 DOI: 10.35382/TVUJS.16.2.2026.164

ADDRESSING PRONUNCIATION PROBLEMS OF STOP CONSONANTS AMONG NON-ENGLISH MAJOR STUDENTS THROUGH AUDIO-VISUAL AIDS

Tran Thi Xuan Mai[1], Thach Son Le[1*], Nguyen Thi Phuong Nam[1]

Abstract – *This study investigated the effectiveness of audiovisual aids in addressing pronunciation difficulties among non-English majors in English stop consonants. Forty non-English major students at Tra Vinh University, Vietnam, participated in the research, which was conducted primarily from January to May 2025. Before the intervention, a questionnaire was administered to examine learners' perceptions of their ability to recognize and produce stop consonants, as well as their perceived difficulties in distinguishing minimal pairs. Then, the participants were asked to take a pronunciation test, whose results were compared with the posttest results, and to draw a conclusion about the intervention's effectiveness. An eight-week intervention using audio-visual materials was subsequently implemented. After that, a posttest was conducted to assess its effectiveness. The questionnaire findings revealed that /p/, /d/, and /k/ were perceived as more challenging than /b/, /t/, and /g/ in both recognition and production. In the pronunciation test, learners performed better at recognizing the sounds /p/, /b/, /d/, /k/, and /g/. They also produced the sounds /p/, /b/, /d/, and /g/ more accurately at the single-word level. Additionally, students showed fewer errors in producing /p/ at the sentence level. These results suggest that audio-visual aids can effectively enhance learners' recognition and production of some English stop consonants. Pedagogical implications, recommendations, and limitations are also presented to guide future research.*

Keywords: audio-visual aids, pronunciation of stop consonants, pronunciation problems, stop consonants.

I. INTRODUCTION

Pronunciation is an essential aspect of foreign language learning, as it directly affects learners' listening comprehension, speaking fluency, and confidence in communication [1]. Fraser [2] stated that when mispronounced, messages may be misunderstood, which causes breakdowns in communication. However, Levis [3] confirmed that although modern teaching methods emphasize communicative competence, pronunciation instruction has not received adequate attention in many language classrooms. Therefore, while focusing on communicative competence, pronunciation should also be regarded as an integral component of language learning.

English as a Foreign Language (EFL) learners in Vietnam often experience considerable difficulties in pronunciation, especially in producing consonant sounds [4]. They tend to omit consonants in clusters and substitute English sounds with similar Vietnamese sounds and delete final consonants [5, 6]. Among these errors, stop consonants such as /p/, /b/, /t/, /d/, /k/, and /g/ frequently occur. This indicates that stop sounds are among the most problematic areas for Vietnamese learners. Consequently, identifying an effective pedagogical intervention to help learners improve their pronunciation, particularly of stop consonants, is essential.

As Gilakjani et al. [7] pointed out, with the rapid development of educational technology, teachers today have a wide range of tools to enhance pronunciation teaching. Among these, audiovisual aids are considered highly effective

[1]Tra Vinh University, Vietnam
*Corresponding author: tsle@tvu.edu.vn
Received date: 16 March 2026; Revised date: 1 May 2026; Accepted date: 24 June 2026

due to their visual appeal, ease of use, and motivational potential. Supporting this view, Harmer [8] noted that audio-visual aids help learners visualize the relationship between articulatory movements and produced sounds, thereby supporting both recognition and imitation of pronunciation.

For these reasons, the present study employs audiovisual aids as a pedagogical intervention to improve the pronunciation of English stop consonants among non-English-major students at a university in Mekong Delta, Vietnam. This study aimed to address the following questions:

1. What are learners' perceptions of their ability to recognize and produce stop consonants before the intervention?

2. How does the application of audio-visual aids help students enhance their pronunciation of stop consonants?

II. LITERATURE REVIEW

A. Theoretical frameworks

Speech learning theory

The speech learning theory proposed by Flege [9] explored how a learner perceived and produced sounds in a second language (L2) and the influence of the first language (L1) on this process. Learners compared L2 sounds to their existing L1 phonetic categories. The comparison can either enhance or undermine pronunciation accuracy. Additionally, if a sound in L2 closely resembled a sound in L1, learners might perceive them as identical. This led to classification as equivalent and other pronunciation problems. Another key element of this theory was the emphasis on continuous exposure. It argued that practice in the L2 contributed to improved pronunciation over time.

In short, this theory could be applied to explain pronunciation problems that learners encounter based on their connection to their first language. Additionally, it created the belief that errors in second-language pronunciation could be corrected regardless of age.

Multimedia learning theory

Multimedia learning theory, proposed by Mayer [10], posits that learners acquire knowledge more effectively from a combination of words and images than from words alone. This framework is grounded in the assumption that the human brain processes information through two interconnected channels: the auditory and the visual. Effective learning can occur when both channels are engaged simultaneously in a balanced manner.

The theory is guided by three core principles. First of all, pairing verbal input with relevant visuals can facilitate learners' comprehension by enabling them to build richer mental models. Second, the limited capacity principle emphasizes that each channel has a finite processing capacity. In other words, instructional materials should focus on essential content and exclude extraneous details to avoid cognitive overload. Finally, the active processing principle refers to meaningful learning, which can occur when learners actively select, organize, and integrate new information with their prior knowledge. These principles provide a foundation for the use of audio-visual aids across educational contexts. Well-designed tools, such as narrated slide shows, animated demonstrations, or instructional videos, distribute cognitive demands across auditory and visual channels, thereby enhancing comprehension and retention.

Multisensory learning theory

According to Mercer et al. [11], multisensory learning theory originates from the multisensory approach, which posits that learning becomes more effective when information is presented through multiple sensory modalities such as visual, auditory, kinesthetic, and tactile channels. This approach, also known as VAKT (visual–auditory–kinesthetic–tactile), emphasizes engaging learners through a range of sensory experiences. Activities in this approach integrate sensory reinforcement techniques, enabling it to address a wide range of learners rather than a single group [12]. Zhan et al. [13] confirmed that empirical studies have demonstrated the role of

multisensory learning in assisting learners' cognitive load, which supports language learning. In particular, Hirata et al. [14] found that combining auditory and visual cues can improve phoneme learning in a second language, highlighting the value of multimodal input in pronunciation training.

B. Pronunciation definitions and assessment

In English, pronunciation is a crucial component of oral communication, referring to the production and perception of speech sounds [15]. There are two components of pronunciation: segmental features, including individual phonemes such as consonants and vowels, and suprasegmental features, such as stress and intonation.

From the above definitions, it is clear that assessing pronunciation ability involves two main components: recognition and production. Munro et al. [16] affirmed that listener rating is an assessment that involves subjective judgments by trained or untrained raters. It is used to evaluate the aspects of intelligibility, comprehensibility, and accentedness. Derwing et al. [17] acknowledged that trained raters, such as linguists or teachers, usually provide more consistent evaluations, whereas untrained raters, such as native speakers, judge pronunciation based on their everyday communication experience. As Kang et al. [18] pointed out, while this approach reflects real-world communication, factors such as rater bias, familiarity with accents, and linguistic background can affect the results. For this reason, researchers often combine listener assessments with objective measures, such as phonetic transcription or audio analysis, to provide a more complete evaluation.

