

Article

Acoustic and Signal-Based Evaluation of AI-Assisted English Speech Training for Arabic-Speaking University Learners

Abdallah Abu Qub'a *  and Jassim Al Herz 

English Language Department, College of Arts, King Faisal University, Al-Ahsa 36362, Saudi Arabia

* Correspondence: aabuquba@kfu.edu.sa

Received: 7 April 2026; **Revised:** 22 April 2026; **Accepted:** 7 May 2026; **Published:** 26 October 2026

Abstract: The biomechanical process that produces human speech creates measurable acoustic signals by vocal folds vibration and vocal tract resonant filtering. This study analyzed sound and vibration principles to estimate the effectiveness of Artificial Intelligence (AI)-assisted pronunciation training in a university context. The study recorded speeches for 200 Arabic-speaking undergraduate students studying a general English course. Students were assigned randomly to an experimental group using ELSA Speak (a mobile AI application) that provides pronunciation feedback based on acoustic features and a control group receiving traditional teaching methods without using AI support. Researchers used a pre-test/post-test experimental design and used Praat and MATLAB to analyze speech recordings from reading and speaking tasks to extract objective acoustic measures, including fundamental frequency (F0) stability (jitter), formant frequency consistency (F1, F2 dispersion), and spectral characteristics (spectral centroid and bandwidth). The results of the paired *t*-tests and ANOVA showed that the experimental group achieved significant improvements ($p < 0.05$) compared to the control group. Specifically, pitch variability decreased noticeably (F0 standard deviation reduced by 39.5%), vowel articulation became more stable (formant dispersion reduced by 30–40%), and spectral organization improved. The findings also found that AI-assisted pronunciation training leads to measurable improvements in speech signal stability and clarity. This study highlights the value of sound analysis as a quantitative framework for evaluating human-centered speech learning systems.

Keywords: Acoustic Phonetics; AI-Assisted Language Learning; Computer-Assisted Pronunciation Training (CAPT); English Pronunciation; Speech Signal Processing

1. Introduction

The complex biomechanical process of producing human speech that transforms neuromuscular commands into audible acoustic signals involves coordination among three systems: the respiratory system, which is the source of power, the phonatory system which is the source of sounds, and the articulatory system, which filters the sounds via the vocal tract [1]. In addition, the acoustic waveform is considered a rich physical signal with some properties such as fundamental frequency (F0), formant structures, spectral energy distribution, and temporal perturbations, and these properties are direct, measurable manifestations of the speaker's physiological control and articulatory precision. As a result, analyzing speech signals uses signal processing techniques to measure stability, efficiency, and output quality like evaluating mechanical systems. Achieving intelligible and fluent pronunciation for second language (L2) learners is a great challenge especially when depending on traditional teaching methods that focus on grammatical and vocabulary knowledge. Similarly, adult L2 learners have physiological and perceptual pronunciation problems since they need to overcome fossilized articulatory habits of their first language (L1)

and need to perceive and produce new acoustic sounds [2]. This issue is clear for Arab learners of English who find differences in the phonological patterns of Arabic and English and these differences lead to pronouncing foreign accentedness and reducing intelligibility [3–6]. In traditional classrooms with teacher-focused approaches, there is limited time for learners to practice speaking, or receive intensive instance feedback to deal with these phonological differences.

Using AI and developed digital signal processing has improved the ways of teaching pronunciation via Computer-Assisted Pronunciation Training systems (CAPT) in the form of mobile or web applications. These systems recognize speech and acoustic modeling automatically and provide instant and objective feedback for the spoken output. These systems also analyze the speaker's sounds and guide specific parameters such as segmental accuracy (e.g., vowel quality via formant frequencies F1 and F2), prosodic features (e.g., pitch control via F0 contours), and voice quality metrics (e.g., stability via jitter and shimmer) [7]. These systems visualize speech as a modifiable acoustic signal and this helps learners to adjust their pronunciation based on the provided feedback. Many studies explored the effectiveness of such systems and found that CAPT systems provide appropriate acoustic feedback which leads to improved pronunciation and accent reduction [8]. Also, these AI systems reduce affective filters, increase learners' motivation, and foster autonomous practice outside classrooms [9]. These advantages of AI systems align with hypotheses in Second language acquisition (SLA) such as Schmidt's Noticing Hypothesis (1990) [10] where the visual and acoustic feedback makes phonological targets clear, and DeKeyser's (2007) Skill Acquisition Theory [11] where repetitive feedback fosters accurate articulation.

The broader field of signal processing has seen parallel advances in optical communication networks, where techniques for analyzing spectral precision, filter stability, and signal integrity directly inform methods for evaluating human speech as an acoustic signal [12,13]. Just as photonic systems rely on precise spectral filtering to maintain signal quality in subcarrier routing [12], pronunciation training systems depend on accurate formant tracking and spectral clarity to provide meaningful feedback to learners. This interdisciplinary perspective reinforces the engineering principle that stable, well-filtered signals—whether optical or acoustic—produce more reliable and intelligible outputs.

