BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Voice Onset Time in the Production of English Stops: A Comparative Study of Native and Arabic-Speaking EFL Learners

Hesham Al-Damen [1], Mohamad Almashour [1], Mutasim Al-Deaibes [2*], Rami AlSharefeen [3]

[1] *Department of English Language and Literature, School of Foreign Languages, University of Jordan, Amman 11942, Jordan*

[2] *Department of English Language and Literature, Faculty of Arts, Yarmouk University, Irbid 21163, Jordan*

[3] *Academic Affairs Section, Rabdan Academy, Abu Dhabi 22401, United Arab Emirates*

ABSTRACT

This study examines the production of Voice Onset Time (VOT) in English stop consonants by native speakers of American English and Arabic-speaking English as a Foreign Language (EFL) learners at two proficiency levels. VOT, an acoustic parameter, is an essential feature in distinguishing between voiced and voiceless stops. Drawing on Flege's Speech Learning Model (SLM), the research investigates whether learners differentiate between voiceless and voiced stops (/p/ vs. /b/) and apply appropriate aspiration in /sp/ clusters, and whether proficiency influences VOT patterns. Data were collected from 29 native English speakers and 58 Arabic-speaking learners, who produced minimal pairs and /sp/ cluster words embedded in carrier sentences. All tokens were annotated manually in Praat and analyzed using linear mixed effects models. Results showed that native speakers maintained robust VOT distinctions, while Novice-High learners exhibited overlapping distributions between /p/ and /b/ and inappropriate aspiration in clusters. Intermediate-High learners produced more target-like patterns, suggesting early stages of L2 category formation. Findings support the SLM's predictions and underscore the need for explicit instruction on VOT contrasts and improvements in AI-assisted pronunciation feedback tools. The study concludes with some pedagogical implications for pronunciation instruction. For example, teachers working with Arabic-speaking learners should highlight the role of aspiration in English voicing contrasts and explicitly address its absence in /sp/ clusters.

Keywords: Voice Onset Time (VOT); Speech Learning Model (SLM); Arabic EFL Learners; Second Language Phonetics; Aspiration; Phonetic Category Formation; Acoustic Analysis; Pronunciation Instruction; AI in Language Learning

*CORRESPONDING AUTHOR:
Mutasim Al-Deaibes, Department of English Language and Literature, Faculty of Arts, Yarmouk University, Irbid 21163, Jordan; Email: mutasim.aldeaibes@yu.edu.jo

ARTICLE INFO
Received: 12 April 2025 | Revised: 13 May 2025 | Accepted: 19 June 2025 | Published Online: 16 July 2025
DOI: https://doi.org/10.30564/fls.v7i7.9466

CITATION
Al-Damen, H., Almshour, M., Al-Deaibes, M., et al., 2025. Voice Onset Time in the Production of English Stops: A Comparative Study of Native and Arabic-Speaking EFL Learners. Forum for Linguistic Studies. 7(7): 747–757. DOI: https://doi.org/10.30564/fls.v7i7.9466

1. Introduction

Voice Onset Time (VOT) is one of the most widely studied acoustic cues for distinguishing voicing in stop consonants [1–7] (to mention a few). In English, this distinction is primarily realized through aspiration: voiceless stops such as /p/, /t/, and /k/ are produced with a long-lag VOT, while voiced stops such as /b/, /d/, and /g/ have short-lag or near-zero VOT. Mastery of these contrasts is essential for intelligibility, as misarticulated voicing distinctions can lead to misunderstandings and contribute to accentedness in second language (L2) speech [6–8]. In their groundbreaking research [9], VOT across eleven different languages and categorized these languages into two primary groups based on their VOT contrasts. The first group consists of languages such as English, German, and Swedish, which exhibit a long-lag (positive) VOT for voiceless stops and a short-lag or zero VOT for voiced stops. The second group includes languages like Dutch, Russian, and Swedish, in which voiceless stops show a short positive VOT while voiced stops display a lead (negative) VOT. They noted that the lead VOT values for voiced stops range from -125 to -75 ms, indicating that voicing begins before the stop burst occurs. For voiceless unaspirated stops, VOT values fall between 0 and 25 ms, meaning voicing starts simultaneously with the stop burst. In contrast, voiceless aspirated stops have a VOT of 60 to 100 ms due to the prolonged aspiration that characterizes these sounds. In a subsequent study on Spanish word-initial stops [10], found that using VOT to distinguish between voiced and voiceless stops was both noticeable and effective.