C. Definitions and phonetic features of stop consonants

Stop consonants, or plosives, are fundamental sounds of English consonants, which include six sounds: /p/, /t/, /k/, /b/, /d/, and /g/. They are produced with complete closure of the vocal tract followed by a sudden release of air, making them essential to speech production across languages. Ladefoged et al. [19] confirmed that these sounds are characterized by three phonetic features: place of articulation, manner of articulation, and voicing. Place of articulation divides them into bilabial (/p/, /b/), alveolar (/t/, /d/), and velar (/k/, /g/) stops, each classified by the organs of speech involved. Manner of articulation describes the complete blockage and release of airflow. Voicing, or the vibration of the vocal cords, distinguishes stop consonants: voiced stops (/b, d, g/) involve vibration, while voiceless stops (/p, t, k/) do not.

Additionally, aspiration is a key phonetic feature of voiceless stop consonants. It is referred to as a burst of air that follows the release of the closure before the onset of vocal fold vibration. Ladefoged et al. [19] emphasized that voiceless stops (/p, t, k/) are typically aspirated when they occur at the beginning of a stressed syllable, for example, in the words pat [p^hæt], top [t^hɑp], or cat [k^hæt]. However, aspiration is either absent or reduced when these stops follow an /s/ sound. In terms of position in a word, Ladefoged et al. [19] also stressed that voiceless stops are strongly aspirated when they occur in the initial position. However, when these stops appear in the medial position, aspiration depends on stress patterns. In word-final position, aspiration tends to be weaker or absent. Particularly in casual speech, the stop may be unreleased or only lightly aspirated.

D. Definitions of audio-visual aids

Audio-visual aids are instructional materials that use both audio and visual components to enhance the learning process and improve comprehension [20]. Venkataiah [21] defined them as the teaching tools that combine auditory and visual stimuli, such as videos, charts, models, and multimedia presentations, to facilitate effective communication and learning. In this research, audio-visual aids refer to the instruments teachers use to support learners in combining sound and visual aids.

Shams et al. [22] confirmed that the use of audiovisual aids in language teaching and learning has become popular. It is consistent with

multisensory learning theory, which emphasizes engaging multiple sensory modalities to enhance students' learning and retention. Mayer [10] pointed out that when learners receive information through both auditory and visual channels, their comprehension and knowledge retention improve compared to single-mode instruction.

E. Related studies on solutions to pronunciation problems

The effectiveness of explicit instruction in developing pronunciation accuracy and intelligibility has been widely acknowledged. For instance, Naeini et al. [23] compared the effects of explicit versus implicit pronunciation instruction among Iranian EFL learners and found that the explicit group significantly outperformed the implicit group in terms of accuracy and fluency. Their findings highlight that direct articulatory guidance and awareness-raising activities play a crucial role in improving learners' intelligibility. Furthermore, Linebaugh et al. [24] reported that explicit articulatory training targeting the production of the problematic sounds /p/ and /b/ enhanced Arab learners' perceptual discrimination between these sounds. This result reinforces the interconnection between perception and production, suggesting that articulatory awareness can improve both auditory discrimination and oral performance.

Audio-visual aids have increasingly been recognized as effective tools for pronunciation instruction because they allow learners to see and hear speech simultaneously, thereby reinforcing perception and production. Their contribution is highlighted in the increase of phonetic awareness, articulation, and intelligibility. Naasse et al. [25] emphasized that audio-visual aids foster learners' ability to recognize and accurately produce sounds, while also strengthening their overall phonetic competence.

The effectiveness of audio-visual aids has demonstrated considerable potential for supporting pronunciation development. Gamboa et al. [26] investigated a variety of audiovisual materials, including videos, songs, and computer-assisted programs, and found positive effects on ESL learners' phonetic accuracy and speech perception. Turgunboev [27] reported that authentic audio-visual materials significantly improved EFL learners' pronunciation accuracy in terms of segmental and suprasegmental features while also increasing their speaking confidence. More technology-driven approaches have also emerged. Bui et al. [28] developed Pteacher, a computer-aided pronunciation training system that provides exaggerated audiovisual corrective feedback, resulting in significant gains in learners' pronunciation accuracy. Moreover, Zhi et al. [29] showed that multimodal input, particularly visual cues from speakers' faces, facilitated phonetic learning by sharpening learners' ability to detect acoustic differences.

Audio-visual training improves both perception and production, especially when learners observe mouth movements and facial cues alongside auditory input [30]. Li [31] provided evidence that audio-visual high-variability phonetic training leads to notable improvement in the perception and production of difficult sounds. In a comprehensive review, Toyama et al. [32] concluded that multimodal approaches are generally effective for pronunciation instruction, although some segmental sounds remain resistant to improvement.

III. RESEARCH METHODOLOGY

A. Research design

This study used an action research design because it focused on a realistic classroom issue of pronunciation and aimed to find ways to address the problems through pedagogical intervention. Action research is often considered suitable for teachers, as it allows them to examine their own practice and make changes to support students' learning in authentic settings [33, 34].

The study followed a cyclical action research process, including planning, action, observation, and reflection [34]. In the planning phase, learners' difficulties and perceptions of their ability to recognize and produce stop consonants were identified by using a questionnaire and

a pronunciation test. This phase directly addressed research question 1, which aimed to examine learners' perceptions of their pronunciation problems before the intervention. During the action phase, the audio-visual intervention was implemented over eight sessions. The observation phase consisted of monitoring learners' participation and performance through classroom observations. Also in this phase, a pronunciation test was administered after the intervention. Finally, in the reflection phase, the results of pre- and post-pronunciation tests were compared to evaluate the effectiveness of the intervention. This process provided the basis for answering research question 2, which aimed to examine how the audio-visual intervention helped learners to address their pronunciation problems.

B. Research participants

This study involved 40 non-English-major students (35 males and 5 females) enrolled in the General English 2 course during the second semester of the 2024–2025 academic year at Tra Vinh University, Vietnam. According to the Sport Management curriculum, students are required to complete General English courses from Levels 1 to 4 [35]. Based on the university's conversion regulations, General English 2 is considered equivalent to the A2 level of the Common European Framework of Reference for Languages (CEFR) [36]. Therefore, the participants' English proficiency level can be considered A2. At the time of the study, the students were freshmen and had studied English for approximately 10 years, including three years in primary school, four years in lower secondary school, and three years in upper secondary school.

C. Research instruments

This study employed two research instruments. The first instrument was a questionnaire designed to assess learners' self-perceptions of their ability to recognize and produce English stop consonants. The second instrument was the pronunciation test, which provided an objective measure of their actual recognition and production performance and aimed to examine the effectiveness of audio-visual aids. A summary of the instruments is presented in Table 1.

Table 1: Summary of research instruments

Instruments	Purposes	Structure	Number of items	Scoring methods
Questionnaire	To examine students' self-perceptions of recognition and production difficulties	Ability to recognize and produce single sounds	54 items	5-point Likert scale
		Problems with differentiating minimal pairs		
Pronunciation test	To diagnose students' actual recognition and production difficulties, and track improvement after intervention	Part 1: Recognition (minimal pairs, short conversations, single words)	51 recognition items	Dichotomous (0/1) for recognition & word-level
		Part 2: Production (single words and sentences)	30 production words	Error counting

Questionnaire

This study employed a questionnaire adapted from Jahara [37]. It was divided into two main sections. Section 1 focused on students' self-reported recognition and production of single sounds across the three positions. The items were written in a positive tone (e.g., 'I can recognize the sound /p/ correctly in the initial position'). In this section, there were six single sounds. Each item was used to assess learners' perceptions and abilities to recognize and produce every sound in three positions within words. Therefore, each sound had six items. In total, the first part consisted of 36 items.