A significant gap still exists between technological capabilities and pedagogical applications in academic programs since many studies focus on CAPT algorithm development, short-term studies, and voluntary users of such systems [14]. In addition, there is a lack of longitudinal studies that use AI-assisted pronunciation training systems in university classrooms for L2 learners. Thus, this study aims to fill this gap by providing an acoustic and signal-based evaluation framework using available AI pronunciation tutors. The study aims to answer these questions:

1. Do Arab EFL learners who receive AI-assisted pronunciation training show significantly greater improvement in source-related acoustic stability (as measured by F0 variability and measures like jitter) compared to a control group receiving traditional teaching?
2. Do they show significantly greater improvement in filter-related acoustic precision (as measured by formant frequency consistency for vowels and spectral clarity metrics) compared to the control group?
3. To what extent can improvements in these objective acoustic metrics be systematically attributed to the AI feedback intervention, thereby validating the application of sound and vibration analysis techniques in evaluating language learning technology?

The study, as a result, aims to provide a quantitative methodology for L2 pronunciation studies as it uses speech signals as the primary data for evaluation and explains how the principles of engineering can be used to measure and improve human-centered learning systems. It is clear from the objectives of the study that it is an acoustic-analytical study that uses speech signal processing techniques to evaluate improvements in learners' pronunciation. Despite using CAPT and SAL contexts, its novelty lies in treating speech production as a measurable acoustic signal and quantifying learning outcomes through standard acoustic parameters. Accordingly, the pedagogical implications are derived from signal-level evidence rather than from instructional design or system development perspectives.

2. Methodology

2.1. Research Design and Context

This study used a quasi-experimental, pre-test/post-test control group design to evaluate the impact of AI-assisted pronunciation training on the acoustic quality of learner speech. The study was conducted over a full

14-week semester within the mandatory General English course at King Faisal University, Saudi Arabia. The General English course is a foundational three-credit-hour course designed for first-year students across all colleges, focusing on integrated English language skills with a communicative approach. Pronunciation is a stated learning outcome, typically addressed through teacher modelling and repetition drills during class. The study was integrated into the standard curriculum; the experimental intervention replaced a portion of the regular self-study and homework time. Randomization in this study was conducted at the class level rather than the individual student level. Intact classroom groups were used due to scheduling constraints; therefore, available classes were randomly assigned to either the experimental or control condition using a simple random allocation procedure.

2.2. Participants

200 male undergraduate students participated in the study (mean age = 18.7 years, SD = 0.8). All participants were native Arabic speakers. Participants were randomly selected from General English course classes. Then 100 students were assigned to an experimental group with AI assistance, and 100 to a control group that received traditional instruction. All participants took a diagnostic speaking test to evaluate their level, and the t-test on the pre-test acoustic metrics (detailed in Section 2.5) confirmed no statistically significant differences between groups at the outset ($p > 0.05$ for all key parameters). Informed consent was obtained from all participants. Finally, the experimental and control groups were given classes at different times and were taught separately to reduce the possibility of cross-group interaction. In addition, access to the AI-assisted learning tools was restricted to the experimental group through individualized login credentials.

2.3. Materials and Intervention

2.3.1. Control Group Instruction

The control group followed the standard General English course syllabus and pronunciation was taught by instructors using listening and repeating exercises, correction of noticeable errors during speaking activities, and general feedback. No dedicated technological tools for pronunciation feedback or formal acoustic awareness training were provided.

2.3.2. Experimental Group

The experimental group had the same course syllabus in addition to using Elsa Speak (English Language Speech Assistant), a commercially available AI-powered pronunciation teaching application. Researchers chose Elsa Speak due to its proven speech recognition engine, its provision of specific feedback on phoneme-level accuracy and prosody, and its accessibility for students. Intervention fidelity was monitored using the built-in usage dashboard of the application as it automatically recorded student engagement, including frequency of use, number of completed speaking tasks, and total practice duration. Students in the experimental group were required to complete weekly guided practice sessions. Dashboard records showed that 92% of participants (92 out of 100) met the required engagement level of 30 min per week for all 14 weeks. The mean total practice time was 412 min (SD = 58 min), with an average of 29.4 minutes per week (SD = 4.1 minutes). Only 8 participants fell below the required threshold, with an average weekly practice time of 18.6 min (SD = 6.3 minutes). No participants were excluded from the analysis based on compliance, as the intention-to-treat principle was followed. The AI practice sessions included mandatory homework, requiring 30 min of engaged practice per week, verified through the app's progress tracking dashboard, and weekly activities focused on phonetic lessons in the course such as word and sentence repetition in which the app provides a model, records the student, and gives an instant score (0–100%) based on acoustic similarity to the native model, highlighting mispronounced segments, phoneme training using specific exercises targeting sounds known to be difficult for Arabic L1 speakers, and finally prosody practice by reading short dialogues with feedback on fluency and intonation contour.