The VOT values found in different spoken Arabic dialects show variability and inconsistency based on the specific dialect. The results can be categorized into two primary groups: The first group of studies indicates that voiced plosives are articulated with lead (negative) voicing, while voiceless plosives are produced with short-lag (positive) voicing, seen in Lebanese and Jordanian Arabic [7,11,12]. In contrast, the second group of studies reveals that in the Ghamdi dialect of Saudi Arabic, voiced stops have led (negative) voicing, whereas voiceless stops are articulated with long-lag (positive) voicing. For Arabic-speaking learners of English, acquiring accurate VOT patterns presents unique challenges. While Arabic exhibits a voicing contrast between /b/ and /t/, it does not utilize aspiration as a phonemic feature. Consequently, Arabic learners may have difficulty perceiving and producing the aspirated–unaspirated contrast that characterizes English voiceless stops. These difficulties may be further compounded in specific phonological environments, such as /s/-stop clusters (e.g., speak, spin), where English suppresses aspiration despite the presence of a voiceless stop. In such environments, learners often overgeneralize the aspiration rule, leading to unnatural production patterns [13].

This study is informed by Flege's Speech Learning Model (SLM), which posits that L2 learners' ability to acquire new phonetic categories depends on the perceived similarity between L1 and L2 sounds [10,14,15]. According to the model, when an L2 sound is perceived as sufficiently different from any L1 sound, learners are more likely to form a new category. Conversely, when the perceived difference is small, learners may assimilate the L2 sound into an existing L1 category, resulting in compromised perception and production. Applying this framework, Arabic-speaking learners may either fail to differentiate English /p/ from /b/ due to assimilation or gradually form a distinct /p/ category with increased proficiency and exposure.

Despite the importance of accurate stop production, much remains unknown about how Arabic EFL learners at different proficiency levels negotiate English VOT contrasts, particularly in cluster environments. Furthermore, with the growing use of AI-based pronunciation tools, there is an urgent need for annotated L2 data that reflects real learner variability and informs the development of fine-grained acoustic feedback systems.

This study investigates VOT production in three stop contexts—voiceless /p/, voiced /b/, and /p/ in /sp/ clusters—among native English speakers and Arabic-speaking learners at Novice-High and Intermediate-High proficiency levels. It addresses the following research questions:

1. Do native speakers and L2 learners maintain contrasts between the VOTs of voiceless /p/ and voiced /b/?
2. Do native speakers and L2 learners maintain contrasts between the VOTs of aspirated /p/ and unaspirated /p/ in /sp/ clusters?
3. Are VOT values influenced by learners' experience with English?

4. Do Arabic EFL learners show evidence of forming new phonetic categories for English stops, or do they assimilate them to existing L1 categories based on VOT distribution patterns?

By addressing these questions through a controlled production task and quantitative analysis, the study contributes to current understandings of phonetic acquisition in L2 speech and offers implications for both pedagogy and AI-based pronunciation feedback systems.

2. Theoretical Framework

This study adopts Flege's [10,15] Speech Learning Model (SLM) as its principal theoretical framework for interpreting the acquisition of English stop contrasts by Arabic-speaking learners. The SLM posits that second language (L2) phonetic learning is shaped by the perceived phonetic similarity between the learner's first language (L1) and the target L2 sounds. It emphasizes the role of experience and perception in shaping production patterns and accounts for both successful category formation and persistent cross-linguistic interference.

A key premise of the SLM is that the formation of a new L2 phonetic category depends on whether learners perceive the L2 sound as sufficiently different from any existing L1 category. If the perceived phonetic distance is great enough, learners may establish a new category; if not, the L2 sound is likely to be assimilated to an L1 category, leading to reduced accuracy in perception and production. This model allows for gradient development and variability in learner outcomes, depending on factors such as age of acquisition, amount of L2 input, and the nature of the contrast in question.

In the case of Arabic-speaking learners of English, the acquisition of VOT contrasts is particularly informative. Although Arabic includes a range of voiced and voiceless stop consonants, it does not use aspiration as a phonemic feature, and the voiceless bilabial plosive /p/ is absent from its phonemic inventory. Consequently, English voiceless aspirated stops may be assimilated into existing Arabic categories that lack a long-lag VOT distinction. This perceptual assimilation can hinder learners from identifying aspiration as a contrastive cue in English, particularly at lower proficiency levels. With increased exposure and targeted instruction, learners may begin to form new phonetic categories that reflect the aspirated–unaspirated distinction, but this development is gradual and not guaranteed.

The present study also considers the specific phonological context of /sp/ clusters, where English suppresses aspiration for voiceless stops. This context poses additional challenges for learners, as it requires the inhibition of a rule (aspiration of voiceless stops) that is otherwise consistently applied. The ability to suppress aspiration in these environments reflects more advanced phonological knowledge and finer-grained control over production.

While the Perceptual Assimilation Model for L2 learners (PAM-L2 [16]) offer complementary insights, especially for perception studies, the SLM remains better suited for analyzing L2 production data. The model's integration of perceptual and articulatory dimensions, its emphasis on category formation, and its compatibility with developmental phonetics research make it an appropriate framework for interpreting the VOT patterns observed in this study.

