OPEN ACCESS

International Journal of Language Learning and Teaching Research
IJLLTR
Volume 1 (2026)
https://doi.org/10.65553/IJLLTR.260106

Integrating Perception and Production through Technology-Supported Mandarin Pronunciation Instruction: A Quasi-Experimental Classroom Study

Liu Shi
School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China
Yanping Yan
School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China
Jingjing Xing
The Ontario Institute for Studies in Education, University of Toronto, Toronto, Canada
Jie Zhang
School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China
Corresponding author: Zhangjie1@ncwu.edu.cn

Abstract

Digital speech technologies can extend pronunciation practice beyond limited class time, but their pedagogical value depends on principled integration with instruction and feedback. Grounded in the revised Speech Learning Model (SLM-r), this study developed and piloted a four-phase perception–production model for Mandarin pronunciation instruction. The model was implemented through Hanyin Yusheng, a WeChat mini-program combining auditory

Submission date: May 5, 2026; Revision date: November 20, 2026; Online: December 15, 2026

Copyright: © 2026 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

discrimination, recorded production, automated speech evaluation, practice resources, and teacher-accessible learning analytics. The project comprised two connected studies. First, 40 intermediate learners completed parallel perception and production tasks targeting Mandarin lexical tones, the unaspirated–aspirated onset contrast represented by Pinyin d–t (/t/–/tʰ/), and the finals /i/, /a/, and /u/. Perception and production scores were strongly correlated (Spearman's $\rho = .755$, $p < .001$). Second, 32 learners in intact classes received either technology-enhanced perception–production instruction (n = 16) or conventional teacher-led instruction (n = 16). Across a four-week, eight-lesson intervention, both groups completed a pretest, an immediate posttest, and a seven-day follow-up test. A 2 × 3 mixed analysis showed significant main effects of group, $F(1, 30) = 21.821$, $p < .001$, $\eta p^2 = .421$, and test occasion, $F(2, 60) = 160.883$, $p < .001$, $\eta p^2 = .843$, but no significant group-by-time interaction, $F(2, 60) = 1.555$, $p = .222$, $\eta p^2 = .049$. The findings provide preliminary evidence for the feasibility of integrating perception, production, automated feedback, and spaced practice. Because the interaction was nonsignificant, however, they do not warrant a strong causal claim that the technology-enhanced condition produced greater gains. Implications are discussed for technology-enhanced Mandarin pronunciation pedagogy and more rigorous future trials.

Keywords: Mandarin pronunciation; second language speech learning; automated speech evaluation; perception–production relationship; technology-enhanced language learning; classroom intervention

Introduction

Pronunciation is central to communicative success in an additional language because inaccurate segmental and suprasegmental patterns can impede intelligibility, lexical access, and interaction. Mandarin pronunciation is especially demanding. In addition to consonantal and vowel contrasts, learners must attend to lexical tone, where changes in fundamental-frequency height and contour distinguish word meanings. Learners therefore need both perceptual categories for relevant contrasts and articulatory and laryngeal routines for producing them. Pronunciation teaching, however, is often constrained by limited contact time, large classes, delayed feedback, and an emphasis on imitation without systematic perceptual learning.

Digital speech technologies may alleviate some of these constraints. Mobile platforms can distribute short listening and speaking tasks, provide repeated exposure to models, collect spoken responses, and return immediate automated feedback. Automatic speech recognition (ASR) and automated speech evaluation can increase practice frequency and privacy, potentially supporting learner autonomy while reducing the cost of individualized feedback. Evidence from computer-assisted pronunciation training is broadly positive but heterogeneous. A meta-analysis of 15 ASR pronunciation studies reported a medium overall effect and identified feedback type, treatment duration, linguistic target, and peer involvement as moderators (Ngo et al., 2024). Individual studies have likewise reported benefits from ASR-supported speaking practice (Bashori et al., 2024; Sun, 2023), mobile feedback (Dai & Wu, 2023), and automated speech evaluation (Zou et al., 2023a). Technology, however, is

not itself a pedagogy: opaque scores, accent-related recognition errors, limited diagnosis, and decontextualized repetition may fail to establish robust perceptual categories (Wu et al., 2025).

Recent syntheses reinforce the importance of aligning technological affordances with pronunciation pedagogy and assessment. In a systematic review of computer-assisted pronunciation training, Amrate and Tsai (2025) documented variation in instructional targets and assessment practices and noted that relatively few studies used broader measures such as intelligibility and comprehensibility. This observation highlights the need to specify both the phonetic targets and the scope of the outcomes when evaluating technology-supported pronunciation instruction.

This limitation is consequential in Mandarin. Tone realization varies with prosodic and phonological context, and perception and production may share some mechanisms while remaining partly skill-specific (Hao, 2018; Li & DeKeyser, 2017). Visual-acoustic feedback can facilitate both Mandarin tone perception and production (Chen, 2024), while perception–production relationships appear stronger for critical contour cues than for every acoustic dimension (Leung & Wang, 2020). Effective instruction should therefore coordinate discrimination, guided production, interpretable multimodal feedback, and repeated opportunities to revisit difficult categories.

To address this need, the present study developed a technology-supported perception–production model for teaching Mandarin tones and selected segmental contrasts. The model was grounded in SLM-r (Flege & Bohn, 2021), which conceptualizes second-language speech learning as a dynamic, input-driven process shaped by the quantity and quality of experience and by individual differences. It was implemented through Hanyin Yusheng (literally, Chinese Phonetics and Speech), a project-developed WeChat mini-program connecting diagnostic assessment, adaptive lesson planning, multimodal classroom work, automated feedback, and a seven-day review cycle.

The study makes three contributions. First, it extends technology-enhanced pronunciation research beyond its predominantly English-learning contexts by examining Mandarin as an additional language, where learners must coordinate lexical tones and segmental contrasts. Second, it extends the application of the revised Speech Learning Model (SLM-r) by considering technology-mediated learning environments as a new context for phonetic experience, in which diagnostic input, repeated production opportunities, and feedback cycles may shape the development of perceptual and productive categories. Third, it treats perception and production as coordinated but separately trainable objectives rather than assuming that improvement in one automatically transfers to the other.

Theoretical Framework and Literature Review

The perception–production relationship in second-language speech

The original Speech Learning Model proposed that second-language production accuracy is constrained by how learners perceive L2 sounds and classify them in relation to established first-language categories (Flege, 1995). SLM-r retains the interaction between first- and additional-language phonetic systems but gives greater attention to dynamic change across the lifespan, input distributions, and individual differences (Flege & Bohn, 2021). Importantly, the revised account does not require

a simple perception-first sequence. Perceptual and production learning may develop in parallel, at different rates, and under different task demands.

Empirical work supports both association and relative independence. In Mandarin tone learning, overall identification and read-aloud accuracy may correlate even when error patterns differ by tone context and task (Hao, 2018). Leung and Wang (2020) found production–perception correspondence for critical F0 contour cues but not for lower-weighted height cues. Training can also be channel-specific: learners tend to perform best in the modality they practiced, with incomplete transfer to the other modality (Li & DeKeyser, 2017). A correlation therefore neither makes perception and production identical constructs nor implies that training only one will necessarily improve the other.

Instruction should accordingly include high-quality models, contrastive listening and categorization, explicit articulatory explanation, controlled production, and feedback directing attention to relevant acoustic or articulatory properties. Learners should move iteratively between input and output: perception provides targets and error awareness, whereas production sharpens attention to contrasts and tests whether emerging categories can be used in speech.

While SLM-r provides a theoretical account of how perceptual and productive phonological categories develop through experience, the present study extends this perspective by examining how technology-mediated environments may increase opportunities for noticing, feedback, practice, and revision during pronunciation learning. Thus, technology is conceptualized not as a replacement for phonological learning mechanisms but as an infrastructure that facilitates interaction between perception, production, and instructional mediation.

