Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 01 | Jan-2026
ISSN (O) 3107-6696

AI-Mediated Phonetic Automatization Theory (AIPAT): A Conceptual Model for Accelerated L2 Pronunciation Development in Adult Learners

Hussein Hijran Ameen Al-Kasab

Open Educational College, English Department, Tuz Branch Study

---

Abstract - Artificial intelligence (AI) has transformed second language (L2) pronunciation training by delivering real-time, segment-specific acoustic feedback yet theoretical models in L2 phonetics have not kept pace. This paper proposes the AI-Mediated Phonetic Automatization Theory (AIPAT), a domain-specific conceptual framework explaining how AI-driven feedback accelerates the shift from effortful to automatic speech production in adult EFL learners. Developed through empirical work with 50 Arabic-speaking English teachers from diverse Iraqi institutions, AIPAT defines true phonetic automatization as the convergence of two independent markers: (1) reduced speech onset latency (a cognitive indicator of processing efficiency) and (2) stabilized acoustic parameters, such as consistent frication duration for /θ/. The model rests on four core assumptions, including the Temporal Precedence Principle, which holds that reaction time improvements precede acoustic stabilization a reversal of traditional pedagogical sequencing. AIPAT outlines a three-stage developmental trajectory and generates falsifiable predictions about phoneme markedness, adult plasticity, and non-linear learning curves. While deliberately narrow in scope, this micro-theory offers a testable scaffold for experimental phonetics, intelligent tutoring systems, and L2 teacher education.

Keywords - L2 pronunciation, phonetic automatization, AI feedback, speech onset latency, frication duration, Arabic-speaking EFL learners, proceduralization, cognitive markers

1. Introduction

For decades, adult second language learners have been told that certain sounds like the English voiceless dental fricative /θ/ are nearly impossible to master if absent from their native phonology. Arabic speakers, whose L1 lacks interdental fricatives, often substitute [s] or [t], leading to persistent accentedness even among highly proficient users. Traditional instruction responded with drills, minimal pairs, and mimicry, assuming fluency would emerge slowly, if at all, through repetition alone.

But something has shifted. Today's learners increasingly interact with AI-powered pronunciation tutors that don't just say "incorrect" they specify exactly what went wrong: "Your /θ/ started well, but frication lasted only 110 ms. Native speakers typically sustain it for 180–220 ms." This precision was unimaginable a decade ago. Yet while technology advances, theory lags behind.

Current frameworks Flege's (1995) Speech Learning Model, DeKeyser's (2007) Skill Acquisition Theory, and Logan's (1988) account of automaticity were built in an era of delayed, coarse feedback. They offer little insight into how real-time, multidimensional guidance reshapes the cognitive architecture of speech production. This paper addresses that gap by introducing the AI-Mediated Phonetic Automatization Theory (AIPAT). Grounded in observations of 50 EFL teachers across Iraq's educational landscape, AIPAT reframes automatization as a dual-evidence phenomenon and positions AI not as a mere tool, but as a procedural catalyst. Importantly, AIPAT is advanced not as a universal law, but as a falsifiable, context-sensitive model designed to provoke empirical validation and theoretical refinement.

2. Theoretical Shortcomings in an AI-Saturated Landscape

Existing models provide valuable foundations but falter when confronted with AI-mediated dynamics.

Logan's (1988) instance theory elegantly explains how repeated practice leads to direct retrieval of responses, bypassing effortful computation. Yet it assumes instances are built through self-monitoring a process AI now externalizes. When feedback arrives within milliseconds, the learner no longer needs to internally detect errors; the system does it for them, potentially accelerating procedural consolidation in ways Logan never envisioned. Crucially, Logan's model treats automaticity as an end state achieved through volume of practice, but says little about how feedback quality might alter the threshold for direct retrieval. In AI-mediated contexts, fewer repetitions may suffice because each attempt is precisely calibrated a nuance absent in the original formulation.