In contrast, section 2 focused on minimal pairs and was phrased negatively (e.g., 'I often get confused when hearing the sounds /p/ and /b/ at the beginning of words'). Because there were three minimal pairs, which aimed to assess at all three positions by recognizing and producing, each pair had six items. Totally, there were 14 items in section 2. In general, the revised questionnaire comprised 54 items and was administered to 40 participants.

Pronunciation test

The test consisted of a recognition section and a production section. For sound recognition, three tasks were included. First, students discriminated minimal pairs in the initial position (e.g., pat vs. bat). Second, minimal pairs were embedded in

a short conversation to examine recognition in context. Third, students listened to single words containing the target stop consonants in initial, medial, and final positions and identified whether each sound they heard was correct.

In the sound production section, two levels were assessed, including word and sentence. At the single-word level, students worked in pairs. One student read aloud a list of words, and the partner judged whether each pronunciation was correct or incorrect; the researchers verified the responses. At the sentence level, students were shown pictures and asked to answer five questions that were posed by the teacher. To ensure familiarity with the content, the pictures were based on eight topics covered in the course: jobs, sports, health, entertainment, places, appearance, transportation, and books. Besides, the questions were designed to elicit sentences containing all six target consonants (/p, b, t, d, k, g/). For example, when shown a picture of a girl drinking water, the teacher might ask, 'What is the girl doing?' to elicit responses in full sentences, such as 'The girl is drinking water.' This ensured that the target sounds would appear in natural speech.

D. Intervention design

The intervention was designed to improve students' pronunciation of English stop consonants through contextualized practice supported by audio-visual aids. It followed the sequence, a framework widely used in communicative language teaching to introduce new language, provide guided practice, and extend use in meaningful communication [8, 15]. This design reflects the view that pronunciation taught in context is more effective for long-term retention and communicative transfer than isolated drilling [7].

The themes are aligned with the course content for General English 2. Since the intervention was implemented at the end of the semester, when the participants had completed most of the units, eight random topics were chosen as the main themes for the eight lessons. The selected topics contained words with minimal pairs, which facilitated the researchers' design of the contexts. A summary of the intervention schedule is shown in Table 2.

Table 2: Overview of intervention lessons

Lessons	Focused sounds	Themes	Week of semester	Unit in course book
1	/p/ and /b/	Jobs	Week 1	Unit 3
2	/t/ and /d/	Sports	Week 1	Unit 1
3	/k/ and /g/	Health	Week 2	Unit 9
4	/p, t, k/	Entertainment	Week 2	Unit 4
5	/b, d, g/	Places	Week 3	Unit 7
6	/p/ and /b/	Appearance	Week 3	Unit 2
7	/t/ and /d/	Transportation	Week 4	Unit 5
8	/k/ and /g/	Books	Week 4	Unit 12

To ensure consistency and pedagogical coherence across the sessions, a general teaching scenario was designed and applied to all lessons (see Table 3). Each lesson followed three main stages: presentation, practice, and production, with audio-visual aids integrated throughout.

Table 3: Teaching scenario

Stage	Time	Procedure
Presentation	20 minutes	The teacher introduces target sounds using phonetic videos with slowed speed for clarity. Students then practice pronouncing target words using a speech recognition tool (e.g., Speechnotes) until the software confirms accurate articulation.
Practice	25 minutes	*Recognition task:* Students listen to recordings and complete fill-in-the-blank exercises with minimal pairs (e.g., rob vs. rope). The teacher wrote the script and used the web OpenAI.fm (https://www.openai.fm/) to generate the audio, an interactive demonstration of OpenAI's GPT-4o mini text-to-speech model. It was selected because of its power of steerability, which allowed multi-faceted control over the audio output [38]. *Production task:* Students perform role-plays in pairs, record their dialogues, and provide peer feedback based on teacher-given criteria focusing on target sounds.
Production	5 minutes	Students work in pairs to voice over for the characters. Each student says two sentences containing the target sounds, records them, and uploads the recordings to the learning management system (LMS). The teacher provides individualized feedback.

IV. RESULTS AND DISCUSSION

A. Results

Students' recognition of single sounds across positions

A one-sample t-test with the benchmark value of 4 ('Agree') was used to examine students' self-perceptions of their ability to recognize the sounds (Table 4). The results showed that the three sounds, /p/ (M = 3.55, SD = 0.67, t(39) =

-4.23, p < .001), /d/ (M = 3.68, SD = 0.67, t(39) = -2.97, p = .01), and /k/ (M = 3.70, SD = 0.77, t(39) = -2.46, p = .02), were rated significantly below the benchmark. It can be said that students showed less confidence with those sounds. In contrast, /b/ (M = 3.99, SD = 0.66), /t/ (M = 4.01, SD = 0.59), and /g/ (M = 3.97, SD = 0.62) were not significantly different from value 4, suggesting that students generally believed they could handle these consonants. In short, learners perceived /p/, /d/, and /k/ as more difficult to recognize than /b/, /t/, and /g/.

Table 4: One-sample t-test test results of students' perceptions of their ability to recognize stop consonants compared to the test value of 4

Sound	*M*	*SD*	*t*	*df*	*P*	*Mean difference*
/p/	3.55	.67	-4.23	39	**<.001**	-0.45
/b/	3.99	.66	-.08	39	0.94	-0.01
/t/	4.01	.59	.09	39	0.93	0.01
/d/	3.68	.67	-2.97	39	**0.01**	-0.32
/k/	3.70	.77	-2.46	39	**0.02**	-0.30
/g/	3.97	.62	-.34	39	0.74	-0.03

Students' recognition of the sounds /p/, /d/, /k/ across three positions

As noted in the previous section, /p/, /d/, and /k/ were three sounds learners had lower confidence in recognizing. To delve deeper into the three sounds with which learners showed lower confidence, a one-sample t-test was conducted. The mean scores of each sound across the three positions were compared with the test value of 4 (Table 5).

Table 5: One-sample t-test results of students' perceptions of their ability to recognize /p/, /d/, and /k/ against the test value of 4

Sound	Position	*M*	*SD*	*t*	*df*	*P*	*Mean difference*
/p/	Initial	3.90	0.84	-0.75	39	.46	-0.10
	Medial	3.65	0.95	-2.33	39	**.03**	-0.35
	Final	3.10	1.19	-4.77	39	**<.001**	-0.90
/d/	**Initial**	3.48	1.04	-3.20	39	**<.001**	-0.53
	Medial	4.03	0.97	0.16	39	.87	0.03
	Final	3.55	1.01	-2.81	39	**.01**	-0.45
/k/	Initial	3.95	0.93	-0.34	39	.74	-0.05
	Medial	3.73	0.99	-1.76	39	.09	-0.28
	Final	3.43	1.06	-3.43	39	**<.001**	-0.58

As shown in Table 5, for the sound /p/, significant differences were found in the medial (M = 3.65, SD = 0.95, t(39) = -2.33, p = .03) and final positions (M = 3.10, SD = 1.19, t(39) = -4.77, p < .001). The final position had the lowest confidence rating, which revealed that learners found it most difficult to recognize /p/ in this position. For /d/, significant differences were observed in the initial (M = 3.48, SD = 1.04, t(39) = -3.20, p < .01) and final positions (M = 3.55, SD = 1.01, t(39) = -2.81, p = .01), suggesting that students felt less confident recognizing /d/ at the beginning and end of words. For /k/, only the final position (M = 3.43, SD = 1.06, t(39) = -3.43, p < .001) was significantly below the benchmark value. Thus, it can be concluded that the final position was the most problematic for the recognition of /k/. In brief, learners perceived greater difficulty in recognizing /p/, /d/, and /k/ in word-final position.