Explicit and technology-supported pronunciation instruction

Pronunciation instruction is effective overall, although effects depend strongly on what is taught and how outcomes are measured. Lee et al. (2015) reported a large overall effect across 86 reports, with stronger effects for longer treatments, fuller feedback, and controlled measures. Saito and Plonsky (2019) showed that conclusions vary according to whether outcomes target specific versus global pronunciation, use human or acoustic ratings, and elicit controlled or spontaneous speech. Gains on narrow word-reading tests should therefore not be equated with gains in spontaneous-speech comprehensibility.

Computer-assisted pronunciation training enables repetition and immediate response. Successful systems must align technical functions with learning principles rather than foreground novelty (Neri et al., 2002). ASR can support learner autonomy by allowing learners to monitor recognition results and revise their output (McCrocklin, 2016; McCrocklin & Levis, 2025), but feedback design matters (Wang & Young, 2015), and automated feedback is not inherently superior to teacher or peer feedback (Dai & Wu, 2023). Ngo et al. (2024) similarly found larger ASR effects with explicit feedback, moderate or longer treatments, and peer involvement. Recent systematic evidence also emphasizes the need to align computer-assisted pronunciation training with explicit pedagogical and assessment principles (Amrate & Tsai, 2025). Automated evaluation should thus function as one source of formative evidence in a teacher-guided cycle, not as an unquestioned replacement for professional judgment.

Recent empirical studies have begun to examine AI-supported pronunciation

practice in more targeted instructional settings. Aly and Alsehibany (2025) investigated the integration of ChatGPT and SpeechAce in EFL pronunciation instruction and reported improvements in selected pronunciation and motivational outcomes. Xodabande et al. (2025) likewise found that a ChatGPT-supported intervention was associated with sustained pronunciation gains among Iranian EFL learners. More directly relevant to Mandarin, Cheng et al. (2025) developed an AI-enhanced multimodal tone-training program combining automatic speech recognition, dynamic time warping, and visual feedback; their findings indicated improvements in both tone production and perception among learners from non-tonal language backgrounds. Liu (2026) examined voice-based ChatGPT for CSL learners and reported improvement in comprehensibility and tone, although gains in initials and finals were comparatively smaller. These studies suggest growing potential for AI-mediated pronunciation practice while also showing that outcomes may vary by linguistic target, learner background, and feedback design.

Recent work also highlights the motivational and affective conditions of technology use. Wang and Xue's (2024) intervention found that AI-driven chatbots enhanced Chinese EFL learners' academic engagement, suggesting that interactive affordances can affect participation as well as linguistic practice. Liu et al. (2025) identified distinct emotional and autonomous-learning profiles in AI-enhanced data-driven language learning, underscoring the heterogeneity concealed by group averages. In a robot-assisted environment for students preparing to teach Chinese internationally, Zhang and Yan (2025) found that grit and enjoyment contributed to engagement, with enjoyment emerging as the stronger predictor. Together, these findings imply that the effectiveness of digital pronunciation work may depend on how technology structures agency, emotion, and sustained participation, not merely on scoring accuracy.

Learner responses to automated feedback are also receiving closer attention. Huang and Pan (2026), using a longitudinal design, found that learners' perceptions of the diagnostic value of AI pronunciation feedback were associated with subsequent trust, pronunciation self-efficacy, and pronunciation performance. In addition, Zou et al. (2025) examined learners' acceptance of an AI speech-evaluation program, underscoring the importance of user acceptance in technology-supported speaking practice. These findings reinforce the view that the pedagogical value of automated pronunciation tools depends not only on technical feedback delivery but also on whether learners regard the feedback as useful, trustworthy, and actionable.

Multimodal input, learning analytics, and spaced practice

Speech learning is inherently multimodal: learners coordinate auditory input, visual representations of articulatory movement or pitch, kinesthetic experience, orthographic representations, and feedback on their own speech. In Mandarin tone teaching, a four-week study found larger perceptual and production gains from Praat-generated audiovisual feedback than from researcher feedback alone (Chen, 2024). Multimedia can support attention and retention when learner proficiency and cognitive resources are considered (Teng, 2022, 2023), but poorly aligned displays may increase processing demands. Visualizations and videos should direct attention to a linguistic target rather than decorate a lesson.

Digital platforms can also create a learning-analytics feedback loop. Pre-class responses reveal class-wide and individual difficulties and help teachers allocate time

and select examples; post-class and follow-up responses indicate change and retention. This is pedagogically different from merely digitizing worksheets because learner evidence informs subsequent instruction. Human interpretation remains essential. Qin and Dong (2025), for example, showed that learners' lived affective interpretations of GenAI-mediated feedback did not map automatically onto their actual revision behavior. Although their study concerned writing, its sociocultural implication transfers to pronunciation: automated output becomes developmentally useful only when learners interpret it, act on it, and receive appropriate mediation. Accordingly, the present model combines automated evaluation with teacher explanation in a predictable four-phase routine.

Recent research on multimodal generative AI in Chinese-language learning further suggests that engagement and phonological development should not be treated as interchangeable outcomes. Wang (2026), for example, reported that multimodal GenAI supported task preparation and learner participation, while improvements in tonal accuracy and delivery confidence remained limited. This finding is relevant to the present study because it cautions against assuming that greater access to multimodal technology or stronger engagement necessarily translates into equivalent gains across pronunciation dimensions.

Research gaps and questions

Three gaps motivated the project. First, technology-enhanced pronunciation research has concentrated on English, with comparatively little classroom evidence on Mandarin tones and segmental contrasts. Second, many interventions foreground production and automated scoring without a theoretically explicit perceptual component. Third, some studies infer effectiveness from experimental-group pre–post gains even when the relevant group-by-time comparison is absent or nonsignificant. A persuasive evaluation must consider change over time, between-group differences, and their interaction.

The study addressed the following questions:

> To what extent are Mandarin pronunciation perception and production scores associated among intermediate learners completing parallel tasks?
>
> How do pronunciation scores vary by instructional group and test occasion during a four-week classroom intervention?
>
> What pedagogical conclusions about the technology-supported perception–production model are warranted by the observed main and interaction effects?

Study1

Research Design

Study 1 employed a cross-sectional perception–production design to examine the relationship between Mandarin pronunciation perception and production among intermediate learners and to provide empirical evidence for integrating these two dimensions in subsequent technology-supported instruction.

Participants

Forty intermediate learners of Mandarin participated in Study 1, including 20 women and 20 men, with ages ranging from 20 to 25 years. All participants had achieved intermediate HSK proficiency or above and had studied Mandarin in China for approximately 2.5–3 years. At the time of data collection, they were engaged in formal Mandarin learning programs and had received systematic instruction in Mandarin pronunciation and phonological features.

The participants were selected because intermediate learners have acquired basic Mandarin phonological knowledge while still experiencing challenges in accurately perceiving and producing non-native phonological contrasts. This proficiency level was therefore considered appropriate for examining the relationship between pronunciation perception and production, as learners at this stage possess sufficient linguistic experience to complete controlled perception and production tasks while maintaining observable variability in pronunciation performance.

All participants voluntarily participated in the study and completed both perception and production tasks. They were informed of the research procedures and provided consent before data collection. Their participation did not affect their course grades or academic evaluation. The participants were international learners from five Southeast Asian countries: Vietnam, Laos, Cambodia, Thailand, and Indonesia. Their diverse linguistic backgrounds provided a cross-national context for examining Mandarin pronunciation perception and production. The distribution of participants across the five countries is reported separately for each study.