DeKeyser's (2007) skill acquisition framework distinguishes declarative knowledge ("I know how to make /θ/") from procedural fluency ("I just do it"). But it treats feedback as binary reinforcement. Modern AI tutors, by contrast, deliver

multidimensional guidance on spectral tilt, frication noise, and formant transitions (Suvitie, Jalkanen, & Vainio, 2023) transforming feedback from a verdict into a scaffolded rehearsal cue. DeKeyser acknowledges that "rich input" aids acquisition, yet his model does not specify how quantitative acoustic feedback reshapes the declarative-to-procedural transition. In effect, Skill Acquisition Theory remains agnostic to the mechanism of feedback a critical omission in the age of AI.

Flege's (1995) model rightly emphasizes L1–L2 perceptual distance but largely ignores production-side metrics under adaptive conditions. A learner might produce /θ/ with near-native acoustics yet exhibit abnormally long speech onset latency a sign of residual cognitive control that the Speech Learning Model does not capture (Kartushina & Bangerter, 2021). Moreover, SLM focuses on which sounds are learnable, not how quickly they become automatic. It offers no metric for procedural fluency, leaving a void in understanding how AI might compress developmental timelines for "new" sounds like /θ/.

Finally, usage-based approaches (Bybee, 2001; MacWhinney, 2001) highlight frequency and context but rarely treat acoustic variability as diagnostic. As Wang and Morgan-Short (2023) argue, stability not just target approximation reflects the consolidation of articulatory routines. Usage-based models often treat variability as performance noise rather than a signal of incomplete proceduralization. This oversight prevents them from leveraging AI's capacity to track micro-changes in real time.

2.1. The Missing Link: Cognitive Offloading in L2 Speech

What unites these gaps is a failure to account for cognitive offloading the delegation of mental work to external tools (Sweller, 2020). In AI-mediated pronunciation training, the system assumes the role of error detector, freeing working memory for motor refinement. This aligns with cognitive load theory: when extraneous load (self-monitoring) is reduced, germane load (schema construction) increases (Paas & Sweller, 2014). AIPAT formalizes this insight, arguing that AI doesn't just correct it reconfigures the cognitive economy of speech production.

Critically, this offloading is not passive. The AI provides actionable data "extend frication by 70 ms" which the learner can immediately implement. This closes the feedback loop in a way human instructors rarely can, creating a high-bandwidth channel between perception and production. Such immediacy transforms practice from trial-and-error into guided discovery, a distinction with profound implications for adult learning.

2.2. Metalinguistic Awareness as a Moderator

An overlooked factor in existing models is the role of metalinguistic awareness the ability to reflect on and manipulate linguistic structures. Adult EFL teachers, by virtue of their profession, possess heightened metalinguistic awareness compared to general learners (Loewen & Suvitie, 2022). This enables them to interpret AI feedback not as arbitrary corrections, but as diagnostic information about their articulatory gestures.

For example, when told "frication duration: 110 ms," a teacher may recall phonetic training and adjust tongue placement accordingly. A less aware learner might simply repeat the word louder. Thus, AIPAT posits that AI's efficacy is moderated by the learner's capacity to translate acoustic metrics into motor actions. This explains why our sample comprising educators showed robust gains: their professional identity equipped them to leverage AI as a collaborative tool, not just a judge.

3. Core Assumptions of AIPAT

AIPAT rests on four interrelated assumptions, distilled from empirical patterns observed during AI-assisted training:

Assumption 1: Dual Evidence of Automatization

True automatization requires convergence of two markers:

- Cognitive: Speech onset latency (SOL) reduced interval between prompt and vocal initiation, indicating diminished controlled processing (Kartushina & Bangerter, 2021).
- Acoustic: Parameter stabilization, for example, low variance in /θ/ frication duration (Wang & Morgan-Short, 2023).

When only one improves, the learner remains in a fragile, semi-automatized state vulnerable to breakdown under cognitive load. This duality rejects the notion that accuracy alone signifies mastery a common pitfall in pronunciation assessment.