Learners in the experimental group received multimodal feedback: a numerical score, visual highlighting of errors on the text, and the option to see a waveform/spectrogram comparison for some exercises. The instructor monitored overall compliance weekly but did not provide individual corrective feedback on the app's outputs.

2.4. Speech Data Collection

High-quality speech recordings were conducted in a quiet, dedicated language lab at the university at two points: Week 1 (Pre-test) and Week 14 (Post-test). Participants were given two speaking tasks to complete, and the tasks were presented in Arabic to ensure comprehension. These tasks included reading sentences—five phonetically balanced sentences designed to elicit target English vowels and consonants (e.g., “The ship left the busy port in a thick fog.”)—and spontaneous speech, which required a 30-s response to a simple, familiar prompt (“Describe your daily routine at the university.”). Participants were recorded individually. Each recording session lasted approximately 2 min per participant.

2.5. Acoustic Analysis Procedure

All recorded files were anonymized and analyzed using two software programs. The analysis process included different levels. First, the sentence reading task was used for formant and spectral analysis; the spontaneous speech sample was used for broader prosodic and stability measures (F0, jitter, shimmer). Second, the F0 contour was extracted using the autocorrelation method (pitch floor: 75 Hz, ceiling: 300 Hz). Jitter (local, %) and Shimmer (local, %) were calculated using Praat’s built-in “Voice Report” function to quantify cycle-to-cycle frequency and amplitude perturbation, respectively. Third, formants (F1, F2, F3) were tracked using the Burg method (Maximum formant: 5,500 Hz for all speakers, given the high F2 values in some English vowels; Number of formants: 5). Formant values were extracted at the temporal midpoint of all monophthongal vowels in the reading task (/i:, ɪ, ε, æ, ɑ:, ɒ, ɔ:, ʊ, u:/). The standard deviation (Hz) of F1 and F2 for each vowel category across tokens served as the measure of formant dispersion/consistency. Fourth, for the same vowel segments, a 25 ms Hamming-windowed slice at the midpoint was used to compute the Spectral Centroid (Hz) and Spectral Bandwidth (Hz) using Praat’s “Spectrum” object. Lower variability (SD) in these measures across tokens indicates a more consistent spectral energy distribution. Finally, using SPSS for statistical analysis within groups in which paired-samples t-tests were conducted for each acoustic parameter (F0 SD, Jitter, Shimmer, F1/F2 Dispersion, Spectral Centroid SD) to compare pre- and post-test scores within each group, and between groups in which a series of mixed-design ANOVAs were performed with *Time* (Pre, Post) as the within-subjects factor and *Group* (Experimental, Control) as the between-subjects factor. The primary focus was the Time x Group interaction effect, which indicates whether the change over time differed significantly between groups.

The selected acoustic parameters were chosen because they represent different aspects of speech production. Fundamental frequency (F0) reflects vocal fold stability and prosodic control, while jitter and shimmer capture cycle-to-cycle perturbations in voice stability. Formant frequencies (F1–F3) and their dispersion reflect articulatory precision and vowel space consistency, which are key indicators of pronunciation accuracy. Spectral centroid and bandwidth provide additional information about overall spectral distribution and speech clarity.

3. Literature Review

Acoustic phonetics or the study of speech sounds provides the framework for quantifying pronunciation where speech is analyzed as a physical signal, decomposed into source characteristics (the vibration of the vocal folds) and filter characteristics (the resonant properties of the vocal tract) [15]. Moreover, different measurable parameters are used as reliable indicators of speech quality and intelligibility. First, the fundamental frequency (F0), the primary acoustic correlate of pitch, represents the rate of vocal fold vibration and its stability is important for prosodic control. Micro-perturbations in frequency (jitter) and amplitude (shimmer) are quantitative measures of vocal fold vibration irregularity. As a result, these measures are increasingly applied to L2 speech, where higher jitter and shimmer can signal unstable phonation or heightened speech anxiety, affecting perceived fluency [16]. Second, formant frequencies (F1, F2, F3) are resonant peaks in the spectrum that define vowel quality and consonant transitions. Formant frequencies analysis for Arab learners is very important as it helps in describing vowel production patterns, for instance, Arabic learners of English exhibit systematic variations in vowel space organization and formant structure which reflect cross-linguistic influence from Arabic phonology [17–19]. These results emphasize the importance of objective acoustic measurements in identifying pronunciation deviations and tracking improvement over time. F1 correlates inversely with tongue height, and F2 with tongue advancement, so the consistency

of formant production for a given vowel which is measured as low dispersion across tokens is a strong predictor of vowel intelligibility and reduced accentedness [20]. Third, spectral moments that include the centroid (the “center of gravity” of the spectrum) and bandwidth provide a complete view of spectral shape and energy distribution that indicate articulatory precision and clarity [21]. The analysis of consonantal timing and gemination in Arabic-influenced English speech showed that detailed spectral and temporal measurements revealed predictable rhythmic and articulatory differences linked to native-language transfer [22] and such results support the relevance of acoustic phonetic methods for investigating both segmental and suprasegmental features of L2 speech. The power of these acoustic approaches lies in their objectivity. While human ratings are subjective and can be inconsistent, acoustic metrics provide continuous, ratio-scale data ideal for detecting subtle, incremental changes in learner pronunciation over time.