3. Methodology

3.1. Participants

The study included 87 participants divided into two main groups. The control group consisted of 29 monolingual native speakers of American English (15 males, 14 females), aged 18 to 23, all of whom were undergraduate students at a Midwestern university in the United States. The experimental group included 58 Arabic-speaking English as a Foreign Language (EFL) learners enrolled in undergraduate English language programs at a public university in Jordan. Proficiency levels were determined by the primary researcher, who has received extensive training in conducting and rating ACTFL Oral Proficiency Interviews (OPIs). Each learner completed a structured OPI, and their performance was assessed according to ACTFL proficiency guidelines. Based on these interviews, 30 learners were classified as Novice-High and 28 as Intermediate-High. This process ensured consistent and experience-based placement aligned with established standards for oral proficiency assessment. All participants reported normal hearing and no history of speech or lan-

guage disorders. All participants reported no known history of either speech or hearing impairment.

3.2. Materials and Stimuli

The speech stimuli were composed of 20 English words presented in a fixed carrier phrase (see **Appendix A**). These included ten minimal pairs that contrasted voiceless /p/ with voiced /b/ (e.g., peak vs. beak), matched for vowel quality (For more information about vowels in Arabic, see the reference [17].) and is designed to follow a simple CVC syllable structure. Some minimal pairs included low-frequency or nonsense words (e.g., peem, peeg) due to the absence of real English words that systematically contrast /p/ and /b/ across all vowel and coda environments. In such cases, phonotactically legal nonsense items were introduced to preserve phonological balance and allow for controlled comparisons. This approach ensured that VOT measurements reflected learners' phonetic encoding of the target contrast rather than their familiarity with lexical items, and it aligns with established practices in L2 phonetics research where nonce forms are used to elicit segmental production in a controlled manner. The second set of ten items consisted of naturally occurring English words containing /sp/ clusters (e.g., speak, speech, speedy), which were included to examine learners' ability to suppress aspiration in complex onset environments. Syllable structure was not controlled in this set, as only real words were used, resulting in natural variation in syllable length and coda complexity; however, the consistent /sp/ onset across items allowed for systematic investigation of aspiration suppression in this phonological context.

All words were embedded in the fixed carrier sentence "Say ___ once!", allowing consistent prosodic context across participants. All minimal pair items contrasting /p/ and /b/ (e.g., peak–beak, peep–beep, peach–beach) were designed to follow a simple CVC structure, with a single onset consonant, a monophthongal vowel nucleus, and a simple coda. This phonological consistency minimized variability in syllable complexity, ensuring that VOT differences were attributable to stop voicing rather than structural differences. Phonotactically legal nonsense words such as peem and peeg were included to complete the set of contrasts, as English lacks real minimal pairs for certain vowel-consonant combinations involving /p/ and /b/. In contrast, syllable structure was not controlled in the /sp/ cluster word list, which consisted of naturally occurring English words (e.g., speak, speech, speedboat, speedy). These items varied in syllable length and coda structure, reflecting the constraints of using real lexical forms in this condition. While this variability may introduce minor differences in segmental context, the cluster onset /sp/ remained consistent across all items and served as the primary environment of interest for examining aspiration suppression.

3.3. Procedure

Recordings were conducted in quiet, supervised settings with minimal background noise. All participants were recorded using high-quality microphones with a minimum sampling rate of 44.1 kHz. Each participant read the full randomized stimulus list, with each word repeated three times. They were instructed to speak at a natural pace and avoid hyper-articulation. Each word was repeated three times by the participants. VOT was measured for all three repetitions, and the mean value was used as the representative measure for each condition. This approach reduced the influence of token-level variability and ensured a more stable estimate of individual production patterns. All annotations followed a standardized segmentation protocol using Praat.

3.4. Acoustic Analysis

VOT measurements were performed using Praat [18], version 6.4.27). VOT was defined as the interval in milliseconds between the release burst of the stop consonant and the onset of periodic voicing for the following vowel. For each stop, VOT was measured as the period between the beginning of the release burst of the stop and the onset of the glottal vibration [3]. Vowel onset was defined as the apparent emergence of F1; vowel offset was taken as the point at which F2 substantially weakened or disappeared from the spectrogram. VOT was the time between when a plosive is released and when periodicity begins [19,20]. The release burst was identified as a sudden

increase in aperiodic energy in the waveform, while the onset of voicing was marked by the appearance of the first regular pitch periods and the emergence of voicing striations in the spectrogram. All tokens were segmented manually using both waveform and spectrogram views in Praat at a sampling rate of 44.1 kHz. Annotators followed a standardized protocol to ensure consistency across measurements. Manual segmentation was carried out by trained annotators following a standardized protocol. To assess reliability, 10% of the recordings were independently annotated by a second rater who was blind to group assignment. Inter-rater reliability was calculated using the Intraclass Correlation Coefficient (ICC), yielding strong agreement.