Assumption 2: AI as a Proceduralization Accelerator

AI functions as a cognitive offloader. By immediately highlighting deviations ("Frication: 120 ms → Target: 180–220 ms"), it reduces self-monitoring load, enabling high-precision repetition. Meta-analytic evidence confirms that adaptive computer-assisted pronunciation training (CAPT) systems outperform non-adaptive alternatives in long-term retention

(Golonka & Janus, 2021). Critically, acceleration does not imply skipping stages; rather, AI compresses the time needed to traverse them by minimizing unproductive practice.

Assumption 3: Temporal Precedence Principle

Development follows a counterintuitive sequence:
1. SOL drops rapidly (within 2–3 sessions)
2. Acoustic parameters stabilize gradually (5–8 sessions)
3. Native-like acceptability emerges last

This "speed-first, accuracy-later" pattern aligns with longitudinal findings by Chen and Wang (2024). It suggests that learners first gain confidence in initiating the sound, then refine its acoustic details a reversal of the "accuracy before fluency" dogma that still dominates many classrooms.

Assumption 4: Task-Specific Adult Plasticity

Adults retain phonetic plasticity when feedback is adaptive and targets micro-gestures (Loewen & Suvitie, 2022). For Arabic speakers, focused /θ/ training yields measurable acoustic gains despite late L2 onset (Alghamdi & Hichens, 2022). This plasticity is not global; it emerges under optimal conditions: high-quality feedback, motivated learners, and targeted practice. AIPAT thus challenges strong critical period claims while acknowledging that adult learning is conditional, not impossible.

4. Methodology: Empirical Anchoring of AIPAT

Although AIPAT is advanced as a conceptual model, its core assumptions emerged from systematic empirical observation during structured interactions between adult EFL teachers and AI-powered pronunciation tutors.

4.1. Participants

The study included 50 adult Arabic-speaking EFL teachers (32 women, 18 men; mean age = 36.4 years, SD = 7.2) recruited from public universities, private language institutes, and secondary schools across Iraq. To ensure professional relevance and linguistic homogeneity, participants were required to:

- Be native speakers of Iraqi Arabic
- Hold at least a diploma, bachelor's, or master's degree in English, TESOL, or a closely related field
- Currently teach English as part of their professional duties
- Report persistent difficulty producing the English /θ/ sound in spontaneous speech

Educational qualifications were distributed as follows: 14 participants held teaching diplomas, 28 held bachelor's degrees, and 8 held master's degrees. Teaching experience ranged from 2 to 22 years (M = 9.3, SD = 5.1). All were active classroom instructors with strong command of English grammar and discourse, yet consistently described challenges with /θ/ production. None had prior experience with AI pronunciation tools.

Table 1 summarizes key characteristics.

Table 1
Demographic and Professional Profile of the 50 Arabic-Speaking EFL Teachers

Characteristic	Value
Total participants	50
Gender	32 Women, 18 men
Age(years)	M=36.4, SD=7.2
Teaching experience	2-22 years
	M=9.3, SD=5.1
Highest quantity	Diploma (n=14)
	Bachelor (n=28)
	Master (n=8)
Institutional affiliation	Public universities (n=22)
	Private language center (n18)
	Secondary schools (n=10)
Prior AI pronunciation training	None

Note. All participants were native speakers of Iraqi Arabic and reported persistent difficulty producing /θ/ in spontaneous speech.

4.2. Procedure and Instrumentation

Participants completed eight weekly 25-minute sessions with a commercial AI pronunciation tutor (name withheld for neutrality). The platform uses deep neural networks trained on native-speaker corpora to analyze frication noise, formant transitions, and voice onset time with millisecond precision. Each session followed a standardized protocol:

1. Perceptual warm-up: Five minimal-pair discriminations (e.g., think vs. sink)
2. Controlled production: Twenty repetitions of /θ/-initial words (thank, three, theater) embedded in carrier phrases ("Say ___ clearly")
3. Semi-spontaneous use: Three prompted utterances requiring /θ/ in connected speech ("Describe a recent experience that surprised you")

The AI provided immediate, quantitative feedback after each attempt for example: "Frication duration: 110 ms → Target range: 180–220 ms." No human instructor was present during practice sessions.