Students' recognition of minimal pairs

To explore students' perceived difficulties in distinguishing minimal pairs involving stop consonants, a one-sample t-test was conducted with a neutral reference value of 3.0, given the range of the mean scores. Because the questionnaire items were phrased negatively (e.g., 'I get confused between . . . '), mean scores that are higher than the test value of 3 will reflect greater perceived difficulty, whereas lower ones indicate fewer difficulties. The findings are presented in Table 6.

Table 6: One-sample t-test results of students' perceptions of their ability to recognize minimal pairs against the test value of 3

Sound	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
/p/–/b/	3.14	0.88	1.015	39	.32	.14
/k/–/g/	2.57	0.76	-3.533	39	<.001	-.43
/t/–/d/	2.51	0.88	-3.627	39	<.001	-.49

As illustrated in Table 6, the minimal pairs /t/–/d/ (M = 2.51, SD = 0.88, t(39) = –3.63, p < .001) and /k/–/g/ (M = 2.57, SD = 0.76, t(39) = –3.53, p < .001) were rated significantly below the neutral point. These results showed that learners generally did not perceive much difficulty in distinguishing these sounds. In contrast, the pair /p/–/b/ (M = 3.14, SD = 0.88, t(39)

= 1.02, p = .32) was not significantly different from the neutral value. In other words, learners were uncertain about their ability to discriminate between /p/ and /b/.

Given this uncertainty, a one-sample t-test with a test value of 3 was conducted for each position of the pair /p/–/b/ (Table 7). The results showed a significant difference in the word-final position (M = 3.45, SD = 1.06, t(39) = 2.69, p = .01), whereas no significant differences were found in the initial or medial positions. Learners expressed the above-average agreement with the statement 'I was confused when hearing the pair /p/–/b/ at the end of a word'. It can be said that this position caused the most difficulty for them.

Table 7: One-sample t-test results comparing students' perceptions of their recognition of /p/–/b/ across positions in words to the test value of 3

Position	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
Initial	2.95	1.06	-.30	39	.77	-.05
Medial	3.03	1.23	.13	39	.90	.03
Final	3.45	1.06	2.69	39	**.01**	.45

In conclusion, learners' self-perceptions highlighted notable variation in their ability to recognize the six target stop consonants. They felt less confident with /p/, /d/, and /k/, especially when these sounds occurred in word-final positions, while /b/, /t/, and /g/ were generally perceived as manageable. The analysis of minimal pairs reinforced the trend of the sound /p/. Although students reported minimal confusion with /t/–/d/ and /k/–/g/, they expressed uncertainty with /p/–/b/, especially at word ends. This points to particular challenges with /p/ in the final position. Overall, students demonstrated clear awareness of their strengths and weaknesses in sound recognition, with consistent difficulties in recognizing/p/, /d/, and /k/.

Students' perception of their production of single sounds across positions

A one-sample t-test with a test value of 4 was conducted to examine students' perceptions of their ability to produce the six target sounds. As shown in Table 8, three sounds /p/ (M = 3.53, SD = 0.66, t(39) = –4.46, p < .001), /d/ (M = 3.44, SD = 0.65, t(39) = –5.46, p < .001), and /k/ (M = 3.70, SD = 0.78, t(39) = –2.44, p = .02) were rated significantly below the benchmark value. This shows that learners reported lower confidence in producing these consonants. In contrast, /b/ (M = 3.99, SD = 0.50, t(39) = –0.11, p = .91), /t/ (M = 3.90, SD = 0.65, t(39) = –0.98, p = .33), and /g/ (M = 3.87, SD = 0.72, t(39) = –1.18, p = .25) did not differ significantly from the test value. These findings suggest that students generally believed they could produce the three sounds /b/, /d/, and /g/ successfully.

Table 8: One-sample t-test results for students' perceptions of their production of stop consonants compared to the test value of 4

Sound	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
/p/	3.53	0.66	-4.46	39	**<.001**	-.47
/b/	3.99	0.50	-.11	39	.91	-.01
/t/	3.90	0.65	-.98	39	.33	-.10
/d/	3.44	0.65	-5.46	39	**<.001**	-.56
/k/	3.70	0.78	-2.44	39	**.02**	-.30
/g/	3.87	0.72	-1.18	39	.25	-.13

In short, students perceived /p/, /d/, and /k/ as more difficult to produce, while pronouncing /b/, /t/, and /g/was viewed as more confident. This pattern is consistent with their perceptions of recognition.

Students' perception of the production of the sounds /p/, /d/, and /k/ Since students reported lower confidence in producing /p/, /d/, and /k/ than in the other target sounds, their perceptions of these three consonants were further examined across word positions as presented in Table 9. A one-sample t-test with a test value of 4 was performed for the initial, medial, and final positions of each sound.

For /p/, significant differences were observed in the medial (M = 3.70, SD = 0.91, t(39) = –2.08, p = .04) and final positions (M = 3.13, SD = 0.99, t(39) = –5.58, p < .001), but not in the initial position (M = 3.78, SD = 0.83, t(39) = –1.70, p = .10). These findings indicate that learners reported lower confidence when producing /p/ in the medial and final positions, with the greatest difficulty occurring in the final position.

Table 9: One-sample t-test results for students' perceptions of their production of /p/, /d/, and /k/ compared to the test value of 4

Sound	Position	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
/p/	Initial	3.78	.83	-1.7	39	.10	-.23
	Medial	3.70	.91	-2.08	39	.04	-.30
	Final	3.13	.99	-5.58	39	<.001	-.88
/d/	Initial	3.15	1.12	-4.79	39	<.001	-.85
	Medial	3.73	.99	-1.76	39	.09	-.28
	Final	3.45	.85	-4.11	39	<.001	-.55
/k/	Initial	4.10	.84	.75	39	.46	.10
	Medial	3.65	1.05	-2.11	39	.04	-.35
	Final	3.35	1.10	-3.74	39	<.001	-.65

For /d/, both the initial (M = 3.15, SD = 1.12, t(39) = –4.79, p < .001) and final positions (M = 3.45, SD = 0.85, t(39) = –4.11, p < .001) were rated significantly below the benchmark, while the medial position was not. In other words, learners felt less confident producing /d/ at the beginning and end of words.

For /k/, significant differences were found in the medial (M = 3.65, SD = 1.05, t(39) = –2.11, p = .04) and final positions (M = 3.35, SD = 1.10, t(39) = –3.74, p < .001), but not in the initial position (M = 4.10, SD = 0.84, t(39) = 0.75, p = .46). These results suggest that students expressed lower confidence when producing /k/ in medial and final contexts. Briefly, students perceived word-final positions as the most challenging for producing /p/, /d/, and /k/, with additional difficulties in the medial position for /p/ and /k/.

Students' perceptions of the production of minimal pairs

To further investigate students' perceptions of producing stop consonants in contrastive contexts, a one-sample t-test with a test value of 3 ('neutral') was conducted for three minimal pairs. Because of using the negative statements (e.g., 'I get confused when producing . . . '), mean scores significantly lower than 3 indicate fewer perceived difficulties, whereas scores not significantly different from 3 suggest uncertainty. The results are presented in Table 10.