3.5. Statistical Analysis

All statistical analyses were conducted in Python using the statsmodels package. A linear mixed effects model was fitted with VOT (in milliseconds) as the dependent variable. Fixed effects included Group (Control, Intermediate-High, Novice-High), Condition (Voiced /b/, Voiceless /p/, /sp/ Cluster), and their interaction. Random intercepts were specified for participants and word items to account for repeated measures and subject-level variability. While the inclusion of random slopes for within-subject predictors such as Condition is often recommended, the statsmodels package in Python does not currently support fully crossed random slope structures in the same way as R's lme4. Therefore, we adopted a parsimonious and stable model specification appropriate for the capabilities of the Python-based environment and consistent with prior L2 phonetics research. No outliers were removed, as all VOT values fell within expected ranges for each group and condition. All reported effects are accompanied by standard errors, p-values, and 95% confidence intervals. Effect sizes were also calculated and interpreted where appropriate.

4. Results

This section presents the findings of both descriptive and inferential analyses that examine Voice Onset Time (VOT) production across three stop conditions, i.e., voiceless /p/, voiced /b/, and /p/ in /sp/ clusters across three groups: native English speakers (Control), Intermediate-High EFL learners, and Novice-High EFL learners.

4.1. Descriptive Statistics

Descriptive statistics were calculated for each Group × Condition combination to examine general patterns in VOT production. As shown in **Table 1**, native English speakers produced long-lag VOT for voiceless /p/ (M = 90.00 ms), short-lag VOT for /b/ (M = 13.83 ms), and appropriately short VOT for /p/ in /sp/ clusters (M = 19.21 ms). Intermediate-High learners showed intermediate values, suggesting developing phonetic categories, while Novice-High learners exhibited overlap between /p/ and /b/, indicating incomplete acquisition of the voicing contrast.

Table 1. Descriptive Statistics by Group and Condition.

Group	Condition	Mean VOT (ms)	SD	Min	Max	N
Control	/sp/ Cluster	19.207	5.778	10	28	29
Control	Voiced /b/	13.828	5.751	5	25	29
Control	Voiceless /p/	90.0	15.09	69	120	29
Intermediate-High	/sp/ Cluster	18.607	5.633	8	26	28
Intermediate-High	Voiced /b/	13.179	5.099	4	23	28
Intermediate-High	Voiceless /p/	77.321	12.769	56	97	28
Novice-High	/sp/ Cluster	8.6	2.686	5	13	30
Novice-High	Voiced /b/	8.267	4.323	0	15	30
Novice-High	Voiceless /p/	9.1	2.857	5	13	30

Visual inspection of VOT distributions further illustrates these trends. As depicted in the boxplot in **Figure 1**, native speakers show distinct VOT bands for each condition. Intermediate-High learners display partial separation, particularly between /p/ and /b/, while Novice-High learners show substantial overlap between categories and a wider spread in /sp/ cluster tokens, suggesting variability in the application of aspiration suppression rules.

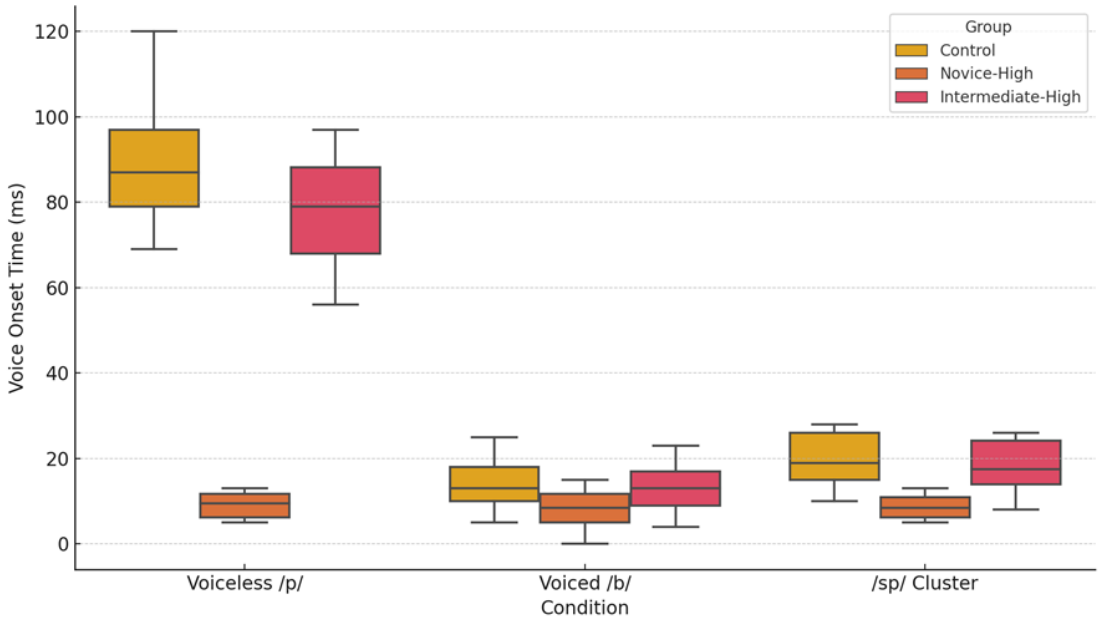

Figure 1. VOT Distribution by Condition and Group.