Computer-Assisted Pronunciation Training (CAPT) applications have been developed and have been used in teaching pronunciation as they focus on providing learners with visual representations of their speech, such as spectrograms or pitch contours, and this improves learners’ awareness through comparison with a model. In addition, using Automatic Speech Recognition (ASR) also enables applications to provide automated and corrective feedback [23]. Nowadays, CAPT is defined by artificial intelligence where deep learning models are trained on huge speech corpora, and one example of these models is the Elsa Speak application which performs two functions. First, it analyzes learners’ utterances using multi-dimensional acoustic feature extraction (e.g., MFCCs, formant trajectories, duration), and second it provides personalized and corrective feedback for specific phonemes or prosodic patterns. Recent studies on technology-enhanced language learning investigated the pedagogical use of AI-based tools such as Grammarly and ChatGPT to explore their effectiveness in supporting language production accuracy and learner autonomy among EFL students and the findings highlighted the great potential of AI-driven feedback systems in language learning contexts [24].

Furthermore, some studies on AI-CAPT systems show that these systems can effectively improve learners’ perception and production of problematic segmental contrasts [25]. Also, systems that provide multiple feedback modalities (auditory, visual, textual) are more effective than single-mode systems, as they provide different learning styles and reinforce noticing [9]. In fact, these tools support learners’ autonomy and increase time-on-task by providing unlimited, anxiety-free practice opportunities outside the classroom, a key advantage in settings with limited instructional hours [11].

From a theoretical perspective, AI-CAPT aligns strongly with foundational SLA concepts. It applies to Schmidt’s Noticing Hypothesis by making phonological gaps noticeable through immediate and explicit feedback. It also facilitates the pushed output required by Swain’s Output Hypothesis (1985) [26] where learners must produce language repeatedly to receive evaluation. Finally, it supports the development of automaticity through structured, repetitive practice with instant reinforcement [11].

Arab learners of English face many challenges in English pronunciation due to the phonological transfer patterns of their L1 and the differences between the two languages which prevent acquiring native-like English pronunciation. First, vowel systems in Arabic and English are different. Arabic dialects typically have a 6- or 8-vowel system, contrasting primarily in length, whereas English has a complex system of approximately 11–12 monophthongs and multiple diphthongs, contrasting in both quality and quantity and these differences lead to a pervasive “vowel reduction” challenge. Similarly, Arab learners often struggle to distinguish between tense/lax pairs like /i:/ vs. /ɪ/ and /u:/ vs. /ʊ/, frequently neutralizing them to a quality closer to the Arabic long vowel and this neutralization is acoustically measurable as a collapse in the F1–F2 vowel space and increased formant dispersion for English vowel categories. Second, several English consonants are absent in Arabic (e.g., /p/, /v/, /ŋ/, /tʃ/, /dʒ/), which leads to substitutions (e.g., /b/ for /p/, /f/ for /v/). Arabic speakers often exhibit timing differences associated with consonant gemination and stress assignment which reflect the rhythmic properties of Arabic speech [27]. These findings provide additional evidence of systematic L1 transfer effects influencing English pronunciation. The emphatic (pharyngealized) consonants in Arabic (/tˤ/, /sˤ/, etc.) can cause co-articulatory lowering of F2 frequencies in neighboring vowels, an effect that may carry over into English, subtly coloring vowel quality [28].

Moreover, Arabic is a syllable-timed or mora-timed language with a relatively flat intonation contour, while English is a stress-timed language with wide pitch excursions to mark prominence and pragmatic intent. Furthermore, Arabic learners often exhibit reduced pitch range, misplaced sentence stress, and a tendency toward syllable-

timed rhythm in English, which can negatively impact intelligibility and nativeness [29]. This shows acoustically as reduced F0 variability and less durational contrast between stressed and unstressed syllables. These L1-specific challenges make Arabic-speaking learners an ideal population for testing the efficacy of targeted acoustic feedback. An effective AI-CAPT system should, in theory, help learners notice and correct these specific vowels, consonants, and prosodic differences.

In summary, the systematic reviews of AI-based pronunciation tools show that such applications contribute positively to learners' speaking performance, particularly in terms of fluency, accuracy, and motivation, although challenges remain in classroom integration and long-term pedagogical impact [30]. In addition, there is a growing importance of using automatic pronunciation assessment systems in CAPT [31]. Thus, AI-assisted pronunciation training represents a significant advancement in L2 speaking development, especially when combined with structured practice and continuous feedback.