Table 10 reveals that the pairs /p/–/b/ (M = 2.93, SD = 0.96, t(39) = –0.44, p = .66) and /k/–/g/ (M = 2.73, SD = 0.98, t(39) = –1.71, p = .10) did not differ significantly from the neutral value. These findings suggested that learners were generally uncertain about their ability to produce these pairs accurately. In contrast, the pair /t/–/d/ (M = 2.67, SD = 0.99, t(39) = –2.14, p = .04) was rated significantly below 3, suggesting that learners perceived fewer difficulties in producing them. In summary, students expressed greater confidence in producing /t/–/d/, while their perceptions of /p/–/b/ and /k/–/g/ remained inconclusive.

Table 10: One-sample t-test results for students' perceptions of their production of minimal pairs compared to the test value of 3

Sound	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
/p/–/b/	2.93	.96	-.44	39	.66	-.07
/t/- /d/	2.67	.99	-2.14	39	.04	-.33
/k/-/g/	2.73	.98	-1.71	39	.10	-.27

Another one-sample t-test with a test value of 3 was conducted to examine students' perceptions of producing the two sound pairs (/p/–/b/ and /k/–/g/) across three-word positions.

Table 11: One-sample t-test results for students' perceptions of their production of /p/–/b/ and /k/–/g/ compared to the test value of 3

Pair	Position	*M*	*SD*	*t*	*df*	*p*	*Mean difference*
/p/-/b/	Initial	2.73	1.15	-1.5	39	.14	-.28
	Medial	2.90	1.19	-.53	39	.60	-.10
	Final	3.18	1.08	1.02	39	.31	.18
/k/-/g/	Initial	2.60	1.22	-2.08	39	.04	-.40
	Medial	2.65	1.08	-2.06	39	.05	-.35
	Final	2.95	1.06	-.30	39	.77	-.05

As can be seen in Table 11, regarding /p/–/b/ pair, none of the three-word positions differed significantly from the test value: initial (M = 2.73, SD = 1.15), t(39) = -1.50, p = .14; medial (M = 2.90, SD = 1.19), t(39) = -0.53, p = .60; final (M = 3.18, SD = 1.08), t(39) = 1.02, p = .31. This indicates that learners reported a similar level of difficulty across initial, medial, and final positions. For /k/–/g/ pair, the results showed significant differences in the initial and medial positions: initial (M = 2.60, SD = 1.22), t(39) = -2.08, p = .04; medial (M = 2.65, SD = 1.08), t(39) = -2.06, p = .05. The final position did not differ significantly from the test value

(M = 2.95, SD = 1.06), t(39) = -0.30, p = .77. These results suggested that learners experienced greater difficulty distinguishing /k/ and /g/ in the final position than in initial and medial positions.

Overall, students perceived that /p/, /d/, and /k/ posed greater challenges to produce than /b/, /t/, and /g/. Word-final positions were generally the most challenging, particularly for /p/, /d/, and /k/, while medial positions added difficulty for /p/ and /k/. In terms of minimal pairs, learners felt most confident with /t/–/d/, showed consistent difficulty with /p/–/b/ across all positions, and experienced greater challenges with /k/–/g/ in final position. These findings suggest that both the type of consonant and its position within a word influence students' perceived production difficulties.

Students' recognition of sounds before and after the intervention

A paired-samples t-test was conducted to compare the mean scores for each target sound on a 10-point scale. The purpose was to determine whether the intervention led to momentous improvements in learners' sound recognition. The results in Table 12 revealed that among the six stop consonants, five sounds /p/, /b/, /d/, /k/, and /g/ showed significant improvement after the intervention, while significant changes were not found for /t/.

Table 12: Significant improvements in sound recognition after the intervention

Sounds	*M (Pretest)*	*M (Posttest)*	*Mean difference*	*SI*	*t*	*df*	*p*
/p/	6.69	7.64	-.95	2.21	-2.73	39	.01
/b/	7.43	8.41	-.97	1.80	-3.40	39	.002
/d/	5.97	6.81	-.84	2.50	-2.12	39	.04
/k/	6.70	7.72	-1.02	2.0	-3.22	39	.003
/g/	6.69	7.92	-1.22	1.64	-4.73	39	<.001

To examine which positions contributed to these improvements, a paired-samples t-test was performed for each sound across the three positions. Because the number of occurrences differed, scores were converted to a 10-point scale for comparability. The results are presented in Table 13.

Accordingly, a meaningful improvement was observed in the medial position of the /p/ sound, t(39) = -2.24, p = .03. Similarly, learners showed better recognition when /b/ appeared in the final position, t(39) = -2.50, p = .02. The sound /k/ also showed a momentous improvement in the initial position, t(39) = -2.45, p =.02. Meanwhile, /d/ showed noteworthy change in final position, t(39) = -4.93, p < .001. Finally, /g/ emerged with improvement in initial position, t(39) = -6.74, p < .001.

Table 13: Significant improvements by sound position after the intervention

Sound	Position	*M (Pretest)*	*M (Posttest)*	*Mean difference*	*SD*	*t*	*df*	*p*
/p/	Medial	5.75	7.38	–1.63	4.58	–2.24	39	.031
/b/	Final	7.50	9.0	–1.50	3.79	–2.50	39	.017
/k/	Initial	7.54	8.25	–0.71	1.83	–2.45	39	.019
/d/	Final	2.25	5.25	–3.00	3.85	–4.93	39	<.001
/g/	Initial	6.44	9.06	–2.63	2.46	–6.74	39	<.001

Summary of students' recognition of sounds before and after the intervention

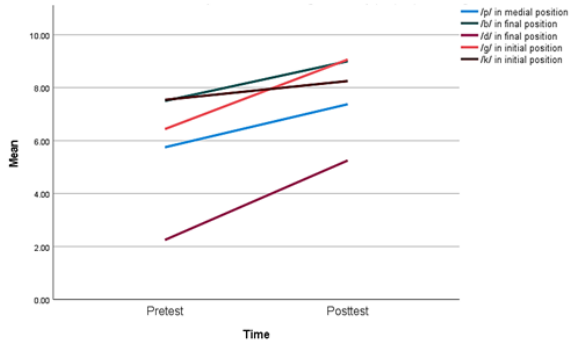

Fig. 1: Students' improvement in recognition of /p/, /b/, /d/, /k/, and /g/ in specific positions of words

Figure 1 illustrates students' improvement in recognizing the target stop consonants after the implementation of audio-visual aids. Overall, /p/, /b/, /d/, /g/, and /k/ showed higher mean scores, suggesting a positive impact of the intervention on learners' recognition of these sounds. Among the five target sounds, /d/ in the final position showed the most remarkable gain, rising from approximately 2.5 in the pretest to nearly 6.0 in the posttest. This substantial increase suggests that visual and auditory support helped learners

perceive voiced sounds in word-final positions more accurately. Besides, /g/ in the initial position and /b/ in the final position also displayed considerable improvement, both reaching mean scores above 8.5 in the posttest. In contrast, /p/ in the medial position and /k/ in the initial position showed more modest gains. However, they still reflected consistent upward trends. These results collectively indicate that the use of audio-visual materials effectively enhanced learners' ability to distinguish English stop consonants, particularly those that posed greater difficulty in the pretest.

Students' production of sounds before and after the intervention

In the word-level pronunciation task, learners were asked to read 30 words containing the target stop consonants. Although the number of occurrences was the same, the phoneme positions varied and were distributed unevenly. Therefore, all mean scores were normalized to a 10-point scale to ensure consistency.