Instruments: Perception and Production Tasks

Study 1 employed parallel perception and production tasks designed to assess learners' ability to perceive and produce the same Mandarin phonological contrasts. The tasks targeted three major pronunciation domains: (a) the four Mandarin lexical tones, (b) the alveolar stop aspiration contrast represented by Pinyin d and t (approximately /t/ and /t^h/), and (c) the finals i, a, and u. These targets were selected because they represent fundamental yet challenging features of Mandarin pronunciation for second-language learners, involving both suprasegmental and segmental contrasts.

The test materials consisted of 24 matched syllable items. Specifically, the materials included 16 tone-related items representing the four lexical tones, 4 aspiration-contrast items targeting the d–t distinction, and 4 final-vowel items involving i, a, and u. This balanced item structure enabled direct comparison between learners' perceptual and productive performance across the same phonological targets. The stimuli included combinations such as dī, dí, dǐ, dì, tī, tí, tǐ, tì, dā, dá, dǎ, dà, dū, dú, dǔ, and dù. The matched-item design enabled direct comparison between learners' perceptual and productive performance on the same phonological targets.

The perception task consisted of 24 forced-choice identification items. For each item, participants listened to a recorded Mandarin syllable and selected the corresponding option from the provided choices. The task was designed to evaluate learners' ability to distinguish the target phonological contrasts, including tonal differences, aspiration contrasts, and final-vowel distinctions.

The production task consisted of 24 read-aloud items corresponding to the perception stimuli. Participants were required to produce the same target syllables

orally, allowing their production accuracy to be compared with their perceptual performance. By using parallel task designs, the study examined whether learners who demonstrated stronger perceptual sensitivity to Mandarin phonological contrasts also tended to achieve higher production accuracy.

Data Collection

The perception task, delivered through Wenjuanxing, contained 24 forced-choice identification items recorded at a natural pace by a provincial broadcaster certified at Putonghua Proficiency Test Level 1-B. Items were separated by 10 seconds; replay, skipping, and repetition were disabled, and administration was synchronously proctored via Zoom. The production task comprised 24 read-aloud items based on the same targets. Participants read each item twice and submitted recordings within a fixed period.

The production recordings were segmented and denoised in Audacity before scoring. Two trained Mandarin pronunciation instructors independently evaluated all recordings based on a predefined scoring rubric. The rubric focused on three target dimensions: lexical tone accuracy, aspiration contrast accuracy, and final-vowel accuracy. Each item was coded as either accurate production (1) or inaccurate production (0). The production score was calculated as the percentage of correctly produced items within each domain and as an overall accuracy score across all target items.

To ensure the reliability of the scoring procedure, interrater agreement between the two raters was examined using the intraclass correlation coefficient (ICC). The results indicated excellent agreement between the raters (ICC = .91, $p < .001$), suggesting high consistency in production scoring. Any disagreements were discussed and resolved through joint review of the recordings before the final dataset was established.

For statistical analysis, accuracy scores were first summarized by target domain. Because the score distributions were treated as nonparametric, Spearman's rank correlation analysis was conducted to examine the association between overall perception and production performance.

Results

To address Research Question 1, Study 1 first examined learners' performance across the three target pronunciation domains and then investigated the association between pronunciation perception and production. Descriptive accuracy patterns were summarized by target domain, followed by a nonparametric correlation analysis to determine whether learners' perceptual and productive abilities were systematically related.

Table 1 presents descriptive statistics for learners' perception and production accuracy across the three target domains. Overall, production accuracy was descriptively higher than perception accuracy in all domains. The largest difference was observed for finals, with production accuracy exceeding perception accuracy by 18.5 percentage points, followed by lexical tones (13.6 percentage points) and onset contrasts (3.5 percentage points). These descriptive patterns suggest that learners demonstrated relatively stronger performance in producing the target sounds than in identifying them under the current task conditions. However, these differences should not be interpreted as evidence that production is inherently easier than

perception, because the two tasks involved different response demands and were not designed for direct within-domain comparison.

Table 1. *Perception and production accuracy by target domain in Study 1 (n = 40)*

Target domain	Perception		Production	
	M	SD	M	SD
Mandarin lexical tones	57.0	18.6	70.6	16.4
Pinyin d–t onsets	67.5	15.2	71.0	14.8
Finals i, a, and u	64.0	17.1	82.5	13.5
Overall accuracy	62.8	13.7	74.7	12.9

Note. Accuracy scores represent the percentage of correctly identified or produced target items.

Beyond these descriptive differences, the study further examined whether learners' perceptual and productive performances were statistically associated. Because the score distributions were treated as nonparametric, Spearman's rank correlation analysis was conducted between overall perception and production scores.

As shown in Table 2, learners achieved a mean perception accuracy of 62.8% (SD = 13.7) and a mean production accuracy of 74.7% (SD = 12.9). Spearman's rank correlation analysis revealed a strong positive association between perception and production performance ($\rho = .755$, $p < .001$, 95% CI [.580, .863]). This finding indicates that learners who demonstrated higher sensitivity to Mandarin phonological contrasts also tended to produce these contrasts more accurately.

Table 2. *Spearman correlation between overall perception and production scores in Study 1 (n = 40)*

Variables	M	SD	Spearman's ρ
Perception accuracy	62.8	13.7	-
Production accuracy	74.7	12.9	.755***

Note. Spearman's rho was used because score distributions were treated as nonparametric. ***p < .001.

The findings provide empirical support for a close relationship between Mandarin pronunciation perception and production. However, the correlation should be interpreted cautiously. The observed association does not establish the causal direction between the two abilities, nor does it indicate that improvement in one modality will automatically result in improvement in the other. Rather, the findings suggest that perception and production are interconnected but partially distinct dimensions of second-language speech learning. This result provides an empirical rationale for integrating perceptual training and productive practice within the subsequent technology-supported perception–production instructional model.

The two studies were designed to address complementary questions rather than to compare learners at the same proficiency level. Study 1 examined whether perception

and production were sufficiently associated to justify their joint inclusion in an instructional model. The intermediate learners were able to complete parallel perception and production tasks while still showing meaningful individual variation. The observed association therefore provided a rationale for coordinating the two skill dimensions, but it did not establish a perception-to-production transfer effect or identify an optimal instructional sequence. Study 2 extended this rationale to a less experienced learner group, for whom explicit support in establishing perceptual categories and articulatory routines may be particularly relevant. The different proficiency profiles should consequently be understood as serving distinct purposes within the two-study design, not as evidence that Study 1 directly predicts the intervention outcomes in Study 2.

Study 2

Research Design

Study 2 employed a four-week quasi-experimental classroom design to examine the implementation and preliminary outcomes of the technology-supported perception-production instructional model developed based on the findings of Study 1. Specifically, the study compared learners receiving technology-enhanced perception-production instruction with those receiving conventional teacher-led pronunciation instruction. Because participants were assigned according to pre-existing intact classes and no individual random assignment was conducted, the study was classified as quasi-experimental rather than a randomized controlled experiment (Ballance, 2024). This design reflects authentic classroom conditions while allowing the researchers to examine changes in pronunciation performance across different instructional conditions. However, because participants were assigned by pre-existing intact classes rather than through individual random assignment, observed differences should be interpreted cautiously and should not be taken as definitive evidence of causal effects.

A 2 (instructional group: technology-enhanced perception–production instruction vs. conventional instruction) × 3 (test occasion: pretest, immediate posttest, and seven-day follow-up test) mixed design was employed. The pretest was administered before the intervention to establish baseline pronunciation performance, the immediate posttest was conducted after the four-week intervention to examine short-term learning gains, and the follow-up test was administered seven days later to evaluate the retention of pronunciation improvement.

