RESEARCH ARTICLE

Videogame-based phonetic training for the perception of American English vowels by native Greek listeners

[version 1; peer review: awaiting peer review]

Eleni Vlahou[1], Aaron Seitz[2], Norbert Kopčo[3]

[1]Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
[2]Department of Psychology, Northeastern University, Boston, USA
[3]Institute of Computer Science, P. J. Šafárik University in Košice, Košice, Slovakia

v1 **First published:** 08 Aug 2026, **6**:285
https://doi.org/10.12688/openreseurope.24222.1
Latest published: 08 Aug 2026, **6**:285
https://doi.org/10.12688/openreseurope.24222.1

Abstract

This study examined the effectiveness of a videogame-based phonetic training program for improving Greek listeners' perception of two American English vowel contrasts (/ɪ/—/iː/and/æ/—/ʌ/). The game implemented a high-variability phonetic training paradigm that exposed listeners to multiple voices and lexical items, with speech stimuli generated using synthetic voices. Training consisted of approximately 1.5 hours of gameplay conducted in both quiet and multitalker babble, at progressively lower target-to-masker ratios. Before and after training, identification accuracy was assessed in quiet and in babble. The experimental group showed clear performance gains, whereas a no-training control group showed no improvement. Learning was observed for both vowel contrasts, under both quiet and babble conditions. Partial generalization was observed to untrained voices, but not to untrained lexical items. The results suggest that brief game-based phonetic training with synthetic speech can effectively induce phonetic learning and limited generalization across both quiet and adverse listening conditions. Broader and more robust transfer may require more extended training and greater variability in training materials.

Keywords

nonnative speech perception, high-variability phonetic training, gamification, multitalker babble

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

This article is included in the Horizon Europe gateway.

Corresponding author: Eleni Vlahou (evlahou@gmail.com)

Author roles: Vlahou E: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Seitz A**: Conceptualization, Formal Analysis, Methodology, Validation, Writing – Review & Editing; **Kopčo N**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: E.V. was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Post Doctoral Researchers" (Project Number 00447). E.V. and N.K. were supported by EU HORIZON-MSCA-2022-SE-01 grant No. 101129903. N.K. was also supported by APVV projects APVV-23-0054 and SK-AT-23-0002.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Vlahou E, Seitz A and Kopčo N. **Videogame-based phonetic training for the perception of American English vowels by native Greek listeners [version 1; peer review: awaiting peer review]** Open Research Europe 2026, **6**:285 https://doi.org/10.12688/openreseurope.24222.1

First published: 08 Aug 2026, **6**:285 https://doi.org/10.12688/openreseurope.24222.1

1. Introduction

Accurate perception of nonnative speech sounds can be challenging for adult learners, especially when the target contrasts are absent from their native language. These difficulties often persist even after years of experience with a second language, impairing communication and limiting the attainment of second-language proficiency (Best & Tyler, 2007; Flege, 2003; McCandliss et al., 2002).

The most widely studied case of segmental difficulty involves Japanese learners' struggle to distinguish English /r/ and /l/, a phenomenon first examined over fifty years ago (Goto, 1971; Miyawaki et al., 1975) and still actively researched today (e.g., Saito et al., 2022; Shinohara and Iverson, 2021). Another well-known example is the difficulty non-Hindi speakers face in discriminating between dental and retroflex stops (Golestani & Zatorre, 2004; Werker & Tees, 1983). While less prominent in the literature, English vowel contrasts present a strong challenge for many nonnative listeners. Learners from languages with smaller vowel inventories frequently struggle to acquire the vowel distinctions found in English, which has a large and densely packed vowel space (Georgiou, 2019; Ladefoged & Johnson, 2011). Greek learners are no exception.

Modern Greek has a simple five-vowel system (/i e a o u/), with acoustically well-separated vowel categories (Arvaniti, 2007). In contrast, American English includes approximately three times as many vowels, with phonemes that occupy overlapping regions of the vowel space (Arvaniti, 2007; Ladefoged & Johnson, 2011). This mismatch often leads Greek listeners to assimilate multiple English vowels into a single native-language category (Georgiou, 2019; Iverson & Evans, 2007; Lengeris, 2009). Common examples include /ɪ/–/iː/ (*ship–sheep*) and /æ/–/ʌ/ (*lack–luck*), which are frequently confused and miscategorized (Georgiou, 2019; Lengeris, 2009; Lengeris & Hazan, 2010).