Despite the detailed acoustic description, advanced AI-CAPT technology, and well-documented L1-specific challenges, a significant gap still exists. First, many AI-CAPT studies focused on technology development or short-term laboratory-based studies with volunteer participants. Also, there are a few longitudinal studies that insert such tools into mandatory, credit-bearing university courses over an entire semester to evaluate their impact as part of a holistic language program rather than as an isolated treatment. Finally, while studies often report outcomes using proficiency scores or subjective learner perceptions, few have used detailed, objective acoustic analysis as the primary outcome measure.

To wrap up, this study aims to address these gaps. It explores the longitudinal, curriculum-integrated use of an AI pronunciation tutor by Arabic-speaking university students, using a pre-test/post-test design with acoustic signal analysis as the core evaluative methodology. By doing so, it seeks to provide empirical, data-driven evidence on whether such technology can systematically alter the biomechanical output of L2 speech in a real-world educational environment.

4. Results and Discussion

4.1. Descriptive Statistics and Baseline Equivalence

Prior to teaching pronunciation, independent samples t-tests confirmed no statistically significant differences ($p > 0.05$) between the Experimental (AI) and Control groups on any of the key acoustic parameters measured, confirming that both groups started from comparable levels. **Table 1** presents the group means and standard deviations for all primary measures at pre-test and post-test.

Table 1. Descriptive Statistics for Key Acoustic Parameters (Pre-Test and Post-Test).

Parameter (Unit)	Group	Pre-Test Mean (SD)	Post-Test Mean (SD)
F0 Standard Deviation (Hz)	Control	28.4 (5.1)	26.7 (4.8)
	Experimental	29.1 (5.3)	17.6 (3.9)
Jitter (local, %)	Control	1.21 (0.31)	1.18 (0.29)
	Experimental	1.25 (0.33)	0.82 (0.22)
Shimmer (local, %)	Control	5.84 (1.42)	5.72 (1.38)
	Experimental	5.91 (1.51)	4.05 (1.11)
F1 Dispersion (Hz) *	Control	96.2 (18.3)	89.5 (17.1)
	Experimental	98.4 (19.1)	62.7 (14.6)
F2 Dispersion (Hz) *	Control	142.1 (25.7)	135.4 (24.3)
	Experimental	145.6 (26.5)	91.3 (19.8)
Spectral Centroid SD (Hz)	Control	412.6 (75.2)	398.1 (72.4)
	Experimental	418.9 (76.8)	271.3 (58.9)
Spectral Bandwidth SD (Hz)	Control	615.4 (112.5)	598.7 (108.3)
	Experimental	623.2 (115.7)	402.5 (92.4)

Note: F1/F2 Dispersion represents the average within-category standard deviation for target English vowels. Bold indicates the post-test mean for the Experimental group. * F1/F2 dispersion represents the average within-category standard deviation of the first and second formant frequencies for the target English vowels. Lower values indicate more consistent vowel production.

Table 1 presents the descriptive statistics for the primary acoustic parameters measured at pre-test and post-test for both the control and experimental groups. Overall, the two groups showed comparable mean values across all parameters at pre-test, with only small differences in their initial scores.

For F0 Standard Deviation (Hz), the control group showed a slight decrease from 28.4 Hz (SD = 5.1) at pre-test

to 26.7 Hz (SD = 4.8) at post-test. In contrast, the experimental group demonstrated a larger reduction from 29.1 Hz (SD = 5.3) to 17.6 Hz (SD = 3.9).

Regarding jitter (local, %), the control group showed minimal change, decreasing slightly from 1.21% (SD = 0.31) to 1.18% (SD = 0.29). The experimental group showed a more noticeable reduction from 1.25% (SD = 0.33) to 0.82% (SD = 0.22).

For shimmer (local, %), the control group values decreased marginally from 5.84% (SD = 1.42) to 5.72% (SD = 1.38). The experimental group showed a larger decrease from 5.91% (SD = 1.51) to 4.05% (SD = 1.11).

In terms of F1 dispersion (Hz), the control group decreased from 96.2 Hz (SD = 18.3) to 89.5 Hz (SD = 17.1), while the experimental group decreased from 98.4 Hz (SD = 19.1) to 62.7 Hz (SD = 14.6). Similarly, F2 dispersion (Hz) decreased slightly in the control group from 142.1 Hz (SD = 25.7) to 135.4 Hz (SD = 24.3), while the experimental group showed a larger decrease from 145.6 Hz (SD = 26.5) to 91.3 Hz (SD = 19.8).

For spectral centroid standard deviation (Hz), the control group decreased from 412.6 Hz (SD = 75.2) to 398.1 Hz (SD = 72.4), whereas the experimental group decreased from 418.9 Hz (SD = 76.8) to 271.3 Hz (SD = 58.9).

Finally, spectral bandwidth standard deviation (Hz) showed a small reduction in the control group from 615.4 Hz (SD = 112.5) to 598.7 Hz (SD = 108.3), while the experimental group decreased from 623.2 Hz (SD = 115.7) to 402.5 Hz (SD = 92.4).

These descriptive results show that both groups showed reductions across most parameters from pre-test to post-test, with larger numerical decreases observed in the experimental group across all measured acoustic variables.