4.2. Inferential Statistics

To statistically test the influence of Group and Condition on VOT, a linear mixed effects model was fitted. The model included fixed effects for Group, Condition, and their interaction, with random intercepts for Participant ID to account for repeated measures.

As summarized in **Table 2**, there was a significant main effect of Condition. Voiceless /p/ produced substantially higher VOT values than voiced /b/ and /sp/ clusters. There was also a significant main effect of Group, with Novice-High learners producing significantly lower VOTs overall compared to native speakers. The Intermediate-High group did not differ significantly in overall VOT from the Control group, but showed notable interaction effects with Condition.

The Group × Condition interaction was statistically significant, indicating that group-based differences varied by phonetic context. As shown in **Table 3**, Novice-High learners did not produce a reliable VOT distinction between /p/ and /b/, and also showed a tendency to over-aspirate /p/ in /sp/ clusters (marginally significant, $p = 0.072$). Intermediate-High learners displayed more native-like behavior in the /sp/ context and showed significant separation between /p/ and /b/, although VOT values for /p/ remained lower than those of native speakers.

Table 2. Main Effects from Linear Mixed Effects Model.

Effect	Coefficient	Std. Error	*p*	CI Lower	CI Upper
Intercept	13.828	1.416	0.0	11.051	16.604
Group[T.Intermediate-High]	-0.649	2.021	0.748	-4.61	3.312
Group[T.Novice-High]	-5.561	1.986	0.005	-9.454	-1.668
Condition [T./sp/ Cluster]	5.379	2.003	0.007	1.453	9.305
Condition [T.Voiceless /p/]	76.172	2.003	0.0	72.246	80.098
Group Var	0.0	0.062	1.0	-0.122	0.122

Table 3. Interaction Effects from Linear Mixed Effects Model.

Effect	Coefficient	Std. Error	*p*	CI Lower	CI Upper
Group[T.Intermediate-High]: Condition [T./sp/ Cluster]	0.049	2.858	0.986	-5.552	5.651
Group[T.Novice-High]: Condition [T./sp/ Cluster]	-5.046	2.809	0.072	-10.552	0.46
Group[T.Intermediate-High]: Condition [T.Voiceless /p/]	-12.03	2.858	0.0	-17.631	-6.428
Group[T.Novice-High]: Condition [T.Voiceless /p/]	-75.339	2.809	0.0	-80.845	-69.833

Taken together, these results suggest a clear developmental trajectory. Novice learners appear to assimilate L2 stops to L1 categories, while intermediate learners begin to form distinct phonetic categories, consistent with the predictions of the Speech Learning Model [10,15].

5. Discussion

This study examined the production of Voice Onset Time (VOT) in English stop consonants by native English speakers and Arabic-speaking learners at two proficiency levels. By analyzing the VOT values of voiced /b/, voiceless /p/, and unaspirated voiceless /p/ in /sp/ clusters, the study aimed to assess learners' acquisition of English stop contrasts and to evaluate whether these patterns reflected assimilation to L1 categories or the emergence of new phonetic categories. The results support the predictions of Flege's Speech Learning Model (SLM), which posits that the perceived similarity between L1 and L2 sounds governs the likelihood of category formation in L2 phonetic acquisition.

5.1. Development of Phonetic Categories

The Control group produced VOT values consistent with established norms for native English speakers: long-lag VOT for voiceless /p/, short-lag for voiced /b/, and short VOT for /p/ in /sp/ clusters. These contrasts were robust and clearly separated across conditions. In contrast, the Novice-High group exhibited substantial overlap between /p/ and /b/, and showed limited evidence of modulating aspiration in /sp/ clusters. These patterns suggest that Novice-High learners had not yet formed distinct L2 phonetic categories for voiceless aspirated stops and were instead assimilating English /p/ into an L1-like unaspirated stop category.

Interestingly, the Novice-High group produced highly similar VOT values across all three stop contexts, with average durations clustered around 9 milliseconds. This overlap likely reflects the learners' reliance on a single L1-influenced stop category with short-lag voicing, into which both English /p/ and /b/ are assimilated. While such convergence is consistent with the predictions of the Speech Learning Model, the near-total absence of VOT differentiation may also indicate reduced phonetic sensitivity or limited awareness of aspiration as a contrastive feature. It is also possible that the controlled reading task masked subtle articulatory distinctions, highlighting the need for complementary perception data in future work.