These perceptual challenges are typically interpreted through influential models of second-language speech perception including the Perceptual Assimilation Model, the Speech Learning Model and the Native Language Magnet model (Best, 1994; Best & Tyler, 2007; Flege, 2003; Flege & Bohn, 2021; Kuhl et al., 2008). The idea that language-specific phonological categories function as a "phonological sieve" through which nonnative sounds are filtered, can be traced back to early structural phonology (Trubetzkoy, 1969; see discussion in Flege, 2003). Contemporary models of second-language speech perception build on this intuition. The Native Language Magnet model, in particular, emphasizes the warping of perceptual space due to early exposure to ambient native-language input, leading to neural commitment to native sound patterns even before the end of the first year of life (Kuhl et al., 2008). This early specialization can explain the persistent difficulties that adults face in perceiving nonnative speech. However, not all contrasts are equally difficult. According to the Perceptual Assimilation Model, when two second-language sounds are assimilated into the same native-language category—a process known as single-category assimilation—discrimination is expected to be particularly poor (Best, 1994). This pattern may also occur in Greek learners of English, who appear to assimilate the /ɪ/–/iː/ and /æ/–/ʌ/ contrasts to their native /i/ and /a/, respectively (Georgiou, 2019). The Speech Learning Model, and its revised version (Flege, 2003; Flege & Bohn, 2021) highlight the role of age, experience, and linguistic input. While acknowledging age-related decline in phonetic plasticity, it rejects the idea of a strict critical period and maintains that adults retain the ability to acquire new phonetic categories. Together, these models help explain why certain vowel contrasts remain challenging even after extensive exposure to a second language, and why targeted phonetic training may be necessary to support phonetic learning.

It is well-documented that phonetic training can lead to significant perceptual improvements, even for challenging contrasts (Iverson et al., 2005; Lively et al., 1993; Logan et al., 1991; McCandliss et al., 2002). A widely used method is high-variability phonetic training, which emphasizes the use of multiple talkers and lexical items during training (Bradlow, 2008; Iverson et al., 2005; Lively et al., 1993; Logan et al., 1991; Shinohara & Iverson, 2021; Uchihara et al., 2025). In a seminal study, Lively et al. (1993) showed that training with multiple speakers promotes generalization to untrained voices, whereas training with a single voice often results in talker-specific representations. Subsequent studies have replicated and extended these findings across a range of nonnative contrasts and learner populations (Lengeris & Hazan, 2010; Leong et al., 2018; Uchihara et al., 2025; Wang et al., 1999).

In terms of training structure, studies have typically employed explicit identification and discrimination tasks, with immediate performance feedback (Bradlow, 2008). More recent work has explored alternative approaches, including implicit and unsupervised learning paradigms, as well as the use of gamification in phonetic training (Guevara-Rukoz et al., 2019; Lim et al., 2019; Lim & Holt, 2011; Saito et al., 2022; Vlahou et al., 2012; Vlahou et al., 2019; Wade & Holt, 2005). Gamification involves incorporating game-like elements such as visual feedback, points, or levels into non-game contexts to enhance engagement and motivation (Deterding et al., 2011; Hamari et al., 2014). Work by Holt and colleagues has demonstrated that game-based environments can produce perceptual gains comparable to, and in some cases exceeding, those from traditional methods (Lim & Holt, 2011; Wade & Holt, 2005). In this context, sound

categories are not explicitly taught; rather, learning occurs incidentally, as successful performance in the game depends on the ability to distinguish the relevant speech sounds (Lim et al., 2019; Lim & Holt, 2011; Saito et al., 2022; Wade & Holt, 2005).