Participants

Thirty-two international learners of Mandarin participated in Study 2, including 16 women and 16 men, with ages ranging from 22 to 25 years. Compared with the intermediate learners recruited in Study 1 (N = 40), the participants in Study 2 had relatively limited Mandarin learning experience, with approximately one year of exposure to Mandarin. Before the intervention, none of the participants had received systematic instruction specifically targeting Mandarin pronunciation. At the time of data collection, they were enrolled in formal Mandarin language courses and were receiving regular classroom-based Mandarin instruction.

The participants were selected because learners with limited Mandarin learning

experience often encounter difficulties in establishing accurate perceptual categories and articulatory routines for Mandarin phonological contrasts. This learner profile was therefore considered appropriate for examining the feasibility and pedagogical potential of the technology-supported perception–production instructional model, as these learners had sufficient Mandarin learning experience to participate in classroom pronunciation activities while still demonstrating considerable potential for pronunciation development.

Participants were assigned according to their pre-existing intact classes rather than individual random assignment. One intact class consisting of 16 learners (8 women and 8 men) received technology-enhanced perception–production instruction, whereas the other intact class consisting of 16 learners (8 women and 8 men) received conventional teacher-led pronunciation instruction. Both groups followed the same instructional schedule and targeted the same Mandarin pronunciation features during the four-week intervention.

All participants voluntarily participated in the study and completed the pretest, immediate posttest, and seven-day follow-up test. They were informed of the research procedures and provided informed consent before data collection. Participation was voluntary and had no influence on their course grades or academic evaluation.

Instructional Model and Technological Infrastructure

Findings from Study 1, together with the theoretical assumptions of SLM-r and previous pronunciation training research, motivated the development of a four-phase, eight-step technology-supported perception–production instructional model. The model was designed to address production-heavy pronunciation instruction in which learners repeatedly imitate sounds without sufficient diagnostic perception training or systematic follow-up review.

Phase 1—Pre-class diagnosis. Learners (a) watched short pronunciation videos and completed perception discrimination and read-aloud tasks through the mini-program; the teacher then (b) reviewed the diagnostic results and adjusted instructional sequence, practice duration, and target difficulty accordingly.

Phase 2—Classroom instruction and practice. The teacher (c) introduced pronunciation targets using class-level performance data, learner recordings, and acoustic/pitch visualizations, explaining phonological contrasts through auditory, visual, and articulatory cues. Learners then (d) engaged in balanced perception and production activities, including identification, discrimination, listen-and-act tasks, modeling, imitation, and controlled production practice.

Phase 3—Immediate reassessment and feedback. Learners (e) completed parallel post-lesson perception and production tasks; the platform and teacher subsequently (f) provided item-level performance results, corrective feedback, and individualized practice guidance.

Phase 4—Spaced review and follow-up. Learners (g) completed brief perception and production activities over the following seven days and (h) completed a follow-up test covering the same pronunciation targets in a reshuffled order.

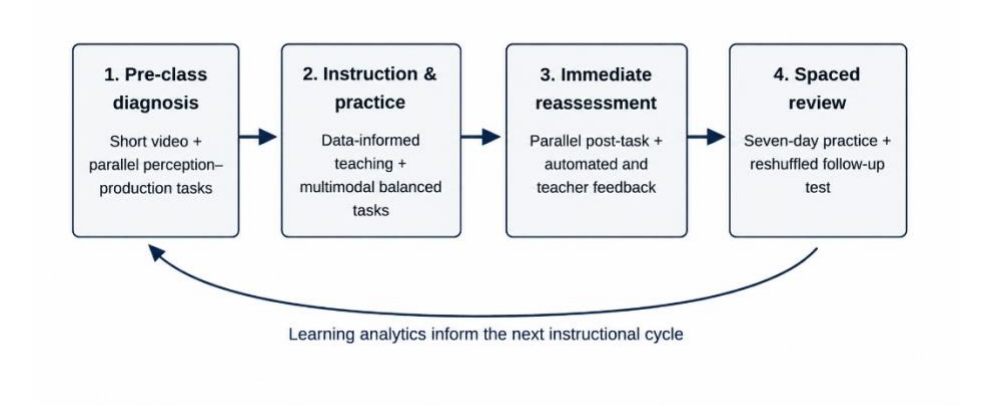


Figure 1. *The Four-Phase, Eight-Step Technology-Supported Perception–Production Instructional Model.*

The model conceptualizes digital tools as an infrastructure for diagnostic assessment, formative feedback, adaptive practice, and follow-up learning rather than as substitutes for teacher explanation or pedagogical mediation. The same learner evidence that supports individual learning also informs teachers' decisions regarding subsequent instructional adjustments.

In Study 2, Hanyin Yusheng served as the technological infrastructure for the intervention group by supporting diagnostic assessment, multimodal pronunciation practice, automated speech evaluation, learning records, and follow-up review. The digital functions complemented rather than replaced teacher explanation, feedback, and classroom interaction. Importantly, automated speech evaluation scores were used as formative feedback indicators rather than as independent outcome measures. The automated evaluation function was intended to help learners identify possible pronunciation deviations and support subsequent practice, while instructional decisions remained guided by teacher interpretation and classroom evidence.

Data Collection

The same teacher taught both groups through eight scheduled lessons over a four-week intervention period (360 minutes in total). The instructional content, learning objectives, and planned teaching time were held constant across groups. Target pronunciation features included the four Mandarin lexical tones, the Pinyin d–t aspiration contrast, the finals i, a, and u, and syllable blending. Both groups were expected to develop declarative knowledge of the target features and improve their accuracy in identifying and producing the target items.

The technology-enhanced group followed the four-phase perception–production instructional model. Before class, learners watched pronunciation videos and completed perception and production diagnostic tasks through the mini-program. The teacher used diagnostic results to identify learner difficulties and adapt instructional sequencing, practice duration, and target difficulty. Classroom activities included pronunciation modeling, discrimination tasks, appropriate acoustic and pitch visualizations, guided imitation, and balanced perception–production practice. After each lesson, learners completed parallel post-lesson tasks, received item-level

feedback, and engaged in seven days of brief follow-up practice before completing the delayed follow-up test.

The control group received the same pronunciation targets, learning objectives, total instructional time, and teacher. Conventional teacher-led instruction included explicit explanation, gesture-based demonstration, exaggerated modeling, contrastive examples, physical teaching aids, imitation, repetition, and corrective feedback. Compared with the technology-enhanced condition, the control condition did not involve mini-program-based diagnostics, automated speech evaluation, or the seven-day platform-guided practice cycle.

All participants completed the same pronunciation assessments at three time points: pretest, immediate posttest, and seven-day follow-up test. The same target items were used across the three assessments, with item order reshuffled in the follow-up test to reduce potential memory effects.

Data Analysis

A two-way mixed-design ANOVA was conducted to examine the effects of instructional condition and test occasion on learners' Mandarin pronunciation performance. Instructional group (technology-enhanced perception–production instruction vs. conventional instruction) was treated as a between-subject factor, whereas test occasion (pretest, immediate posttest, and seven-day follow-up test) was treated as a within-subject factor.

Before the analysis, descriptive statistics were calculated to summarize learners' pronunciation accuracy across groups and test occasions. The ANOVA examined the main effects of instructional group and test occasion, as well as the interaction effect between group and test occasion. The interaction effect was of particular interest because it indicated whether learners receiving technology-enhanced perception–production instruction demonstrated a different pattern of pronunciation development compared with those receiving conventional instruction.

Effect sizes were reported using partial eta squared (ηp^2). When significant effects were identified, pairwise comparisons of marginal means were conducted to further examine differences across instructional groups and test occasions.

