no. 1, pp. 1–15, 2022, doi: 10.1186/s13636-022-00251-w.
- [6] M. Amrate and P. H. Tsai, "Computer-assisted pronunciation training: A systematic review," *ReCALL*, vol. 37, no. 1, pp. 22–42, 2025, doi: 10.1017/S0958344024000181.
- [7] M. Y. C. Jiang *et al.*, "Integrating automatic speech recognition technology into vocabulary learning in a flipped English class for Chinese college students," *Front. Psychol.*, vol. 13, art. no. 902429, 2022, doi: 10.3389/fpsyg.2022.902429.
- [8] Y. El Kheir, A. Ali, and S. A. Chowdhury, "Automatic pronunciation assessment — A review," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, doi: 10.18653/v1/2023.findings-emnlp.557.
- [9] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7262–7266, doi: 10.1109/ICASSP43922.2022.9746743.
- [10] B.-C. Yan *et al.*, "An effective pronunciation assessment approach leveraging hierarchical transformers and pre-training strategies," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 2024, pp. 1737–1747, doi: 10.18653/v1/2024.acl-long.95.
- [11] W. Dong, C. Cucchiarini, R. van Hout, and H. Strik, "Predicting accentedness and comprehensibility through ASR scores and acoustic features," *Comput. Speech Lang.*, vol. 95, art. no. 101858, 2026, doi: 10.1016/j.csl.2025.101858.
- [12] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification," in *Proc. INTERSPEECH*, Dublin, Ireland, 2023, pp. 166–170, doi: 10.21437/Interspeech.2023-173.

- [13] M. Kim, "Digital enhancement of pronunciation assessment: Automated speech recognition and human raters," *Phonet. Speech Sci*, vol. 15, no. 2, pp. 13–20, 2023, doi: 10.13064/ksss.2023.15.2.013.
- [14] Y. Gao, B. M. L. Srivastava, and J. Salsman, "Spoken English intelligibility remediation with PocketSphinx alignment and feature extraction improves substantially over the state of the art," in *Proc. IMCEC*, 2018, doi: 10.1109/IMCEC.2018.8469649.
- [15] J. Wang and Y. Zheng, "Assessing intelligibility as conceptualised within the CEFR-companion volume framework using adaptive comparative judgement," *Lang. Test*, vol. 43, no. 1, pp. 9–36, 2025, doi: 10.1177/02655322251338605.
- [16] S. Ali, S. Tanweer, S. Khalid, and N. Rao, "Mel frequency cepstral coefficient: A review," in *Proc. 2nd Int. Conf. ICT Digit, Smart, Sustain. Develop*, 2021, doi: 10.4108/eai.27-2-2020.2303173.
- [17] F. W. S. dos Santos and H. K. de Souza, "Examining vowel intelligibility in Brazilian Portuguese EFL learners: A study using automatic speech recognition," *Ilha do Desterro*, vol. 77, pp. 243–274, 2024, doi: 10.5007/2175-8026.2024.e98870.
- [18] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, 1st ed. Englewood Cliffs, NJ, USA: PTR Prentice Hall, 1993.
- [19] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, 1975, doi: 10.1002/j.1538-7305.1975.tb02840.x.
- [20] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: A comparative analysis," *EURASIP J. Adv. Signal Process*, vol. 2015, no. 1, Dec. 2015, doi: 10.1186/s13634-015-0277-z.
- [21] J. Ramírez, J. C. Segura, C. Benítez, A. De La Torre, and A. Rubio, "Improved voice activity detection combining noise reduction and subband divergence measures," in *Proc. INTERSPEECH*, Jeju Island, Korea, 2004, pp. 961–964, doi: 10.21437/Interspeech.2004-347.
- [22] Z. Zaland *et al*, "Multichannel speech enhancement for automatic speech recognition: A literature review," *PeerJ Comput. Sci*, vol. 11, art. no. e2772, 2025, doi: 10.7717/peerj-cs.2772.
- [23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn*, 2023, pp. 28492–28518. [Online]. Available: <http://arxiv.org/abs/2212.04356>

- [24] V. K. Singh, "Transcripto: Fine-tuning multilingual ASR for Indian grievance feedback calls," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 42s, 2025, doi: 10.52783/jisem.v10i42s.7880.
- [25] E. M. N. Saeed, B. Al-sellami, M. Tawfik, and A. A. M. Abushariah, "KANWhisper: Leveraging learnable activation functions for interpretable and efficient Arabic automatic speech recognition," *Sci. Rep.*, vol. 16, art. no. 55863, 2026, doi: 10.1038/s41598-026-55863-5.
- [26] A. Babaeian, "Speaking to machine: Validity of pronunciation assessment by technology-driven systems," *Integr. J. Res. Arts Humanit.*, vol. 5, no. 3, pp. 276–279, 2025, doi: 10.55544/ijrah.5.3.31.
- [27] J. Gao, "An investigation of high-proficiency L2 English speakers' oral test performance: A profiling approach," *Front. Commun.*, vol. 7, art. no. 926409, 2022, doi: 10.3389/fcomm.2022.926409.
- [28] S. Inceoglu, W. H. Chen, and H. Lim, "Monitoring student behavior in autonomous automatic speech recognition-based pronunciation practice," *System*, vol. 124, art. no. 103387, 2024, doi: 10.1016/j.system.2024.103387.
- [29] V. Berisha and J. M. Liss, "Responsible development of clinical speech AI: Bridging the gap between clinical research and technology," *NPJ Digit. Med.*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01199-1.
- [30] M. Bashori, R. van Hout, H. Strik, and C. Cucchiari, "Look, I can speak correctly': Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology," *Comput. Assist. Lang. Learn.*, vol. 37, no. 4, pp. 1335–1363, 2024, doi: 10.1080/09588221.2022.2080230.