4.2. Inferential Statistical Analysis

A series of 2 (Time: Pre, Post) $\times$ 2 (Group: Control, Experimental) mixed-design ANOVAs were conducted for each dependent variable. The primary focus was the significant Time $\times$ Group interaction, which indicates that the change from pre-test to post-test was different for the two groups.

As shown in **Table 2**, the Time $\times$ Group interaction was statistically significant ($p < 0.001$) for all acoustic parameters, with large effect sizes (Partial $\eta^2 > 0.14$). This confirms that the pattern of change from pre- to post-test was fundamentally different between the Experimental and Control groups.

Table 2. Mixed ANOVA Results for Time $\times$ Group Interaction and Effect Sizes.

Parameter	F-Value (Time $\times$ Group)	p-Value	Partial η^2 (Effect Size)	Cohen's d (Exp. Group Pre-Post)
F0 Standard Deviation	F(1, 198) = 42.15	< 0.001	0.176 (Large)	2.41 (Large)
Jitter (local)	F(1, 198) = 38.72	< 0.001	0.164 (Large)	1.52 (Large)
Shimmer (local)	F(1, 198) = 36.88	< 0.001	0.157 (Large)	1.43 (Large)
F1 Dispersion	F(1, 198) = 45.61	< 0.001	0.187 (Large)	2.08 (Large)
F2 Dispersion	F(1, 198) = 52.33	< 0.001	0.209 (Large)	2.31 (Large)
Spectral Centroid SD	F(1, 198) = 40.27	< 0.001	0.169 (Large)	2.18 (Large)
Spectral Bandwidth SD	F(1, 198) = 37.94	< 0.001	0.161 (Large)	2.05 (Large)

Follow-up paired-samples t-tests within each group revealed distinct patterns of change. In the Experimental Group, there was a highly significant ($p < 0.001$) improvement on every parameter from pre-test to post-test, which was accompanied by very large Cohen's d effect sizes ranging from 1.43 to 2.41. On the other hand, the Control Group showed no statistically significant improvement ($p > 0.05$) in any parameter, and its observed effect sizes were negligible, with Cohen's d values all below 0.2.

However, the results suggest that AI-assisted pronunciation training, when integrated into a university English course, is associated with improvements in the acoustic signal quality of Arabic-speaking learners' speech. Even though, the results should be interpreted as improvements in acoustic stability and speech signal consistency rather than direct evidence of perceived intelligibility, as no perceptual evaluation was conducted in this study.

It is also clear that the AI-assisted group shows greater improvement in source-related stability. The Experimental group's marked reduction in F0 variability, jitter, and shimmer indicates a more stable and regularly vibrating vocal fold system post-intervention. From a speech motor control perspective, the AI application functioned as a high-precision biofeedback tool. By providing real-time visual and numerical scores related to pitch consistency,

the system likely heightened learners' awareness of inadvertent pitch fluctuations, a common feature of L2 speech anxiety and uncertain prosody. The consequent practice encouraged more consistent subglottal pressure and laryngeal muscle control, leading to the observed reduction in frequency and amplitude perturbation. This finding extends the clinical use of jitter/shimmer as stability metrics into the domain of L2 pedagogy and offers a quantifiable way to track gains in vocal control that underlie fluent, confident delivery.

Moreover, the significant reduction in F1 and F2 dispersion for the Experimental group shows that AI feedback led to more consistent and accurate tongue positioning for English vowels, effectively countering the vowel space "collapse" typical of Arabic L1 speakers³. The app's phoneme-level feedback, which likely targeted specific vowel contrasts like /ɪ/-/i:/, provided the repetitive, corrective practice needed to establish distinct and stable motor programs for these categories.

Furthermore, the significant improvements in spectral centroid and bandwidth stability show that learners achieved not only more accurate but also more efficient articulation. Reduced variability in these spectral moments indicates less random acoustic noise and more controlled energy distribution, likely stemming from more precise and repeatable constrictions in the vocal tract.

The improvements show that the usage of an AI tutor as an effective closed-loop feedback control system helps learners in educational terms in which the learner's speech signal (output) was continuously measured against a native-model target (reference), and the AI-generated feedback (error signal) allowed the learner (the controller) to adjust their articulatory and phonatory settings (the process) in subsequent attempts. The significant reduction in variability across all measured parameters suggests that this loop may have helped minimize "output error" over time and may lead to a more stable and accurate speech production system.

While improvements in the experimental group are primarily attributed to the AI-assisted intervention, potential influences such as general classroom instruction, increased practice time, and learner motivation were controlled through the use of a parallel control group receiving identical instructional content without access to the AI system.