Intermediate-High learners, by contrast, demonstrated more target-like production. Their VOT values for /p/ were significantly longer than those for /b/, indicating that they had begun to develop a productive distinction between these two categories. Although their VOT values for /p/ remained lower than those of native speakers, the reduced overlap and greater consistency suggest emerging category formation. Because VOT values were averaged across three repetitions, the distributions presented reflect central production tendencies rather than token-level fluctuation. This approach minimizes within-speaker noise and enhances the reliability of group-level comparisons. However, it may also obscure momentary instability in learners' productions, particularly at lower proficiency levels where category boundaries may not yet be stable. These developmental differences align with SLM's claim that phonetic categories emerge gradually through increased L2 experience and exposure.

5.2. Effects of Phonological Context

The study also highlights the challenge posed by phonological contexts such as /sp/ clusters. While native speakers correctly suppressed aspiration in these environments, Novice-High learners tended to over-aspirate, and even Intermediate-High learners displayed slightly elevated VOT values compared to the native norm. These findings suggest that L2 learners may overgeneralize the aspiration rule, applying it in contexts where English phonology requires its suppression. This difficulty in context-sensitive application of phonetic rules reflects a broader issue in L2 phonological acquisition and emphasizes the need for instruction that targets not only segmental contrasts but also phonotactic constraints.

5.3. Alignment with the Speech Learning Model

The observed group differences are consistent with SLM's core assumptions. Novice-High learners' overlapping VOT distributions reflect the assimilation of L2 categories into existing L1 categories, due to insufficient

perceived phonetic distance. Intermediate learners' more native-like patterns suggest the beginning of new category formation, as increased experience enables learners to distinguish L2 sounds from L1 equivalents. The model's emphasis on perceptual similarity, experience, and gradual restructuring provides a robust explanatory framework for interpreting these developmental trajectories.

5.4. Implications for Instruction and Technology

These findings have clear implications for pronunciation instruction. Teachers working with Arabic-speaking learners should highlight the role of aspiration in English voicing contrasts and explicitly address its absence in /sp/ clusters. Instruction should include perceptual training, visual feedback such as spectrograms or VOT measurements, and structured practice to improve the timing of stop consonant production. Activities like delayed imitation, spectrogram reading, and minimal pair drills can help learners perceive and produce voicing contrasts more accurately, particularly when their first language does not use aspiration as a contrastive feature.

The results also draw attention to limitations in many current AI-based pronunciation tools, which often overlook fine-grained temporal features such as voice onset time. Popular platforms, including ELSA Speak, Duolingo, and SpeechAce, tend to emphasize broader pronunciation elements such as stress placement or general intelligibility. However, these tools frequently lack the acoustic resolution needed to identify critical timing-based distinctions between aspirated and unaspirated stops. Consequently, learners may receive feedback that is either too general or insufficiently accurate to address specific phonetic challenges tied to their first language background.

Recent developments in AI-enhanced computer-assisted language learning have highlighted the value of personalized, data-informed feedback. Research by Li [8] emphasizes that effective AI-driven pronunciation support must account for precise acoustic cues, not only overall speech quality. Similarly, Levis J.M. and Moyer A. [21] note that subsegmental features like voice onset time are central to intelligibility and accentedness, yet are commonly overlooked in commercial systems.

Improving this situation requires training AI models on annotated learner speech that reflects cross-linguistic influences. Voice onset time can be extracted using forced alignment techniques that match audio with phonetic transcriptions, or through end-to-end machine learning methods such as convolutional neural networks trained on spectrograms or waveform data. These approaches require high-quality, well-labeled training sets. The dataset developed in this study, which contains Arabic-accented English annotated for VOT, offers a valuable resource for such applications.

Incorporating learner-specific data into model training would allow pronunciation tools to detect and respond to aspiration patterns with greater accuracy. For example, systems could provide targeted feedback when a learner over-aspirates /p/ in a cluster environment or fails to distinguish between /p/ and /b/. Feedback based on measurable acoustic contrasts can guide learners toward more accurate production and help instructors monitor phonetic development over time.

Ultimately, aligning phonetic research with AI-assisted instruction requires both carefully annotated speech data and model designs that are sensitive to the specific needs of learners. By contributing detailed, learner-informed acoustic data, studies like the present one help advance the development of AI pronunciation tools that are not only technologically robust but also linguistically and pedagogically meaningful.

6. Conclusions and Pedagogical Implications

From a pedagogical perspective, the findings of this study underscore the importance of integrating explicit pronunciation instruction into EFL curricula [22–24], especially for learners whose L1 lacks aspiration as a phonemic feature. Teachers should prioritize the development of learners' awareness of aspiration contrasts through multimodal approaches that combine auditory, articulatory, and visual cues. For instance, spectrogram-reading activities in software like Praat or user-friendly acoustic analysis tools can help learners visualize aspiration as a burst of aperiodic energy followed by voicing. Structured repetition tasks, delayed imitation, and minimal pair drills can be used to reinforce accurate timing in voiceless stop production.