A paired-samples t-test was conducted to compare the mean scores for each target sound before and after the intervention. The results presented in Table 14 show prominent changes.

Table 14: Significant improvements in word-level pronunciation after the intervention

Sounds	*M (Pretest)*	*M (Posttest)*	*Mean difference*	*SD*	*t*	*df*	*p*
/p/	5.55	6.95	-1.40	3.27	-2.71	39	0.01
/b/	8.0	8.75	-0.75	2.16	-2.20	39	0.03
/d/	3.05	4.25	-1.2	2.75	-2.76	39	0.01
/g/	6.20	7.35	-1.15	2.43	-2.89	39	0.01

The findings indicated improvements for four consonants: /p/, /b/, /d/, and /g/, while /k/ and /t/ did not show any remarkable change. Specifically, the learners' mean posttest score was significantly higher than the pretest score for the sound /p/ (Mdiff = –1.40, SD = 3.27, t(39) = –2.71, p = 0.01). Similarly, the sound /b/ also showed momentous improvement (Mdiff = –0.75, SD = 2.16, t(39) = –2.20, p = 0.03). For the sounds /d/ (Mdiff = –1.20, SD = 2.75, t(39) = –2.76, p = 0.01) and /g/ (Mdiff = –1.15, SD = 2.43, t(39) = –2.89, p = 0.01), the results also showed significant improvement posttest.

To further explore these improvements, another paired-samples t-test was conducted to identify the specific word positions at which learners performed better on the posttest than on the pretest. The positions where the sound changes significantly compared to the pretest are shown in Table 15. Learners pronounced /b/ significantly better in the initial position (M = -1.63, SD = 4.58, t(39) = -2.24, p = .03), /d/ more accurately in the initial position (M = -2.38, SD = 4.53, t(39) = -3.32, p = .002), and /g/ more accurately in the medial position (M = -1.25, SD = 3.35, t(39) = -2.36, p = .02). By contrast, no significant differences were found across the three positions of /p/, suggesting that the improvement was relatively even.

Summary of students' production of sounds before and after the intervention

As shown in Figure 2, students' mean scores in producing the target stop consonants increased notably from the pretest to the posttest. The upward trends in /p/, /b/, /d/, and /g/ demonstrate that the use of audio-visual aids contributed positively to learners' production accuracy. /d/ in the initial position showed the most remarkable improvement, rising from below 3.0 in the pretest to around 5.0 in the posttest. Likewise, /p/ in the medial and final positions, /b/ in the initial position, and /g/ in the medial position all showed consistent improvement, with posttest means ranging from approximately 7.0 to 9.5. The highest performance was observed in /p/ in the final position, which approached the maximum mean of 10, indicating that students were able to produce this voiceless stop with greater accuracy after the training. Overall, the pattern reveals that the audio-visual intervention was effective in enhancing learners' pronunciation of English stop consonants across different word positions.

Students' production of sound at the sentence level

In the sentence pronunciation section, learners are asked to look at a picture and then answer five questions from the teacher. Producing sounds accurately in sentences is often more challeng-

Table 15: Significant improvements by word position in pronunciation

Sounds	Positions	*M (Pretest)*	*M (Posttest)*	*Mean difference*	*SD*	*t*	*df*	*p*
/b/	Initial	6.25	7.88	-1.63	4.58	-2.24	39	**.03**
/d/	Initial	2.50	4.88	-2.38	4.53	-3.32	39	**.002**
/g/	Medial	6.38	7.63	-1.25	3.35	-2.36	39	**.02**

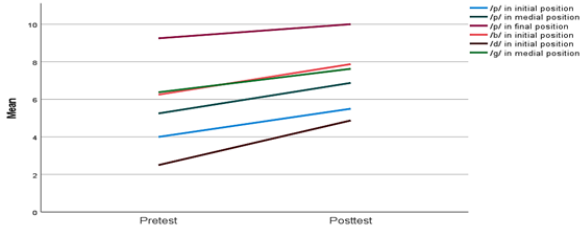

Fig. 2: Students' improvement in production of /p/, /b/, /d/, and /g/ in specific positions of words

ing than at the word level because connected speech involves additional features such as stress, rhythm, and coarticulation [15]. Therefore, an error-counting method was employed to more easily track learners' progress before and after the intervention. Each error was recorded as 1 point. The mean error scores for each sound are then compared before and after the intervention. A decrease in the mean score after the intervention means that the learner made fewer errors. A paired-samples t-test was used to determine whether the mean score differed significantly from the pretest.

Table 16: A paired-samples t-test results for pronunciation errors in target sounds before and after the intervention

Sound	*M (Pretest)*	*M (Posttest)*	*SD*	*t*	*df*	*p*
/p/	0.78	0.48	0.88	2.15	39	**.04**
/b/	0.10	0.13	0.48	-.33	39	.74
/t/	0.15	0.20	0.55	-.57	39	.57
/d/	0.28	0.20	0.53	.90	39	.37
/k/	0.45	0.55	0.84	-.75	39	.46
/g/	0.13	0.20	0.47	-1.00	39	.32

The results displayed in Table 16 indicate that only the sound /p/ demonstrated a statistically significant reduction in errors, with students making fewer mistakes in the posttest (M = 0.48) than in the pretest (M = 0.78). No significant differences were observed for the other sounds. In short, the intervention improved learners' pronunciation of the /p/ sound, while no meaningful changes were observed in the other target sounds.

Summary of students' production of sounds in sentences

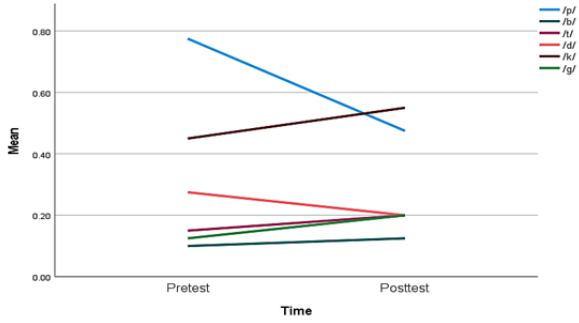

Fig. 3: Mean score of students' errors in pronunciation at the sentence level

Figure 3 illustrates the changes in students' production errors for six English stop consonants at the sentence level before and after the intervention. Overall, the mean error rates decreased slightly, suggesting a modest improvement in learners' ability to produce these consonants accurately in connected speech.

In general, only /p/ showed a significant improvement after the intervention. In other words, learners made fewer errors in the posttest than in the pretest. This suggests that the audio-visual activities were effective in helping students improve their production of /p/.

B. Discussion

Learners' perception of their ability in recognizing and producing stop consonants

before the intervention

The findings show that learners' perceptions of their own abilities to recognize and pronounce stop consonants were consistent. Specifically, in both recognition and pronunciation, /p/, /d/, and /k/ were perceived as more challenging than/b/, /t/, and /g/. In addition, learners reported being more confused by the /p/-/b/ pair than by other pairs when listening. Similarly, both /p/–/b/ and /k/–/g/ were identified as challenging in production. These findings strengthen the conclusion regarding students' perception of and production of single sounds. Although perceptions may not always fully reflect actual performance, the questionnaire results align with previous studies on Vietnamese EFL learners' pronunciation problems. For example, Vietnamese EFL learners often experience difficulties with /p/, as they tend to substitute it with /b/ or /f/ due to interference from their mother tongue [4]. Moreover, they tend to mispronounce the voiced alveolar stop /d/ in words with the 'ed' ending [28]. These learners also tend to omit the /k/ sound in the final position [5].