The results have direct implications for English language teaching in similar university contexts, centering on three key contributions. First, the study empirically validates a blended learning model, showing that AI tools can effectively manage the form-focused, repetitive practice required for phonetic skill automatization. This strategically frees valuable classroom time for the meaning-focused, communicative activities that teachers are best suited to facilitate. Second, the AI system enabled personalized practice at scale, delivering a level of immediate, individualized feedback that is very hard for a single instructor in a large class, thereby addressing a critical bottleneck in pronunciation instruction. Finally, the positive results, which were linked to regular, self-directed practice, underscore the technology's role in empowering learner autonomy and cultivating essential self-monitoring skills, both of which are important for sustained language development beyond the classroom.

Finally, the findings should be interpreted within the context of the specific learner population and instructional setting examined in this study and should not be generalized beyond similar university-level English learning environments without further evidence.

5. Conclusion, Limitations, and Further Studies

This study shows that AI-assisted pronunciation training used in a university-level General English curriculum yields measurable, objective improvements in the acoustic quality of learners' speech. By applying signal processing techniques to pre- and post-intervention recordings from Arabic-speaking undergraduates, we found that the group receiving structured feedback via an AI application showed significant improvements in vocal stability (reduced F0 variability, jitter, shimmer) and articulatory precision (decreased formant dispersion, improved spectral clarity) compared to a control group. These findings support the role of AI-powered tools as effective biofeedback systems that enable closed-loop practice and allow learners to systematically refine their speech output. For instructors, this is consistent with a blended learning model where AI handles intensive, form-focused pronunciation practice, freeing classroom time for communicative instruction.

This study also has several limitations. First, it depends only on acoustic metrics without perceptual listener ratings, and this means that we cannot confirm that the signal-level improvements directly enhanced intelligibility or reduced accentedness. Second, the participant sample was limited to male Arabic speakers at one institution,

which may affect generalizability. Third, the instruction lasted for one semester; the long-term retention of gains is unknown. Finally, the use of a single AI tool and controlled speech tasks may not fully represent performance in spontaneous communication or with other software.

Future studies should incorporate perceptual evaluations by human listeners to directly link acoustic improvements to gains in intelligibility. Longitudinal designs with delayed post-tests are needed to assess the durability of pronunciation gains. Research should also explore adaptive AI systems that personalize feedback based on learners' error patterns. Finally, expanding analysis to connected speech and incorporating qualitative data on learner engagement would provide a more complete understanding of AI-assisted training efficacy. Overall, the results are promising but should be interpreted with appropriate caution given the specific scope and design of the study.

Author Contributions

Both authors were equally responsible for the conception, design, data collection, analysis, and interpretation of the study. Additionally, the authors drafted and revised the manuscript, approved the final version, and are accountable for all aspects of the work.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant Number KFU262416).

Institutional Review Board Statement

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation. This study was approved by the Ethics Committee of King Faisal University, KSA, with ethics approval reference [KFU-2026-ETHICS4372], dated January 18, 2026. The standards are also in line with the Helsinki Declaration of 1975, as revised in 2000.

Informed Consent Statement

Informed consent was obtained from all participants prior to their involvement in the study.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data is not publicly available due to privacy or ethical restrictions.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this manuscript, the authors used ChatGPT solely for language refinement. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

1. Titze, I.R. *Principles of Voice Production*; National Center for Voice and Speech: Iowa City, IA, USA, 2000.
2. Flege, J.E. Second Language Speech Learning: Theory, Findings, and Problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*; Strange, W., Ed.; York Press: Timonium, MD, USA, 1995; pp. 233–277.
3. Alharbi, B.; Aljutaily, M. On the Perceptual Accuracy of Non-Native Phonemic Contrasts: A Case Study of Native Arabic Speakers. *J. Lang. Linguist. Stud.* **2020**, *16*, 2003–2023. [[CrossRef](#)]