A relatively less studied issue concerns the ecological validity of the listening environment during training. In everyday settings, listeners must cope with background noise, reverberation, and competing talkers, yet most phonetic training studies are conducted under idealized, quiet conditions. Some recent work has begun to address this gap. For example, Vlahou et al. (2019) found that gamified training with reverberant speech from various rooms improved implicit learning of a difficult Hindi contrast. Leong et al. (2018) showed that adaptive background noise during high-variability phonetic training produced rapid learning of English speech sounds, with improvements generalizing to untrained sounds and persisting six months later. Despite these promising findings, research on incorporating realistic adverse listening conditions into phonetic training is still at an early stage.

Parallel developments in text-to-speech synthesis and AI voice generation offer new opportunities for research in speech perception (e.g., King, 2014). Synthetic speech offers increased control over acoustic parameters and allows for rapid generation of large stimulus sets without the time-consuming processes required for natural recordings (Nuesse et al., 2019). Although early text-to-speech systems were limited by poor intelligibility, modern text-to-speech technologies produce speech that is comparable to natural recordings (King, 2014; Nuesse et al., 2019). Recent work further shows that participants perform equally well, or even slightly better, with synthetic speech in masked sentence recognition tests and are unable to reliably indicate whether the speech comes from a human or synthetic talker (Calandruccio et al., 2025). These findings are encouraging, but more work is needed to better understand the advantages and possible limitations of using synthetic speech in phonetic training.

Motivated by these converging lines of research, this study investigates whether a videogame-based high variability phonetic training program, with synthetic voices and both quiet and adverse listening conditions, can improve Greek learners' perception of two American English vowel contrasts: /ɪ/–/iː/ (e.g., *ship–sheep*) and /æ/–/ʌ/ (e.g., *lack–luck*). As noted above, these contrasts are often difficult for Greek listeners, due to the limited vowel inventory of their native language which lacks distinct phonemic counterparts for many English vowels (Arvaniti, 2007; Georgiou, 2019; Lengeris, 2009; Lengeris & Hazan, 2010).

Greece demonstrates very high English literacy, as reflected in its 8th-place global ranking in the EF English Proficiency Index (EF Education First, 2023). English is widely taught from an early age through formal education, private tutoring and cultural exposure to English via music and media. However, Greek learners often don't pay attention to English phonemic contrasts not used in their language. While some of them may perform well in quiet, focused listening tasks, we expected their accuracy to decline in fast-paced, noisy conditions—situations that better reflect real-world communication challenges.

To investigate these effects, participants completed a training program that followed a high variability phonetic training approach (Lively et al., 1993), with multiple words presented by four synthesized text-to-speech voices. The task was presented within a custom videogame (PhonemeQuest) that incorporated adaptive difficulty and visual and auditory reward elements. The game is publicly available as an open research prototype (see Data availability statement). As players advanced, background babble was introduced and gradually increased in intensity. Before and after training, participants were tested with trained and untrained words, produced by voices used during training, and with words produced by unfamiliar voices.

We hypothesized that participants who completed the videogame-based training would show significant gains in vowel identification accuracy from pre- to post-test, in both the quiet and the multitalker babble condition for the trained speakers and tokens. Given the high-variability input, we also predicted partial generalization to untrained words and voices, consistent with previous findings from high-variability phonetic training studies (Leong et al., 2018; Lively et al., 1993; Uchihara et al., 2025). In contrast, we expected no significant improvement in a no-contact control group who completed the same tests without any training.

2. Material and methods

2.1 Participants

Eighteen adult native speakers of Greek (9 female, 9 male; age range: 25–46 years) participated in the study. All reported normal hearing and no history of speech or language disorders. English proficiency ranged from upper-intermediate to advanced, based on years of formal instruction, language certifications (CEFR B2 or higher), and, in some cases, studies abroad. Participants were randomly assigned to either an experimental group (n = 9), which completed approximately

1.5 h of gamified phonetic training, or a control group (n = 9), which received no training but completed the same pre- and post-tests.