Results

To address Research Questions 2 and 3, Study 2 examined changes in learners' composite pronunciation scores across instructional conditions and test occasions. Descriptive statistics were first calculated to summarize learners' pronunciation performance across the three assessment points, followed by a mixed-design ANOVA to examine the effects of instructional group, test occasion, and their interaction.

As shown in Table 3, both groups demonstrated continuous improvement across the three assessment occasions. The technology-enhanced perception–production group achieved higher mean scores than the control group at pretest, immediate posttest, and follow-up. Specifically, the technology-enhanced group increased from 2.13 (SD = 0.62) at pretest to 4.88 (SD = 0.89) at immediate posttest and further to 5.69 (SD = 0.48) at follow-up. The control group increased from 1.56 (SD = 0.89) at pretest to 3.63 (SD = 1.26) at immediate posttest and 4.81 (SD = 0.66) at follow-up.

The technology-enhanced group showed a gain of 2.75 points from pretest to immediate posttest and 3.56 points from pretest to follow-up, whereas the

corresponding gains for the control group were 2.07 and 3.25 points, respectively. Although the technology-enhanced group demonstrated higher descriptive gains, the differences in improvement patterns were relatively modest.

Table 3. *Composite pronunciation scores by group and test occasion*

Group	n	Pretest M (SD)	Immediate posttest M (SD)	Follow-up M (SD)
Technology-enhanced	16	2.13 (0.62)	4.88 (0.89)	5.69 (0.48)
Control	16	1.56 (0.89)	3.63 (1.26)	4.81 (0.66)

Note. Values are means and standard deviations of the composite pronunciation score.

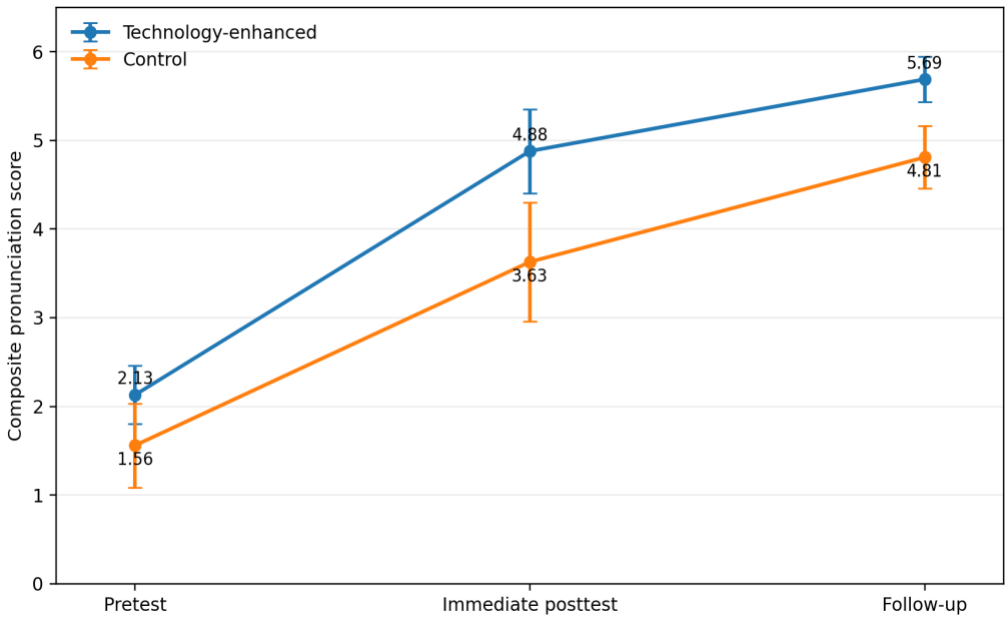


Figure 2. *Mean composite pronunciation scores by instructional group and test occasion. Error bars are omitted because standard errors for model-estimated means were not retained in the source file.*

Mixed -design ANOVA

A two-way mixed-design ANOVA was conducted to examine whether pronunciation development differed according to instructional condition and test occasion. Instructional group was treated as a between-subject factor, whereas test occasion was treated as a within-subject factor.

The analysis revealed a significant main effect of instructional group, $F(1, 30) = 21.821$, $p < .001$, $\eta p^2 = .421$. Averaged across the three assessment occasions, learners receiving technology-enhanced perception–production instruction achieved

significantly higher pronunciation scores than those receiving conventional instruction.

A significant main effect of test occasion was also observed, $F(2, 60) = 160.883$, $p < .001$, $\eta p^2 = .843$, indicating that learners' pronunciation performance improved significantly across the intervention period and follow-up assessment.

However, the interaction effect between instructional group and test occasion was not statistically significant, $F(2, 60) = 1.555$, $p = .222$, $\eta p^2 = .049$. This result indicates that although the technology-enhanced group achieved higher overall scores, the two groups did not demonstrate significantly different patterns of change across the three assessment occasions.

Table 4. *Mixed-design ANOVA of composite pronunciation scores*

Effect	df	F	p	ηp²
Group	1, 30	21.821	< .001	.421
Test occasion	2, 60	160.883	< .001	.843
Group × Test occasion	2, 60	1.555	.222	.049

Given the intact-class assignment and descriptive baseline differences, the significant group main effect should be interpreted cautiously. Future research should employ randomized assignment or baseline-adjusted models to provide stronger evidence for comparative effectiveness.

Pairwise Comparisons

Pairwise comparisons of marginal means were conducted to further examine differences associated with instructional group and test occasion. As shown in Table 5, the technology-enhanced group scored significantly higher than the control group when averaged across test occasions (mean difference = 0.896, $p < .001$).

Regarding test occasion, pronunciation scores significantly increased from pretest to immediate posttest (mean difference = −2.406, $p < .001$), from pretest to follow-up (mean difference = −3.406, $p < .001$), and from immediate posttest to follow-up (mean difference = −1.000, $p < .001$). These results indicate sustained improvement across the intervention and follow-up period.

However, because the group-by-test occasion interaction was not significant, these pairwise comparisons should be interpreted as evidence of overall group differences and temporal improvement rather than evidence that the technology-enhanced condition produced a significantly greater rate of improvement than conventional instruction.

Table 5. *Pairwise comparisons of marginal group and test-occasion effects*

Contrast	Mean difference	SE	t	p	Cohen's d
Technology-enhanced – control	0.896	0.171	5.243	< .001	1.070
Pretest – immediate posttest	−2.406	0.209	−11.498	< .001	−2.875
Pretest – follow-up	−3.406	0.209	−16.277	< .001	−4.069
Immediate posttest – follow-up	−1.000	0.209	−4.779	< .001	−1.195

Overall, Study 2 demonstrated that learners' pronunciation performance improved significantly over time and that the technology-enhanced perception–production group achieved higher overall performance than the conventional instruction group. Nevertheless, the nonsignificant group-by-test occasion interaction suggests that the present findings provide evidence for the feasibility and pedagogical potential of the technology-supported perception–production model rather than definitive evidence of its causal superiority over conventional instruction.

Discussion

Coordinating perception and production

The present study provides empirical support for a close relationship between Mandarin pronunciation perception and production. The strong positive association identified in Study 1 is consistent with the revised Speech Learning Model (SLM-r), which conceptualizes additional-language phonetic development as a dynamic process shaped by experience, input distribution, and interactions between perceptual and productive systems (Flege & Bohn, 2021). The finding also accords with previous Mandarin pronunciation research showing that overall identification accuracy and production performance can be associated, although the strength of correspondence may vary across phonological features and task conditions (Hao, 2018; Leung & Wang, 2020). Rather than suggesting that perception and production represent identical abilities, the present findings indicate that they are interconnected but partially distinct dimensions of second-language speech learning.