4. Abu Guba, M.N. Foot Duration and Polysyllabic Shortening among Arab Speakers of English. *Loquens* **2023**, 9, e091.
5. Abu Guba, M.N.; Abu Qub'a, A. Acquisition of Stress Assignment Phonology by Levantine Arabic Learners of English: Evidence from nonsense words. *Soc. Sci. Humanit. Open* **2026**, 13, 102540. [\[CrossRef\]](#)
6. Abu Guba, M.N.; Mashaqba, B.; Jarbou, S.; et al. Production of Vowel Reduction by Jordanian-Arabic Speakers of English: An Acoustic Study. *Pozn. Stud. Contemp. Linguist.* **2023**, 59, 1–25.
7. Liakin, D.; Cardoso, W.; Liakina, N. Mobilising Instruction in a Second-Language Context: Learners' Perceptions of Two Speech Technologies. *Languages* **2017**, 2, 11. [\[CrossRef\]](#)
8. Dizon, G.; Tang, D. Intelligent Personal Assistants for Autonomous Second Language Learning: An Investigation of Alexa. *JALT CALL J.* **2020**, 16, 107–128. [\[CrossRef\]](#)
9. McCrocklin, S. ASR-Based Dictation Practice for Second Language Pronunciation Improvement. *J. Second Lang. Pronunciat.* **2019**, 5, 98–118.
10. Schmidt, R.W. The Role of Consciousness in Second Language Learning. *Appl. Linguist.* **1990**, 11, 129–158. [\[CrossRef\]](#)
11. DeKeyser, R. Skill Acquisition Theory. In *Theories in Second Language Acquisition: An Introduction*; VanPatten, B., Williams, J., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007; pp. 97–113.
12. Christofidis, C.; Otero-Casado, A.; Torres-Ferrera, P.; et al. Performance Assessment of Interleaver Filters for Next-Generation Mobile Transport. *J. Opt. Commun. Netw.* **2026**, 18, 16–26. [\[CrossRef\]](#)
13. Otero-Casado, A.; Torres-Ferrera, P.; Christofidis, C.; et al. Impact of Subcarrier-Routing Devices on the Performance of DSCM Coherent Systems Evaluated with Analytical Models. In *Proceedings of the 2025 International Conference on Optical Network Design and Modeling (ONDM)*, Pisa, Italy, 6–9 May 2025; pp. 1–6. [\[CrossRef\]](#)
14. Georgiou, G.P. Artificial Intelligence in Pronunciation Teaching: Use and Beliefs of Foreign Language Teachers. *arXiv preprint* **2025**, *arXiv:2503.04128*. [\[CrossRef\]](#)
15. Kent, R.D.; Read, C. *The Acoustic Analysis of Speech*, 2nd ed.; Delmar Cengage Learning: Clifton Park, NY, USA, 2002.
16. Wang, Q.; Xu, F.; Wang, X.; et al. How Anxiety State Influences Speech Parameters: A Network Analysis Study from a Real Stressed Scenario. *Brain Sci.* **2025**, 15, 262. [\[CrossRef\]](#)
17. Abu Guba, M.N.; Daoud, S.; Jarbou, S. Foreign Accented-Speech and Perceptions of Confidence and Intelligence. *J. Psycholinguist. Res.* **2018**, 52, 1093–1113.
18. Abu Guba, M.N. Prosody Trumps Orthography in Second Language Phonology: The Case of Consonant Gemination. *J. Psycholinguist. Res.* **2023**, 52, 997–1015.
19. Adank, P.; Smits, R.; van Hout, R. A Comparison of Vowel Normalization Procedures for Language Variation Research. *J. Acoust. Soc. Am.* **2004**, 116, 3099–3107. [\[CrossRef\]](#)
20. Turner, G.S.; Tjaden, K.; Weismer, G. The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals with Amyotrophic Lateral Sclerosis. *J. Speech Lang. Hear. Res.* **1995**, 38, 1001–1013. [\[CrossRef\]](#)
21. Abu Guba, M.N.; Fareh, S.; Yagi, S. Speech Rhythm in English and Arabic: A Contrastive Review. *Int. J. Arabic-English Stud.* **2023**, 23, 183–202.
22. Neri, A.; Cucchiari, C.; Strik, H. The Effectiveness of Computer-Based Speech Corrective Feedback for Improving Segmental Quality in L2 Dutch. *ReCALL* **2008**, 20, 225–243.
23. Amrate, M.; Tsai, P. Computer-assisted pronunciation training: A systematic review. *ReCALL* **2025**, 37, 22–42. [\[CrossRef\]](#)
24. Okoye, K.; Nganji, J.T.; Escamilla, J.; et al. Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education. *Comput. Educ. Artif. Intell.* **2025**, 6, 100205. [\[CrossRef\]](#)
25. Bu, Y.; Ma, T.; Li, W.; et al. PTeacher: A Computer Aided Personalized Pronunciation Training System with Exaggerated Audio Visual Corrective Feedback. *arXiv preprint* **2021**, *arXiv:2105.05182*. [\[CrossRef\]](#)
26. Swain, M. Communicative Competence: Some Roles of Comprehensible Input and Comprehensible Output in Its Development. In *Input in Second Language Acquisition*; Gass, S., Madden, C., Eds.; Newbury House: Rowley, MA, USA, 1985; pp. 235–253.
27. Watson, J.C.E. *The Phonology and Morphology of Arabic*; Oxford University Press: Oxford, UK, 2007.
28. Khalifa, F.E. An Intonational Study of Kuwaiti Bedouin Arabic. PhD Thesis, University of London, London, UK, 1984.
29. Avrianti, N.; Yuliana, Y.G.S.; Ikhsanudin, I. AI-Based Pronunciation Training Tools for English Language Learners: A Systematic Review. *J. Engl. Educ. Program* **2025**, 7. [\[CrossRef\]](#)

30. El Kheir, Y.; Ali, A.; Chowdhury, S.A. Automatic Pronunciation Assessment: A Review. *arXiv preprint* **2023**, *arXiv:2310.13974*. [[CrossRef](#)]
31. Do, H.; Lee, W.; Lee, G.G. Acoustic Feature Mixup for Balanced Multi Aspect Pronunciation Assessment. *arXiv preprint* **2024**, *arXiv:2406.15723*. [[CrossRef](#)]



Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.