2.2 Stimuli

The speech stimuli consisted of 24 synthesized English words forming 12 minimal pairs (words differing by only one phoneme, e.g., *ran–run*), targeting two vowel contrasts known to be difficult for Greek learners: /ɪ-iː/ (e.g., *bit–beet*) and /æ-ʌ/ (e.g., *lack–luck*). The full word list is shown in Table 1.

All stimuli were generated using six synthesized text-to-speech voices representing adult native speakers of American English (three female), produced using a publicly available text-to-speech service (ttsfree.com). To ensure intelligibility, the stimuli were pre-tested in a forced choice identification task by a native speaker of American English, who achieved 100% accuracy, confirming that the materials were clear and unambiguous.

A multitalker babble masker was used in both training and testing, created from the BKB sentence corpus (Bench et al., 1979). Recordings from 12 different speakers were combined to form a single multitalker babble track. No additional level normalization was applied prior to mixing, as the original BKB recordings are relatively uniform in overall level. The resulting babble was approximately 1 minute in duration; when the available sentence material was exhausted, segments were repeated to reach the target length.

All speech stimuli and the final babble track were amplitude normalized to a fixed RMS level and saved in 44.1 kHz, 16-bit WAV format.

2.3 Procedure

All participants were informed about the nature of the study and gave written informed consent prior to participation.

This experiment followed a standard pre-test–training–post-test design. Participants in the experimental group completed an initial pre-test, three training sessions, and a final post-test, each lasting approximately 30 minutes, for a total experiment duration of about 2.5 hours. Participants in the control group completed the same pre- and post-tests without intervening training. Training sessions were conducted using PhonemeQuest, a Unity-based phonetic training video-game. All sessions, including pre-test, training, and post-test, were completed within a minimum of three and a maximum of five consecutive days.

2.3.1 Training

Figure 1B shows a visual schematic of the training procedure. Participants in the experimental group completed trials within a 3D endless-runner style videogame, in which a character advanced through a desert or green landscape. In each trial, two words forming a minimal pair (e.g., *nut–gnat*) were played in sequence, separated by a short interval. After audio playback, two response pairs appeared in the 3D environment, one positioned behind the other. Each pair contained two floating 3D objects (e.g., a cluster of nuts representing *nut* and a flying insect representing *gnat*), along with the corresponding word labels. Players had to select the correct item from each pair, in the correct order (e.g., first *nut*, then *gnat*), either by jumping through the object or passing underneath it. For training, each participant was exposed to six minimal pairs (three per contrast), with item sets counterbalanced across participants. Four voices (2 female), the same for all participants, were used during training.

Feedback was provided after each trial. A response was scored as correct only if the participant selected both words of the pair in the correct order, resulting in a reward sound, a visual effect and a + 1 score increase (Figure 1B4). All other response outcomes were logged in detail, including partial responses, incorrect selections and omissions at either response position, but were classified as incorrect. Task difficulty adapted over time; three consecutive correct responses led to a level increase, while any incorrect response led to a level decrease.

Table 1. Minimal pairs for the two target vowel contrasts.

Contrast	Minimal pairs
/ɪ/–/iː/	bit–beet; still–steel; ship–sheep; dip–deep; bin–bean; fit–feet
/ʌ/–/æ/	cup–cap; run–ran; luck–lack; fun–fan; bud–bad; nut–gnat