A detailed analysis by word position shows that, for each sound, the position that makes it challenging differs. For example, /p/ is perceived as difficult to recognize in the final position, as in 'pop', while producing it is considered difficult in both medial and final positions. Similarly, /k/ is perceived as difficult to recognize in the final position but difficult to pronounce in both the medial and final positions. In contrast, /d/ is reported as problematic in initial and final positions. Regardless of recognition or production, the final position of a word consistently proves to be the most challenging. These findings from learners' perceptions align with research on the final sound in the context of teaching and learning pronunciation in Vietnam. Previous studies have consistently reported that Vietnamese EFL learners encounter serious difficulties with final consonants, particularly stop consonants. Learners often delete final consonants [28] or fail to release them, leading to miscommunication [5].

Studies have also pointed out the causes of difficulties related to consonants, especially final consonants, in Vietnamese English learners. The main cause is the first-language transfer (L1 transfer). In fact, Vietnamese has only a limited set of final consonants, and all are pronounced in the closed, unaspirated form, while English has a more diverse system of final consonants and requires clear sound release [28]. This difference often causes learners to omit final consonants and simplify consonant clusters [5, 28]. Another cause is the lack of systematic pronunciation instruction in traditional English classrooms, which can lead to a lack of awareness of the phonetic differences between English and Vietnamese [28]. In addition, the difficulty is reinforced by limitations in auditory perception, where learners often fail to distinguish between voiced and voiceless contrasts in word-final position, leading to confusion in minimal pairs [5]. Taken together, these factors explain why problems with consonants, especially in word-final position, persist and are difficult to overcome.

Application of audio-visual aids to enhance students' pronunciation of stop consonants

The findings revealed noticeable improvements in both students' recognition and pronunciation of stop consonants. Regarding sound recognition, students showed positive changes in recognizing the sounds /p/, /b/, /d/, /k/, and /g/ in specific word positions. In the initial position of words, /k/ and /g/ were recognized more correctly, while /p/ was enhanced in the medial position and /d/ in the final position. These improvements can be explained through multimedia learning theory [10], which posits that information presented through both auditory and visual channels is better retained. In pronunciation instruction, combining auditory input with visual cues of articulatory movements strengthens learners' phonemic awareness and helps them map sounds to articulations [30].

In terms of sound production, at the single-word level, students' performance improved significantly for /p/, /b/, /d/, and /g/ compared to the pretest. However, the analysis of positional performance revealed uneven progress. Particu-

larly, while /p/ was produced similarly across all positions, /b/ and /d/ were produced more accurately in the initial position, and /g/ improved notably in the medial position. At the sentence level, only /p/ demonstrated a clear reduction in errors, indicating a significant improvement in pronunciation. These positive outcomes demonstrate the effectiveness of audiovisual aids in improving students' pronunciation of stop consonants. These findings align with Flege's speech learning model, which posits that the accurate perception of sounds precedes the accurate production. Learners must first discriminate new phonetic categories before they can produce them correctly. Thus, improvements in recognizing /p/, /b/, /d/, /k/, and /g/ served as a foundation for enhanced production of /p/, /b/, /d/, and /g/ at the single-word level [9].

The findings of students' improvement in this study are also consistent with previous empirical studies on the effectiveness of audio-visual aids in pronunciation instruction. Specifically, Hardison [30] found that visually assisted pronunciation training enhanced both recognition and production of English consonants compared to auditory training alone. Naasse et al. [25] further confirmed that audio-visual materials strengthen learners' phoneme discrimination and pronunciation accuracy compared to traditional instruction.

V. CONCLUSION

Learners demonstrated consistent perceptions of their ability to recognize and produce stop consonants before the intervention. They identified /p/, /d/, and /k/ as particularly challenging, whereas /b/, /t/, and /g/ were perceived as more manageable. Regarding minimal pairs, the /p/-/b/ contrast was especially problematic in both perception and production, while the /k/-/g/ pair posed difficulty in discrimination, particularly in final positions. These findings indicate that among the three sounds /p/, /d/, and /k/, learners struggle more with /p/ and /k/ in production. These results confirm that learners' self-perceptions are consistent across recognition and production. Understanding learners' perceived ability and challenges provides a foundation for examining the effectiveness of the audio-visual intervention.

The use of audio-visual aids proved effective in enhancing students' recognition and pronunciation of English stop consonants in specific word positions. For recognition, /p/, /b/, /d/, /g/, and /k/ improved significantly after the intervention. In particular, /p/ was recognized more accurately in the medial position, /b/ in the final position, /k/ in the initial position, /d/ in the final position, and /g/ in the initial position. Regarding sound production at the single-word level, /p/, /b/, /d/, and /g/ improved significantly, with /p/ showing consistent progress across three positions, /b/ and /d/ improving in the initial position, and /g/ in the medial position. At the sentence level, only /p/ showed a reduction in pronunciation errors. Overall, students' pronunciation of stop consonants improved with the use of audiovisual aids.

However, several limitations should be noted to provide a foundation for further research on this topic. Firstly, the sample size was relatively small and consisted only of non-English majors. This restricts the generalizability of the findings to broader learner populations such as English majors. Secondly, the intervention spanned only eight sessions. This can be seen as short-term and may not be sufficient for learners to consolidate and retain pronunciation improvements. Thus, it limits the ability to conclude the long-term effectiveness of the instructional approach. Third, the pronunciation assessment relied primarily on teacher judgment, with listening and scoring conducted without the support of audio analysis tools. This reliance on subjective scoring may have introduced bias and reduced the measurement's reliability. Finally, learners' recordings were inconsistently submitted via the LMS during the intervention. Despite repeated reminders, delays persisted. This resulted in learners not being corrected promptly and may have affected the accuracy and comprehensiveness of the results. Generally, the above limitations need to be acknowledged and supported in any further research on this topic.

REFERENCES

[1] Derwing TM, Munro MJ. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*. 2005;39(3): 379–397. https://doi.org/10.2307/3588486.

[2] Fraser H. *Coordinating improvements in pronunciation teaching for adult learners of English as a second language*. Canberra, Australia: DETYA (ANTA Innovative Project); 2000.

[3] Levis JM. *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge, United Kingdom: Cambridge University Press; 2018.

[4] Nguyen TT, Ingram J. A corpus-based analysis of transfer effects and connected speech processes in Vietnamese English. In: Cassidy S, Cox F, Mannell R, Palethorp S (eds.) *Proceedings of the 10th Australian International Conference on Speech Science and Technology*. Sydney, Australia: Macquarie University; 2004. p.516–521. https://assta.org/proceedings/sst/2004/proceedings/papers/sst2004-364.pdf[Accessed16January2026].

[5] Nguyen DH. *A study of interlanguage word-final consonant pronunciation by Vietnamese learners and the degree of intelligibility as judged by native and non-native English speakers*. Dissertation. Bangkok, Thailand: Assumption University; 2019. https://repository.au.edu/handle/6623004553/23980 [Accessed 16 January 2026].

[6] Nguyen HD. Errors in word-final consonant pronunciation in Vietnamese English interlanguage. In: *Proceedings of the 15th International Conference on Humanities and Social Sciences (IC-HUSO 2019)*. 11–12 November 2019; Khon Kaen, Thailand. Khon Kaen, Thailand: Khon Kaen University; 2019. p.304–326.

[7] Gilakjani AP, Sabouri NB. Why is English pronunciation ignored by EFL teachers in their classes? *International Journal of English Linguistics*. 2016;6(6): 195–208. https://doi.org/10.5539/ijel.v6n6p195.