At the same time, the results caution against a simple perception-first account in which accurate perception necessarily precedes successful production. In Study 1, production accuracy was descriptively higher than perception accuracy across all target domains. Such differences may reflect task format, response demands, rating sensitivity, or the possibility that controlled production tasks capture emerging articulatory routines before stable perceptual representations are fully established. This interpretation is consistent with previous training research demonstrating that pronunciation learning can be modality-specific and that improvement in one channel does not automatically transfer to the other (Li & DeKeyser, 2017). Therefore, effective pronunciation instruction should not treat perception and production as separate stages but should instead facilitate iterative movement

between the two: learners identify a contrast, attempt production, receive feedback, revise their output, and return to perception with increased awareness.

The four-phase instructional model developed in this study operationalizes such an iterative perception–production cycle. Pre-class diagnosis makes individual difficulties visible before instruction; classroom activities involving discrimination, articulatory explanation, visualization, and controlled production direct attention toward relevant phonetic features; immediate reassessment creates a short feedback loop; and spaced follow-up practice provides repeated exposure to the same targets. From a theoretical perspective, the model extends SLM-r-informed pedagogy by increasing opportunities for meaningful input, improving the quality of phonetic feedback, and allowing learner performance patterns to inform subsequent instructional decisions.

What the classroom findings do—and do not—show

Study 2 demonstrated that both instructional conditions were associated with substantial improvement in Mandarin pronunciation performance. This pattern is understandable given the focused four-week intervention period, during which both groups received systematic instruction, explicit modeling, repeated practice, and corrective feedback from the same teacher. The significant main effect of test occasion therefore indicates that learners in both instructional conditions improved over the study period and is consistent with broader evidence that structured pronunciation instruction can support second-language speech development and is consistent with previous meta-analytic evidence showing that explicit pronunciation teaching can facilitate second-language speech development (Amrate & Tsai, 2025; Lee et al., 2015; Saito & Plonsky, 2019).

The technology-enhanced group achieved a higher overall mean score across the three assessment occasions. This pattern is consistent with previous findings suggesting that visual feedback, automated speech evaluation, and technology-supported practice may create additional opportunities for pronunciation learning (Chen, 2024; Cheng et al., 2025; Ngo et al., 2024; Xodabande et al., 2025; Zou et al., 2023a). Several mechanisms may explain this pattern. First, diagnostic analytics may help teachers identify learner difficulties and allocate instructional time more efficiently. Second, the platform extends opportunities for individual practice and feedback beyond classroom constraints. Third, the integration of listening and speaking activities may help learners notice discrepancies between intended phonological categories and their own productions. Finally, predictable follow-up practice may support sustained engagement. Previous research showing that AI-supported interaction can enhance learner engagement (Wang & Xue, 2024), that enjoyment contributes to engagement in robot-assisted Chinese-language learning contexts (Zhang & Yan, 2025), and that multimodal GenAI can shape engagement in Chinese-language learning (Wang, 2026) provides plausible motivational context, although these mechanisms were not directly examined in the present study.

However, the nonsignificant group-by-time interaction is a critical finding that requires cautious interpretation. Although the technology-enhanced group achieved higher mean scores at all three assessment occasions, the two groups did not demonstrate statistically different developmental trajectories. Furthermore, the technology-enhanced group showed a higher descriptive pretest mean than the control group (2.13 vs. 1.56), suggesting that initial group differences may have

contributed to the significant main effect of instructional group. Therefore, neither the significant group effect nor the significant improvement over time alone can be interpreted as evidence that the instructional model produced superior pronunciation gains.

The most appropriate interpretation is that the technology-supported perception–production model was feasible and pedagogically promising, and that learners using the model achieved higher overall performance within this pilot context. However, the present quasi-experimental design does not establish incremental causal superiority over high-quality conventional instruction. The lack of a significant interaction effect may be related to several factors, including the strength of the comparison condition, limited intervention duration, relatively small sample size, baseline differences between groups, or the sensitivity of the composite pronunciation measure.

This cautious conclusion remains theoretically meaningful. Technology-enhanced pronunciation instruction may not necessarily outperform well-designed teacher-led instruction within a short intervention period, particularly when both conditions provide intensive practice and feedback. Previous research has similarly shown that the effectiveness of automated speech feedback depends on treatment duration, feedback characteristics, and learner involvement (Ngo et al., 2024). Moreover, automated feedback should be viewed as complementary to, rather than a replacement for, teacher and peer mediation (Dai & Wu, 2023). In the present study, the control condition was an active instructional condition involving modeling, gesture-based explanation, contrastive demonstration, repetition, and correction. The relatively similar developmental trajectories may therefore reflect the effectiveness of both instructional approaches.

Technology as a feedback ecology

The value of the model does not primarily come from the platform itself. Rather, it comes from how learner evidence, automated feedback, and teacher decisions are connected during pronunciation instruction. Hanyin Yusheng connects multiple sources of learning evidence, including learners' perceptual responses, recorded productions, automated speech evaluation, and teacher observation. These sources provide complementary rather than interchangeable information. Automated evaluation can increase feedback frequency and practice opportunities, but it cannot independently explain the phonetic basis of an error or determine the most appropriate pedagogical response. Teachers remain essential in interpreting learner performance, connecting feedback to articulatory and acoustic features, and deciding when individual or class-level intervention is required.

This perspective moves beyond viewing technology as merely a delivery channel and instead conceptualizes digital tools as mediational systems embedded within instructional decision-making. It aligns with recent discussions of actionable automated pronunciation feedback and learner-mediated feedback processing (Huang & Pan, 2026; McCrocklin & Levis, 2025), evidence that AI-mediated interaction can support speaking development when it creates meaningful opportunities for practice (Fathi et al., 2024), and early computer-assisted pronunciation research emphasizing the importance of integrating technological affordances with pedagogical principles (Neri et al., 2002).

This perspective moves beyond viewing technology as merely a delivery channel

and instead conceptualizes digital tools as mediational systems embedded within instructional decision-making. It aligns with previous calls for richer and more actionable feedback mechanisms in AI-based pronunciation learning environments (Zou et al., 2023a; Wang et al., 2025), evidence that AI-mediated interaction can support speaking development when it creates meaningful opportunities for practice (Fathi et al., 2024), and early computer-assisted pronunciation research emphasizing the importance of integrating technological affordances with pedagogical principles (Neri et al., 2002).

The feedback ecology perspective also highlights learner variability. Recent research indicates that learners may experience AI-enhanced learning environments differently depending on their emotional responses, autonomy, and engagement profiles (Liu et al., 2025). Similarly, Qin and Dong (2025) demonstrated that learners' perceptions of AI-mediated feedback may not always correspond directly with their actual behavioral responses. These findings suggest that adaptive pronunciation instruction should not only monitor whether feedback is delivered but also examine how learners interpret and act upon feedback. In the present model, the most important affordance of technology may therefore be its capacity to make learning patterns visible over time and transform these patterns into actionable instructional decisions.

Multimedia components should be understood within the same framework. Short videos, pitch visualizations, auditory models, and learner recordings can enhance attention to pronunciation targets only when they are pedagogically aligned. Previous research has shown that audiovisual feedback can facilitate Mandarin tone learning (Chen, 2024), while learner characteristics influence how multimedia information is processed and retained (Teng, 2022, 2023). Therefore, multimodal resources should not be incorporated simply to increase technological complexity; rather, each representation should serve a clear learning purpose, such as linking pitch contours to perceptual judgments or connecting acoustic information to production attempts.

Theoretical and empirical contributions

The present study contributes to digital pronunciation research by extending technology-supported pronunciation instruction to Mandarin as an additional language. Much previous computer-assisted pronunciation research has focused on English segmental features or broad speaking outcomes. Mandarin pronunciation requires simultaneous attention to lexical tone, aspiration contrasts, vowel quality, and syllable-level coordination. By integrating lexical tones, the Pinyin d–t aspiration contrast, and three finals within a unified instructional cycle, the present study demonstrates a practical approach for coordinating segmental and suprasegmental pronunciation development.