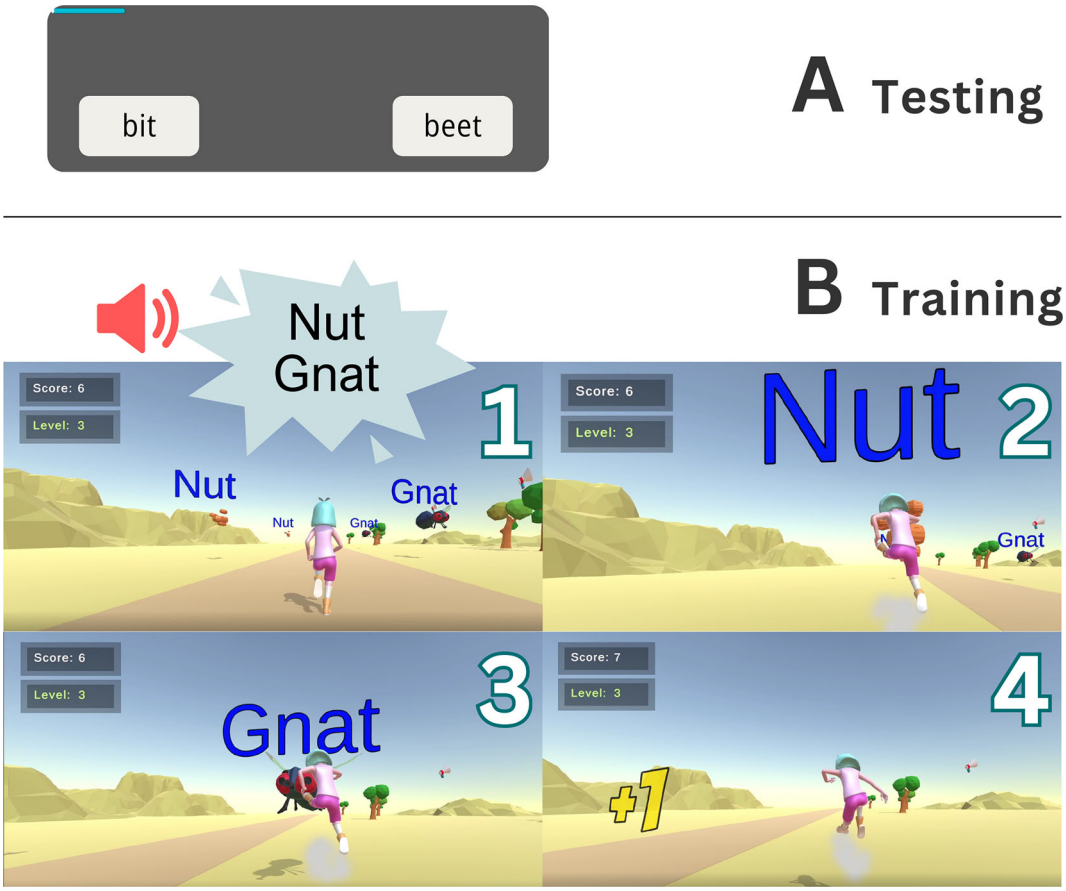

Figure 1. Task schematic of testing and training. A. Testing. In each trial, participants heard a single word from a minimal pair (e.g., *bit* or *beet*) and performed a two-alternative forced-choice task (2AFC) by selecting the correct matching word from two response options displayed on the left and right side of the screen. No feedback was provided during testing. **B. Training**. (1) Each trial began with the player running through the game environment. Two speech sounds (e.g., *nut–gnat*) were presented in sequence. Next, two response pairs appeared along the path, each consisting of the same minimal pair (e.g., *nut–gnat*). The first pair appeared a short distance ahead of the player, while the second farther along the path, behind the first pair. (2–3) The participant had to select the correct words in the correct order by navigating to the appropriate items. (4) Correct responses to both items resulted in a + 1 point score increase and a reward sound.

The game had 8 difficulty levels. As players progressed to higher levels, the game became more challenging along two dimensions. First, character movement speed increased gradually across levels, requiring faster perceptual decisions and responses. Second, multitalker babble was introduced at Level 6 at a target-to-masker ratio (TMR) of approximately +1.5 dB, and increased in intensity at subsequent levels (Level 7: −3.5 dB; Level 8: −8.5 dB). The babble was presented continuously throughout each trial and was not temporally synchronized with target word presentation.

Each training session consisted of two blocks, one for the /ɪ–iː/ contrast and one for the /æ-ʌ/ contrast, with block order counterbalanced across participants. Within blocks, speakers and items were randomized, and word order within each pair was counterbalanced (five presentations in each order). Each session included 240 trials (4 speakers × 2 contrasts × 3 minimal pairs × 10 repetitions).