In addition to segment-level practice, classroom instruction should address phonological context effects, such as the suppression of aspiration in /sp/ clusters. Learners may benefit from explicit instruction that contrasts single-word contexts with cluster environments, helping them identify where aspiration is expected or inhibited. Given the findings that even Intermediate-High learners displayed partial overgeneralization of aspiration, instructional strategies must go beyond isolated word production and incorporate contextualized practice, such as sentence reading or controlled dialogues.

Finally, teacher training programs should provide language instructors with foundational knowledge in acoustic phonetics and tools for pronunciation assessment. Equipping instructors with the ability to analyze learner speech at the segmental and subsegmental levels would enable more targeted, informed intervention. Such training is particularly essential in contexts where pronunciation is often underrepresented in formal language instruction.

This study investigated the production of Voice Onset Time (VOT) in English stop consonants by native English speakers and Arabic-speaking English as a Foreign Language (EFL) learners at two proficiency levels. Through acoustic analysis of voiced /b/, voiceless /p/, and unaspirated /p/ in /sp/ clusters, the study assessed learners' ability to distinguish phonetic categories that are not contrastive in Arabic. The findings confirm that native speakers maintain robust VOT distinctions across contexts, while Novice-High learners show considerable overlap between categories, and Intermediate-High learners demonstrate more target-like, though still variable, patterns.

These results support Flege's Speech Learning Model, which emphasizes the role of perceived L1–L2 similarity in determining whether learners form new phonetic categories or assimilate L2 sounds into existing ones. Novice learners' VOT patterns reflect assimilation and limited phonetic differentiation, while the more distinct and consistent VOTs produced by intermediate learners point to emergent category formation. These developmental trajectories underscore the importance of structured exposure and meaningful phonetic practice in second language instruction.

From a pedagogical perspective, the study highlights the need for explicit instruction on VOT contrasts and phonological rules governing aspiration in English. Teachers should guide learners in distinguishing between voiced and voiceless stops not only through articulatory descriptions but also through perceptual training and visual feedback. Contextual variability, such as the suppression of aspiration in /sp/ clusters, should be addressed through focused practice and awareness-raising activities.

The study also carries implications for AI-enhanced language learning. Current speech recognition tools are limited in their ability to detect subtle, millisecond-level timing differences such as those that define VOT. To be truly effective in second language phonetics training, these tools must evolve to incorporate fine-grained acoustic sensitivity. Annotated datasets such as the one developed in this study may serve as valuable training corpora for future systems, bridging the gap between technological capability and phonetic precision.

Ultimately, this study contributes to a growing body of research emphasizing the complexity of L2 phonetic development and the need for pedagogical and technological tools that are grounded in linguistic theory and acoustic reality. Future research should explore longitudinal changes in VOT production, integrate perception-based measures, and examine how AI-based interventions can support both learners and instructors in navigating the challenges of second language pronunciation.

Author Contributions

Conceptualization, H.A.-D. and M.A.; methodology, M.A. and M.A.-D.; software, M.A. and M.A.-D.; validation, H.A.-D., N.A. and R.A.; formal analysis, H.A.-D.; investigation, M.A. and M.A.-D.; resources, M.A. and M.A.-D.; data curation, R.A.; writing—original draft preparation, R.A.; writing—review and editing, M.A. and M.A.-D.; visualization, H.A.-D.; supervision, M.A. and M.A.-D.; project administration, M.A. and M.A.-D. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board

Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the University of Jordan.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

Data is available upon request.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix A

Stimulus Words Used in the Production Task.

Nonsense words are marked with an asterisk (*).

Table A1. Minimal Pairs Contrasting Voiceless /p/ and Voiced /b/.

/p/ Word	/b/ Word
peak	beak
peat	beat
peep	beep
peem*	beam
peef*	beef
peer*	beer
Peeg*	beeg*
peach	beach
peed	bead
peet*	beat

Table A2. /sp/ Cluster Words.