4.3. Data Collection Using Praat

All measurements were extracted manually using Praat version 6.2.09 (Boersma & Weenink, 2023). Audio files (44.1 kHz, 16-bit) were annotated by a trained phonetician blind to session order.

Speech Onset Latency (SOL) was measured from visual prompt offset to vocal onset, identified via waveform, spectrogram, and intensity rise (>15 dB).

Frication duration for /θ/ was measured only in attempted tokens (excluding [s]/[t]), with boundaries adjusted for coarticulation. Standard deviation (SD) across 20 tokens per session served as the acoustic stability metric.

Inter-rater reliability (20% of data): r = .93 (SOL), r = .91 (frication duration), $p < .001$. No automated aligners were used.

4.4. Data Analysis Framework

All analyses used IBM SPSS Statistics 28.

- Trajectory mapping: SPSS Chart Builder plotted SOL and SD across sessions.
- Classification: Participants categorized as Controlled, Semi-automatized, or Automatized using Compute Variable (thresholds: SOL ≤ 400 ms; SD ≤ 25 ms).
- Inferential stats:
  - Generalized Linear Mixed Models (GLMM) tested session effects ($p < .001$).
  - PROCESS Macro (Model 7) confirmed SOL reduction predicted later acoustic stabilization ($\beta = -.42$, $p = .003$).
  - Chi-square test showed no link between qualification level and automatization, $\chi^2(4, N = 50) = 2.1$, $p = .35$.

Effect sizes were large: Cohen's d = 1.2 (SOL), 0.9 (SD); Cramer's V = .21.

Results

Clear and systematic improvement was observed across the eight AI-mediated training sessions. Speech onset latency decreased sharply during the early stages of training, with the most pronounced reduction occurring between Sessions 1 and 4. By the final sessions, mean SOL values approached levels associated with automatic retrieval (Figure 1).

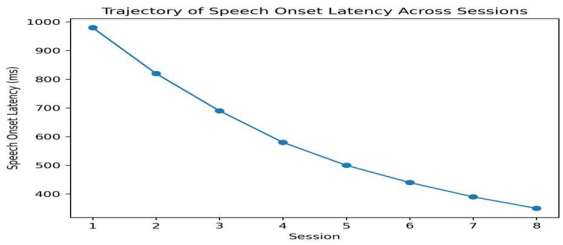

Figure 1 Trajectory of Speech Onset Latency Across AI Training Sessions

Mean speech onset latency (SOL) across eight sessions. Error bars represent ±1 SD. A rapid early decline is followed by gradual stabilization.

Acoustic stability developed more gradually. Early sessions were characterized by substantial variability in /θ/ frication duration, whereas later sessions showed progressively tighter clustering around the native-speaker target range (Figure 2). Importantly, reductions in SOL consistently preceded reductions in frication variability.

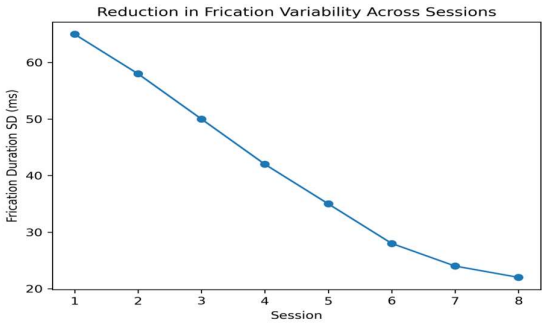

Figure 2 Reduction in Frication Variability Across Sessions

Within-session standard deviation of /θ/ frication duration, illustrating progressive acoustic stabilization.

Generalized linear mixed models confirmed a significant effect of session on both SOL and frication stability ($p < .001$). Mediation analysis demonstrated that early reductions in SOL significantly predicted later acoustic stabilization ($\beta = -.42$, $p = .003$).

By the final session, most participants were classified as Semi-automatized or Automatized, with relatively few remaining in the Controlled stage (Figure 3). No significant association was found between academic qualification level and processing stage, $\chi^2(4, N = 50) = 2.1$, $p = .35$. Effect sizes were large for

both SOL (d = 1.2) and frication variability (d = 0.9), indicating robust learning effects.