2.3.2 Testing

All participants completed the same pre- and post-test sessions, which consisted of a two-alternative forced-choice (2AFC) identification task. On every trial, a single word from a minimal pair was presented through headphones (e.g., *bit*), and participants chose between the two corresponding written alternatives from that pair (e.g., *bit* vs. *beet*) (Figure 1A). No feedback was provided during testing.

Each phonetic contrast included six minimal pairs: the three that were used during training, plus three that served as untrained items. All synthetic voices were presented during testing, the four used in training plus two additional voices (one male and one female). Including untrained voices and items in testing allowed us to assess generalization of learning beyond the specific trained stimuli. Stimuli were presented in two listening conditions: quiet and multitalker babble at a −4 dB signal-to-noise ratio (SNR). Each test session followed a fixed order. First, trained voices were presented in quiet for one phoneme contrast (order counterbalanced), followed by the second contrast. The same sequence was then repeated with babble noise. In the final block, the two untrained voices were presented, first in quiet, then in babble. Trials within each block were randomized. Each pre- and post-test session included 576 trials (6 speakers × 2 phonetic contrasts × 2 listening conditions × 12 words per contrast × 2 repetitions).

2.3.3 Analysis

All analyses were conducted using mixed-effects models, with participants treated as random effects. We report two sets of analyses. First, we examine whether training led to improved performance by analyzing pre- and post-test identification accuracy. These analyses assess effects of condition (experimental vs. control), time (pre vs. post), and stimulus familiarity (trained vs. untrained speakers and items), as well as listening condition (quiet vs. masked) and phonetic contrast (/ɪ–iː/ and /æ–ʌ/).

Second, we analyze performance during training to assess learning dynamics across sessions, using changes in achieved difficulty level over time as the primary outcome measure.

Due to logging errors, one participant in the experimental group completed approximately 25% more pretest trials than intended (distributed evenly across conditions), one participant was missing a single post-test trial, and session-1 training data were not recorded for one participant (although the session was completed).

3. Results

3.1 Testing

To evaluate training effects, we analyzed accuracy from the 2AFC task administered at pre- and post-test (Figure 1A). In this task, participants heard a single word (e.g., *bit*) presented either in quiet or in multitalker babble and selected the matching written alternative from the corresponding minimal pair.

We fitted a generalized linear mixed-effects model (binomial link) implemented in lme4 (Bates et al., 2015). The model included fixed effects of condition (experimental, control), time (pre, post), speaker familiarity and item familiarity (familiarity coded as trained: speakers/words presented during both training and testing, untrained: speakers/words presented only during testing), with all two- and three-way interactions among these factors. We also included listening condition (speech presented in quiet or masked with multitalker babble) and phoneme contrast (/ɪ–iː/vs. /æ–ʌ/) as additive predictors. The model was specified as:

```
accuracy ~ condition * time * speaker * item + listen + contrast + (1|subj)
```

A maximal random-slopes structure resulted in near-perfect intercept–slope correlations, so a random-intercept model was adopted (Baayen et al., 2008; Bates et al., 2015; Matuschek et al., 2017). There was a main effect of Speaker ($\beta = 0.20$, $p = .036$), Listening Condition ($\beta = 0.15$, $p < .001$), and Contrast ($\beta = -0.16$, $p < .001$), as well as a significant Time × Condition interaction ($\beta = 0.27$, $SE = 0.11$, $z = 2.47$, $p = .014$). No other main effects or interactions reached significance (all $p > .09$).

Figure 2 illustrates the Time × Condition interaction: the experimental group showed higher post-test vs. pre-test accuracy, whereas the control group showed no change. Follow-up contrasts derived from the fitted model (computed on the logit scale using the emmeans package; Lenth, 2024) indicated a significant pre-post improvement in the experimental group ($\beta = 0.23$, $p < .001$), but not in the control group ($\beta = 0.02$, $p = .65$).

Although higher-order interactions involving speaker and item familiarity were not statistically significant, we conducted exploratory contrasts to examine whether this interaction differed across combinations of speaker and item familiarity.

The analysis showed that the experimental group improved for trained items from both trained speakers ($\beta = 0.38$, $p < .001$) and untrained speakers ($\beta = 0.34$, $p = .013$), whereas no corresponding improvement was observed in the control group (all $p > .40$). For untrained items, pre–post changes were not significant (all $p > .17$). Figure 3 illustrates these effects.