The study also provides a transferable pedagogical routine for technology-enhanced pronunciation teaching. The innovation does not depend on the specific software architecture of Hanyin Yusheng. Instead, the underlying principle is the alignment of diagnostic assessment, classroom instruction, formative feedback, and delayed review. Instructors using different digital environments, including learning management systems, mobile applications, or institutionally approved speech tools, may adopt similar routines by collecting brief learner responses, identifying emerging difficulties, supporting balanced perception–production activities, and implementing follow-up practice.

More broadly, this study suggests that the educational value of AI-supported pronunciation learning lies not simply in automated scoring accuracy but in how technology reorganizes relationships among learners, teachers, and feedback. By positioning digital tools as supportive infrastructures rather than replacements for teachers, the present model offers a theoretically grounded and practically adaptable pathway for integrating artificial intelligence into Mandarin pronunciation pedagogy.

Pedagogical Implications

First, pronunciation courses should diagnose perception and production jointly. A learner who can imitate an item may still fail to identify it in unfamiliar speech; one who categorizes accurately may lack stable articulatory control. Parallel tasks help teachers distinguish these profiles.

Second, automated feedback should be specific, revisable, and interpreted with teacher involvement. Learners need to know which feature deserves attention and what to do next. A useful sequence is to identify the likely contrast, replay or visualize a model, attempt production, inspect automated output, and seek teacher or peer confirmation when results remain unstable.

Third, analytics should lead to instructional decisions. Class-level error frequencies can determine sequencing, time allocation, and grouping, while individual records can trigger targeted rather than uniform repetition. Such evidence-based use respects the individual variation foregrounded by SLM-r and the heterogeneous AI-learning profiles reported by Liu et al. (2025).

Fourth, follow-up measurement should be built into instruction. Immediate posttests may capture short-term familiarity. Delayed or spaced assessment better indicates whether categories remain available after a lesson. Future studies should include novel and untrained items to test generalization rather than merely recycling trained syllables.

Conclusion

This study developed and examined a technology-supported perception–production instructional model for Mandarin pronunciation through two connected investigations. Study 1 identified a strong positive association between pronunciation perception and production among intermediate Mandarin learners, providing empirical support for coordinating these two dimensions while recognizing their partially independent developmental characteristics. Study 2 further evaluated the classroom implementation of the model through a quasi-experimental pilot. Learners' pronunciation performance improved significantly over time, and the technology-enhanced group achieved higher overall scores than the conventional instruction group. However, the nonsignificant group-by-time interaction indicates that the present findings do not provide sufficient evidence that the model produced significantly greater developmental gains than high-quality conventional instruction.

The study contributes to technology-enhanced pronunciation research by demonstrating how perception and production can be integrated within an SLM-r-informed instructional framework. It further illustrates how diagnostic listening, guided production, automated feedback, teacher interpretation, learning analytics, and spaced review can be organized into a coherent pronunciation learning cycle.

Rather than positioning technology as a replacement for teachers, the study highlights the value of digital tools as mediational infrastructures that support evidence-informed instructional decisions and sustained learner engagement.

Several limitations should be acknowledged. First, Study 2 was based on a small intact-group quasi-experimental design, with one intact class assigned to each condition. Consequently, class-level characteristics were confounded with instructional condition and cannot be separated statistically in the present design. Future research should employ larger samples with multiple classes per condition, randomized or matched-group designs, and baseline-adjusted analyses to strengthen causal inference. Second, the technology-enhanced group completed additional platform-based activities outside class, whereas the control group did not follow an equivalent seven-day platform-guided practice cycle. Unless out-of-class practice time was matched and recorded, the present comparison should be understood as one between instructional packages that differed in both technological support and opportunities for additional practice; the contribution of the platform itself cannot therefore be isolated. Third, the present study primarily examined trained pronunciation targets; future work should include novel-item transfer, connected speech, and human-rated communicative outcomes to evaluate broader learning effects. Fourth, because the intervention combined multiple components, future studies should further disentangle the contributions of automated feedback, adaptive instructional planning, and spaced practice. Finally, future research should incorporate learner affective and behavioral variables, such as engagement, anxiety, motivation, and perceptions of AI-mediated feedback, to better understand how and for whom technology-supported pronunciation instruction produces meaningful benefits.

Funding:

The study reported in this paper represents a contribution to “2027 Special Research Project for Building a Strong Education Province in Henan Province: A Study on the Assessment and Promotion Strategies of Cultural Consciousness Literacy Among Primary and Secondary School Students in Henan Province” and “2027 Special Research Project for Building a Strong Education Province in Henan Province: Research on Digital Collection and Living Inheritance Mechanism of Henan Dialect Speech Resources” sponsored by Henan Provincial Department of Education.

Informed Consent Statement:

Informed consent to participate was obtained from all individual participants included in the study.

Data Availability Statement:

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest:

The authors declare that they have no competing interests.

Acknowledgments：

This work was supported by North China University of Water Resources and Electric Power. The university had no role in the design or implementation of this study. The authors are grateful for the insightful comments offered by the editor and the anonymous reviewers.

Author Contributions：

All listed authors materially participated in the research and manuscript preparation and approved the final version of the article.

References

Ahn, T. Y., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology, 47*(4), 778–786. https://doi.org/10.1111/bjet.12354

Aly, S. M. A., & Alsehibany, R. A. (2025). Integrating AI-powered tools in EFL pronunciation instruction: Effects on accuracy and L2 motivation. *Computer Assisted Language Learning.* Advance online publication. https://doi.org/10.1080/09588221.2025.2534015

Amrate, M., & Tsai, P.-H. (2025). Computer-assisted pronunciation training: A systematic review. *ReCALL, 37*(1), 22–42. https://doi.org/10.1017/S0958344024000181

Ballance, O. J. (2024). Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation. *ReCALL, 36*(1), 58–71. https://doi.org/10.1017/S0958344023000162

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). "Look, I can speak correctly": Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning, 37*(5–6), 1335–1363. https://doi.org/10.1080/09588221.2022.2080230

Chen, M. (2024). Computer-aided feedback on the pronunciation of Mandarin Chinese tones: Using Praat to promote multimedia foreign language learning. *Computer Assisted Language Learning, 37*(3), 363–388. https://doi.org/10.1080/09588221.2022.2037652

Cheng, B., Liao, K., Xiang, Y., Zou, Y., Zhang, X., & Zhang, Y. (2025). Development and validation of an AI-enhanced multimodal training program: Evidence from non-native Mandarin tone learning. *Computer Assisted Language Learning*, 1-25. https://doi.org/10.1080/09588221.2025.2571696

Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning, 36*(5–6), 861–884. https://doi.org/10.1080/09588221.2021.1952272

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System,* 121, 103254. https://doi.org/10.1016/j.system.2024.103254

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Hao, Y.-C. (2018). Contextual effect in second language perception and production of Mandarin tones. Speech Communication, 97, 32–42. https://doi.org/10.1016/j.specom.2017.12.015

Huang, X., & Pan, P. (2026). From feedback appraisal to pronunciation performance: A longitudinal study of learner responses to artificial intelligence-based pronunciation feedback. *International Journal of Applied Linguistics.* Advance online publication. https://doi.org/10.1111/ijal.70296

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Leung, K. K. W., & Wang, Y. (2020). Production–perception relationship of Mandarin tones as revealed by critical perceptual cues. *The Journal of the Acoustical Society of America, 147*(4), EL301–EL306. https://doi.org/10.1121/10.0000963

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition, 39*(4), 593–620. https://doi.org/10.1017/S0272263116000358