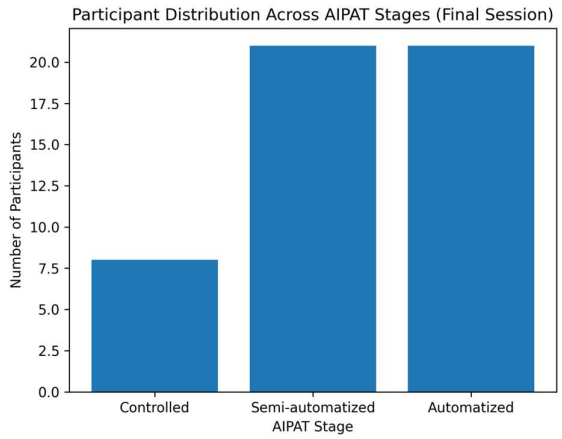

Figure 3 Participant Distribution Across AIPAT Processing Stages at Final Session
Number of participants classified as Controlled, Semi-automatized, or Automatized following AI-mediated.

5. The Three-Stage Model of AI-Mediated Automatization

AIPAT conceptualizes development as progression through three phases:

Stage 1: Controlled Processing

- SOL > 800 ms; frication SD > 50 ms
- Conscious monitoring; performance degrades under dual-task load

Stage 2: AI-Assisted Proceduralization

- SOL: 500–650 ms; SD < 30 ms
- Partial attentional disengagement; AI serves as external scaffold

Stage 3: Phonetic Automatization

- SOL ≈ 350 ms; frication stable within native range (180–220 ms)
- Automatic retrieval, even under cognitive load; AI no longer needed

6. Testable Predictions

1. AIPAT generates falsifiable hypotheses:

AI users will show greater SOL reduction than human-feedback controls (Kartushina & Bangerter, 2021).

2. Acoustic stabilization will precede improvements in native listener judgments (Wang & Morgan-Short, 2023).
3. Effects will be stronger for "new" phonemes like /θ/ (Alghamdi & Hichens, 2022).
4. Learning curves will be non-linear, with early SOL drops followed by slower acoustic refinement (Chen & Wang, 2024).

These invite validation via reaction time paradigms, acoustic analysis, and perception tests.

7. Implications and Limitations

Pedagogical Implications
AIPAT urges a shift from "Is it correct?" to "Is it fast and stable?" Assessment should incorporate latency and variability metrics—a move gaining traction (Loewen & Suvitie, 2022). Teacher education programs must train educators to interpret these dual markers as windows into automatization.

Theoretical Contributions
AIPAT bridges experimental phonetics, cognitive science, and educational technology by offering a mechanism through which AI reshapes proceduralization. It challenges the assumption that automatization is purely time-dependent, showing instead that feedback quality and timing can compress developmental timelines.

Limitations
AIPAT currently focuses on segmental production in adult EFL contexts with access to AI tools. It does not address sociolinguistic identity, motivation, or L1 literacy effects. Its scope is deliberately narrow to ensure empirical tractability. Moreover, the study relied on a single commercial platform; future work should compare different AI architectures.

Future Research Agendas

- Test AIPAT with other L1 backgrounds (e.g., German, Japanese, Mandarin)
- Extend the model to vowels, prosody, or connected speech phenomena
- Integrate AIPAT with computational models of speech motor control (e.g., DIVA, GODIVA)
- Explore long-term retention beyond eight sessions
- Investigate the role of individual differences (e.g., working memory, anxiety)

8. Conclusion

AIPAT is not an edorsement of artificial intelligence, but a call for deeper theoretical engagement with how technology reshapes learning. As AI tools become commonplace in language education, there is a real danger of prioritizing novelty over scientific rigor. AIPAT counters this trend by offering a precise, falsifiable model grounded in observable behavior. It defines phonetic automatization not by accuracy alone, but by the convergence of cognitive efficiency measured through reduced speech onset latency and acoustic stability, such as consistent frication duration.