/sp/ Cluster Word
speak
speech
speaker
speeding
speedy
speakeasy
speechless
speedway
speedboat
speaking

References

[1] Lisker, L., Abramson, A.S., 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word. 20(3), 384–422. DOI: https://doi.org/10.1080/00437956.1964.11659830

[2] Li, F., 2013. The effect of speakers' sex on voice onset time in Mandarin stops. The Journal of the Acoustical Society of America. 133(2), EL142-EL147. DOI: https://doi.org/10.1121/1.4778281

[3] Morris, R.J., McCrea, C.R., Herring, K.D., 2008. Voice onset time differences between adult males and females: Isolated syllables. Journal of Phonetics. 36, 308–317. DOI: https://doi.org/10.1016/j.wocn.2007.06.003

[4] Oh, E., 2019. Effects of gender on voice onset time and fundamental frequency cues in perception and production of English stops. Linguist Res. 36(1), 67–89. DOI: https://doi.org/10.17250/khisli.36.1.201903.003

[5] Cho, T., Ladefoged, P., 1999. Variation and universals in VOT: evidence from 18 languages. Journal of Phonetics. 27(2), 207–29. DOI: https://doi.org/10.1006/jpho.1999.0094

[6] Zampini, M.L., 1998. The relationship between the production and perception of L2 Spanish stops. Texas papers in foreign language education. 3(3), 85–100.

[7] Al-Deaibes, M., Alsharefeen, R., Rabab'ah, G., et al., 2025. Voice onset time in the Emirati Arabic dialect. Cogent Arts Humanit. 12(1), 2483604. DOI: https://doi.org/10.1080/23311983.2025.2483604

[8] Li, M., He, L., Luo, Y., et al., 2021. AI-based pronunciation feedback and L2 fluency: Advancing individualized instruction in computer-assisted language learning. Computer Assisted Language Learning. 34(3), 303–328.

[9] Boersma, P., Weenink, D., 2024. Praat: Doing phonetics by computer [Computer program]. Version 2.7. Available from: https://www.praat.org (cited 3 April 2025).

[10] Flege, J.E., 2020. The revised Speech Learning Model (SLM-r). In: Munro, M.J., Bohn, O.-S. (eds.). The Oxford Handbook of Language Phonology. Oxford University Press: Oxford, UK. pp. 3–24.

[11] Yeni-Komshian, G.H., Caramaza, A., Preston, M.S., 1977. A study of voicing in Lebanese Arabic. Journal of Phonetics. 5(1), 35–48. DOI: https://doi.org/10.1016/S0095-4470(19)31112-X

[12] Mitleb, F., 2001. Voice onset time of Jordanian Arabic stops. The Journal of the Acoustical Society of America. 109(5), 2474. DOI: https://doi.

org/10.1121/1.4744787

[13] Almbark, R., Bouchhioua, N., Hellmuth, S., 2014. Acquisition of English VOT by Syrian Arabic and French speakers. In Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS). Publisher: Glasgow, UK.

[14] Flege, J.E., 1995. Second language speech learning: Theory, findings, and problems. In: Strange W., (eds.). Speech perception and linguistic experience: Issues in cross-language research. York Press: Timonium, MD, USA. pp. 233–277.

[15] Al-Deaibes, M., Jarrah, M., 2023. Production of Arabic geminates by English speakers. J Psycholinguist Res. 52(6), 2877–2902. DOI: https://doi.org/10.1007/s10936-023-10025-w

[16] Best, C.T., Tyler, M.D., 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, O.-S., Munro, M.J. (eds.). Language experience in second language speech learning: In honor of James Emil Flege. John Benjamins: Amsterdam, Netherlands. pp. 13–34.

[17] Mashaqba, B., Huneety, A., Al-Khawaldeh, N., et al., 2023. Acoustics of long vowels in Arabic-speaking children with hearing impairments. Humanit Soc Sci Commun. 10(1), 1–11. DOI: https://doi.org/10.1057/s41599-023-01778-9

[18] Boersma, P., Weenink, D., 2024. Praat: doing phonetics by computer. Available from: https://www.fon.hum.uva.nl/praat/ (cited 2 April 2025).

[19] Aldamen, H., Al-Deaibes, M., 2023. Arabic emphatic consonants as produced by English speakers: An acoustic study. Heliyon. 9(2), 1–12. DOI: https://doi.org/10.1016/j.heliyon.2023.e13401

[20] Aldamen, H., Al-Deaibes, M., 2023. Perception and production of L2 Arabic emphatic consonants: The role of communicative and traditional form-based approaches. Ampersand. 10, 100105. DOI: https://doi.org/10.1016/j.amper.2022.100105

[21] Levis, J.M., Moyer, A., 2021. Intelligibility in Pronunciation: Research and Practice. Cambridge University Press: Cambridge, UK.

[22] Cho, T., Ladefoged, P., 1999. Variation and universals in VOT: Evidence from 18 languages. Journal of Phonetics. 27, 207–229. DOI: https://doi.org/10.1006/jpho.1999.0094

[23] Rabab'ah, G., Cheikh, M., Al-Deaibes, M., 2024. Unraveling Conversational Implicatures: A Study on Arabic EFL Learners. Open Cultural Studies. 8(1), 20240006. DOI: https://doi.org/10.1515/culture-2024-0006

[24] Zampini, M.L., 1998. The relationship between the production and perception of L2 Spanish stops. Tex Papers in Foreign Language Education. 3(3), 85–100.