Liu, S. H. J. (2026). Effects of voice-based ChatGPT on Chinese-as-a-second-language learners' Mandarin pronunciation. *International Journal of Applied Linguistics.* Advance online publication. https://doi.org/10.1111/ijal.70213

Liu, H., Fan, J., & Xia, M. (2025). Exploring individual's emotional and autonomous learning profiles in AI-enhanced data-driven language learning: An expanded SOR perspective. *Learning and Individual Differences, 123,* 102753. https://doi.org/10.1016/j.lindif.2025.102753

McCrocklin, S., & Levis, J. (2025). Automatic speech recognition and pronunciation learning. Language Teaching, 1–17. https://doi.org/10.1017/S0261444825100852

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System, 57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy–technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441–467. https://doi.org/10.1076/call.15.5.441.13473

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL, 36*(1), 4–21. https://doi.org/10.1017/S0958344023000113

Qin, L., & Dong, J. (2025). EFL learners' perceptual perezhivaniya and actual writing revision behaviors mediated by GenAI: A sociocultural theory perspective. *System, 133*, 103769. https://doi.org/10.1016/j.system.2025.103769

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69*(3), 652–708. https://doi.org/10.1111/lang.12345

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology, 14*, 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

Teng, M. F. (2022). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System*, *105*, 102736. https://doi.org/10.1016/j.system.2022.102736

Teng, M. F. (2023). The effectiveness of multimedia input on vocabulary learning and retention. *Innovation in Language Learning and Teaching, 17*(3), 738–754. https://doi.org/10.1080/17501229.2022.2131791

Wang, Y. (2026). Exploring student engagement with multimodal generative AI in task-based Chinese language learning. *Computer Assisted Language Learning.* Advance online publication. https://doi.org/10.1080/09588221.2026.2645428

Wang, Q., & Wang, Y. (2026). Predictors of Self-Regulated Learning in AI-Supported Language Education: A Social Cognitive Theory Perspective. *European Journal of Education*, *61*(3), e70824. https://doi.org/10.1111/ejed.70824

Wang, Y., Derakhshan, A., & Ghasvand, F. (2025). EFL teachers' generative artificial intelligence (GenAI) literacy: A scale development and validation study. *System, 133*, 103791. https://doi.org/10.1016/j.system.2025.103791

Wang, Y., & Wang, Y. (2026b). Explaining learning engagement in AI-assisted learning through teacher support and achievement goals: Insights from goal-setting theory. *Humanities and Social Sciences Communications, 13*(1), 1215. https://doi.org/10.1057/s41599-026-02872-8

Wang, Y., & Xue, L. (2024). Using AI-driven chatbots to foster Chinese EFL students' academic engagement: An intervention study. *Computers in Human Behavior, 159*, 108353. https://doi.org/10.1016/j.chb.2024.108353

Wang, Y.-H., & Young, S. S.-C. (2015). Effectiveness of feedback for enhancing English pronunciation in an ASR-based CALL system. *Journal of Computer Assisted Learning, 31*(6), 493–504. https://doi.org/10.1111/jcal.12079

Wu, J. G., Lee, S.-M., Zou, D., & Xian, J. (2025). A decade of innovation in language education: Insights from eight leading CALL-related journals. *Applied Language Sciences*, *1*(1). https://doi.org/10.65553/ALS.250102

Xodabande, I., Shiri, S., & Zohrabi, M. (2025). Exploring the impacts of an AI-driven instructional intervention on Iranian EFL learners' pronunciation skill development. *Discover Education*, *4*(1), 307. https://doi.org/10.1007/s44217-025-00782-2

Zhang, J., & Yan, Y. (2025). The power of grit and enjoyment on engagement of students majoring in teaching Chinese to speakers of other languages (TCSOL) in a robot-assisted language learning environment: A self-determination theory perspective. *Learning and Motivation, 92*, 102185. https://doi.org/10.1016/j.lmot.2025.102185

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023a). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *SAGE Open, 13*(3), 1–14. https://doi.org/10.1177/21582440231193818

Zou, B., Guan, X., Shao, Y., & Chen, P. (2023b). Supporting speaking practice by social network-based interaction in artificial intelligence-assisted language learning. *Sustainability, 15*(4), 2872. https://doi.org/10.3390/su15042872

Zou, B., Li, H., & Li, J. (2018). Exploring a curriculum app and a social communication app for EFL learning. *Computer Assisted Language Learning, 31*(7), 694–713. https://doi.org/10.1080/09588221.2018.1438474

Zou, B., Lyu, Q., Han, Y., Li, Z., & Zhang, W. (2025). Exploring students' acceptance of an artificial intelligence speech evaluation program for EFL speaking practice: An application of the Integrated Model of Technology Acceptance. *Computer Assisted Language Learning, 38*(5–6), 1366–1391. https://doi.org/10.1080/09588221.2023.2278608

Author Bio

Liu Shi, Ph.D., is a Master's supervisor, Deputy Director and Lecturer in the Department of International Chinese Education, School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China. Her research interests include second language acquisition, international Chinese education, and digital technology-assisted language learning. Her recent works have appeared in SSCI-indexed journals. She has led a number of provincial- and ministerial-level research projects, including grants from the Center for Language Education and Cooperation of the Ministry of Education and the Henan Provincial Philosophy and Social Science Planning Project. She works as an anonymous reviewer for several international journals, such as *Learning and Motivation, SAGE Open,* and *European Journal of Education.*

E-mail: Shiliu910803@ncwu.cn

ORCID: 0009-0008-4361-0198

Yaping Yan, Ph.D., is an Associate Professor of Applied Linguistics. She received her master's and doctoral degrees from the Department of Chinese Language and Literature, Fudan University, Shanghai, China. Her research interests include modern Chinese grammar, rhetoric, and international Chinese education. She has published more than 60 papers in CSSCI source journals, including *Journal of World Chinese Language Teaching, Language Teaching and Linguistic Studies, Chinese Journal, Studies in the Chinese Language, Contemporary Rhetoric, Language Research,* and *Chinese Language Learning,* alongside numerous university journals. In recent years, she has published three academic monographs and co-authored two books. She has presided over several national-level research projects in China and received multiple national academic achievement awards.

E-mail: yanyaping@ncwu.edu.cn

ORCID: 0009-0005-5207-5252

Jingjing Xing, Ph.D. candidate and Research Assistant in the Department of Curriculum, Teaching and Learning, University of Toronto, Canada. She holds an

MA in Language and Literacy Education, an MSc in Intercultural Communication from the University of Warwick, and a BA (Hons) in International Communication Studies from the University of Nottingham Ningbo China. Her research focuses on critical multilingual pedagogy and social justice in ESL education. She has published in peer-reviewed journals and presented at conferences including the Conference on College Composition and Communication. Prior to her doctoral studies, she worked as a university lecturer in China, where she led provincial-level research projects and received excellence awards for teaching and research.

E-mail: jingjing.xing@mail.utoronto.ca

ORCID: 0009-0009-0414-2974

Jie Zhang, Ph.D., is a senior lecturer at the School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China. She has long been devoted to teaching Chinese as a foreign language and has undertaken Chinese language teaching assignments in India, Thailand and South Korea. Currently, she serves as Head of the Department of Chinese as a Foreign Language and has attained fruitful achievements in the field of teaching Chinese as a foreign language. Her publications on language teaching pedagogy, literary criticism, community education and other related themes have appeared in Chinese core journals. She has presided over several provincial-level research projects. Her research interests include international Chinese language teaching, language teaching pedagogy, teacher education, cross-cultural interpersonal factors in educational psychology, and intercultural communication.

E-mail: Zhangjie1@ncwu.edu.cn

ORCID: 0009-0002-6915-4924