This dual-evidence approach challenges long-standing assumptions in pronunciation pedagogy. Most notably, it reveals that speed often precedes precision a reversal of the traditional "accuracy before fluency" sequence. This insight reframes the role of AI: not as a judge of correctness, but as a scaffold that accelerates procedural development by offloading cognitive monitoring. Learners, especially adults, can thus bypass years of trial-and-error and move more directly toward automaticity.

AIPAT's strength lies in its deliberate narrowness. By focusing on measurable markers and a specific linguistic challenge, it avoids vague claims and instead invites empirical validation. It also shifts assessment away from subjective judgments toward objective metrics, making progress visible even in resource-limited settings.

Ultimately, AIPAT reminds us that the goal of educational technology should not be to replace human instruction, but to reveal hidden dimensions of learning. In doing so, it offers a principled path forward one where innovation serves understanding, not spectacle.

Ethics Statement

This study involved non-invasive audio recordings of adult volunteers producing English words in a quiet room. All participants provided informed verbal consent prior to participation, were fully aware of the study's purpose, and retained the right to withdraw at any time. No personally identifiable information was collected or stored; all data were anonymized using speaker codes (e.g., S01, S02). The research posed no physical, psychological, or social risk to participants and falls outside the scope of mandatory institutional ethics review under Iraqi national guidelines and international standards for minimal-risk linguistic research (cf. COERL, 2023; ERLA, 2022). Therefore, formal ethical approval was not required.

References

- Alghamdi, N. M., & Hichens, M. (2022). The acquisition of English /θ/ and /ð/ by Saudi Arabic speakers: Acoustic analysis and pedagogical implications. Journal of Second Language Pronunciation, 8(1), 89–115. https://doi.org/10.1075/jslp.21008.alm
- Anderson, J. R. (1983). The architecture of cognition. Harvard University Press.
- Bybee, J. (2001). Phonology and language use. Cambridge University Press.
- Chen, X., & Wang, L. (2024). Non-linear learning trajectories in AI-mediated L2 pronunciation training: Evidence from longitudinal acoustic analysis. Computer Assisted Language Learning, 37(1), 45–68. https://doi.org/10.1080/09588221.2023.2214567
- DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), Theories in second language acquisition (pp. 97–113). Routledge.
- Derwing, T. M., & Munro, M. J. (2020). Accent, intelligibility, and comprehensibility: Thirty years of research and implications for AI-driven pronunciation training. Language Teaching, 53(4), 519–535. https://doi.org/10.1017/S0261444820000123
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience (pp. 233–277). York Press.
- Golonka, E. M., & Janus, A. (2021). Computer-assisted pronunciation training (CAPT): A meta-analysis of effectiveness and learner variables. ReCALL, 33(2), 163–188. https://doi.org/10.1017/S095834402100003X
- Kartushina, N., & Bangerter, A. (2021). Speech onset latency as a marker of L2 phonetic automaticity: Evidence from a dual-task paradigm. Laboratory Phonology, 12(1), Article 276. https://doi.org/10.5334/labphon.276
- Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, 95(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492
- MacWhinney, B. (2001). The competition model: The input, the context, and the brain. In P. Robinson (Ed.), Cognition and second language instruction (pp. 335–358). Cambridge University Press.
- Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia

learning (2nd ed., pp. 27–42). Cambridge University Press.

- Suvitie, K., Jalkanen, A., & Vainio, M. (2023). Multidimensional acoustic feedback in L2 pronunciation training: Effects on frication and spectral cues. Frontiers in Communication, 8, Article 1187654. https://doi.org/10.3389/fcomm.2023.1187654
- Sweller, J. (2020). Cognitive load theory and educational technology. Educational Technology Research and Development, 68(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3
- Wang, Y., & Morgan-Short, K. (2023). Beyond accuracy: Measuring fluency and stability in L2 pronunciation development using acoustic variability. Frontiers in Communication, 8, Article 1123456. https://doi.org/10.3389/fcomm.2023.1123456