[8] Harmer J. *The practice of English language teaching*. 3rd ed. London, United Kingdom: Pearson Education; 2001.

[9] Flege JE. Second language speech learning: Theory, findings, and problems. In: Strange W (ed.). *Speech perception and linguistic experience: Issues in cross-language research*. Maryland, United States: York Press; 1995. p.233–277.

[10] Mayer RE. *Multimedia learning*. 3rd ed. Cambridge, United Kingdom: Cambridge University Press; 2021. https://doi.org/10.1017/9781316941355.

[11] Mercer CD, Mercer AR. *Teaching students with learning problems*. 4th ed. United States: Merrill Publishing Company; 1993.

[12] Moustafa M. Multisensory approaches to language learning. *Journal of Reading Education*. 1999;24(3): 36–43.

[13] Zhan L, Cheng J. The relationship between multisensory stimulus-integrated foreign language learning models and students' psychological states and language skill development: an empirical analysis using the global learning assessment database. *Frontiers in Psychology*. 2025;16: 1639885. https://doi.org/10.3389/fpsyg.2025.1639885.

[14] Hirata Y, Kelly SD. Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*. 2010;53(2): 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243).

[15] Celce-Murcia M, Brinton DM, Goodwin JM. *Teaching pronunciation: A course book and reference guide*. 2nd ed. Cambridge, United Kingdom: Cambridge University Press; 2010.

[16] Munro MJ, Derwing TM. Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*. 2020;6(3): 283–309. https://doi.org/10.1075/jslp.20038.mun.

[17] Derwing TM, Munro MJ. *Pronunciation fundamentals: evidence-based perspectives for L2 teaching and research*. Netherlands: John Benjamins Publishing Company; 2015. https://doi.org/10.1075/lllt.42?locatt=mode:legacy.

[18] Kang O, Thomson RI, Moran M. Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*. 2017;68(1): 115–146. https://doi.org/10.1111/lang.12270.

[19] Ladefoged P, Johnson K. *A course in phonetics*. 7th ed. United States: Cengage Learning; 2014.

[20] Bradshaw MJ, Hultquist BL, Hagler D. *Innovative teaching strategies in nursing and related health professions*. Massachusetts, United States: John & Bartlett Learning; 2024. https://www.jblearning.com/catalog/productdetails/9781284296808 [Accessed 16 January 2026].

[21] Venkataiah N. *Educational technology: teaching and learning*. New Delhi, India: APH Publishing Corporation; 2008. https://books.google.co.in/books?id=VpyTn9A3NooC&printsec=frontcover#v=onepage&q&f=false [Accessed 19 January 2026].

[22] Shams L, Seitz AR. Benefits of multisensory learning. *Trends in Cognitive Sciences*. 2008;12(11): 411–417. https://doi.org/10.1016/j.tics.2008.07.006.

[23] Naeini B, Adni Z. The impact of explicit vs. implicit instruction on pronunciation intelligibility among Iranian EFL learners. *Journal of Studies in Learning and Teaching English*. 2017;6(2): 103–123. https://en.civilica.com/doc/1660762/ [Accessed 7 January 2026].

[24] Linebaugh G, Roche T. Learning to hear by learning to speak: The effect of articulatory training on Arab learners' English phonemic discrimination. *Aus-*

tralian Review of Applied Linguistics. 2013;36(2): 146–159. https://doi.org/10.1075/aral.36.2.02lin.

[25] Naasse D, Ben Makki A. *Enhancing the EFL learners' pronunciation by using audio-visual aids in teaching pronunciation: A case study of fourth year pupils at Medjedel middle school*. Dissertation. Biskra, Algeria: Mohamed Khider University of Biskra; 2014. http://archives.univ-biskra.dz/handle/123456789/4600 [Accessed 19 January 2026].

[26] Gamboa LF, Sevilla JR. The effectiveness of audio-visual materials in improving ESL learners' pronunciation. *Asian EFL Journal*. 2021;25(6): 120–138.

[27] Turgunboev RI. The effectiveness of authentic audio-visual materials in improving pronunciation and speaking confidence among EFL learners. *International Bulletin of Applied Science and Technology*. 2025;5(12): 33–40. https://doi.org/10.5281/zenodo.17899428.

[28] Bui TTL, Mai TH, Diep HN. Common errors in pronouncing final consonants of English-majored sophomores at Tay Do University, Vietnam. *European Journal of English Language Teaching*. 2021;6(3): 120–161. http://dx.doi.org/10.46827/ejel.v6i3.3640.

[29] Zhi S, Levy RP, Meylan SC. Multimodal input aids a Bayesian model of phonetic learning. *arXiv* [Preprint] 2024. Version 1. https://arxiv.org/abs/2407.15992.

[30] Hardison DM. Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*. 2003;24(4): 495–522. https://doi.org/10.1017/S0142716403000250.

[31] Li Y. Enhancing L2 sound learning through the integration of audio-visual information: Phonetic training in the classroom. *Language Teaching Research*. 2024;30(6): 3355–3369. https://doi.org/10.1177/13621688231219773.

[32] Toyama M, Hori Y. Technology-enhanced multimodal approaches in classroom L2 pronunciation training: A mini review. *Frontiers in Education*. 2025;10: 1552470. https://doi.org/10.3389/feduc.2025.1552470.

[33] Burns A. *Doing action research in English language teaching: A guide for practitioners*. New York, United States: Routledge; 2009. https://doi.org/10.4324/9780203863466.

[34] Kemmis S, McTaggart R. *The action research planner: Doing critical participatory action research*. Singapore: Springer; 2014. https://doi.org/10.1007/978-981-4560-67-2.

[35] Tra Vinh University. *Study plan for the Sport Management program [Kế hoạch học tập ngành Quản lý thể dục thể thao khóa 2024]*. https://gdtc.tvu.edu.vn/index.php/dao-tao/ke-hoach-hoc-tap/ke-hoach-hoc-tap-khoa-2024 [Accessed 19 January 2026].

[36] Tra Vinh University. *Announcement on the selection of non-major foreign languages and English placement testing for full-time undergraduate students admitted in 2025 [Thông báo về việc lựa chọn học ngoại ngữ không chuyên và kiểm tra năng lực tiếng Anh đầu vào đối với sinh viên các ngành đại học hệ chính quy tuyển sinh năm 2025]*. Vinh Long, Vietnam: Tra Vinh University. Announcement number: 2936/TB-DHTV, 2025. https://dvt.tvu.edu.vn/vi/news/tuyen-sinh-dai-hoc/thong-bao-lua-chon-ngoai-ngu-khong-chuyen-va-kiem-tra-nang-luc-tieng-anh-dau-vao-doi-voi-sinh-vien-cac-nganh-dai-hoc-he-chinh-quy-tuyen-sinh-nam-2025-38916.html [Accessed 19 January 2026].

[37] Jahara SM, Abdelrady A. Pronunciation problems encountered by EFL learners: An empirical study. *Arab World English Journal (AWEJ)*. 2021;12(4): 194–212. https://doi.org/10.24093/awej/vol12no4.14.

[38] Skywork. *OpenAI.fm Deep dive: a new era in AI audio creation*. https://skywork.ai/skypage/en/OpenAI.fm-Deep-Dive-A-New-Era-in-AI-Audio-Creation/1972851114648465408 [Accessed 19 January 2026].

Creative Commons Attribution-NonCommercial 4.0 International License.