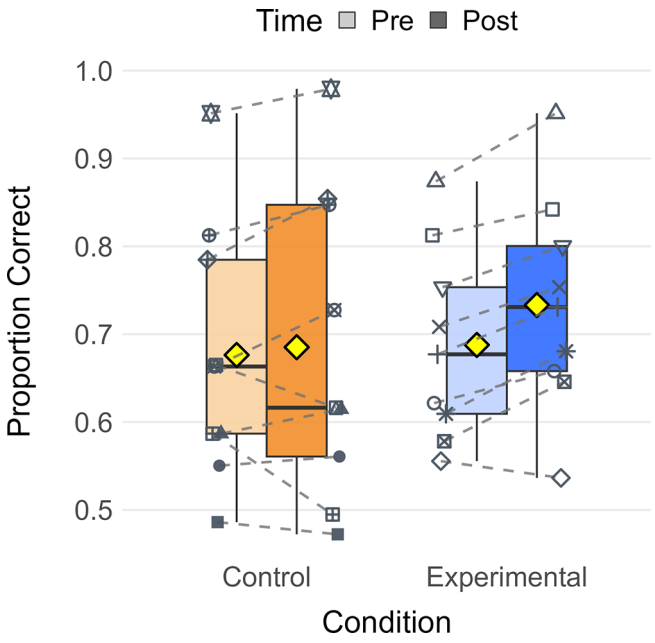

Figure 2. Pre- and post-test accuracy by group. Pre- and post-test accuracy in the 2AFC test for the experimental and control groups, averaged across listening conditions, phoneme contrasts, speakers, and items. Boxplots show the median (thick horizontal line) and interquartile range (box); whiskers extend to the most extreme values within 1.5 × the interquartile range. Diamonds indicate group means. Symbols are consistent across figures to identify the same participants.

The omnibus test showed that performance was higher in the quiet condition (70.9%) than in the masked condition (68.1%), as the multitalker babble degraded speech perception (Figure 4, left panel). The main effect of contrast shows that overall accuracy was lower for the /æ–ʌ/ pair (68%) than for /ɪ–iː/ (71.1%), suggesting that the former is more difficult for Greek listeners (Figure 4, right panel).

To confirm that the training effects were similar across the two vowel contrasts and listening conditions, we ran follow-up analyses using emmeans. There was a significant pre–post improvement for both contrasts (/ɪ–iː/, /æ–ʌ/) and for both listening conditions (quiet, masked; all $p < .0001$). The control group showed no pre–post changes in either contrast or listening condition (all $p > .65$). These effects are visualized in Figures 5 and 6.

Finally, the main effect of speaker (slightly higher accuracy for untrained voices) was unexpected. To explore whether it reflects intelligibility differences present before training, we examined pre-test accuracy only. The two untrained voices (05, 09) were already the most intelligible at baseline (0.70), exceeding all four trained voices (0.65–0.69; see Appendix A, Figure A1). A simple-effects contrast confirmed this difference (trained − untrained: $\beta = -0.14$, $SE = 0.047$, $z = -2.98$, $p = .003$).

3.2 Training

Improvement in performance during training can be quantified in several ways, such as changes in score, number of errors, or level progression. In this analysis, we focused on participants' average level across trials and sessions. Higher levels reflect successful performance under more challenging conditions, as they involved increased movement speed and stronger multitalker babble. If participants improved over time, they should reach higher difficulty levels more quickly in later sessions compared to earlier sessions and sustain them for more trials.

To examine this, trials in each session were grouped into bins of 80, and for each participant we calculated the mean level per bin and session. Figure 7A shows the average level per bin and session at the group-level. Visual inspection suggests that mean levels increased across sessions, and participants tended to reach higher levels earlier in later sessions. This pattern is further illustrated in Figure 7B, which shows the distribution of difficulty levels across bins for each session. Across sessions, higher levels (6–8) became more frequent and were reached earlier in the session.

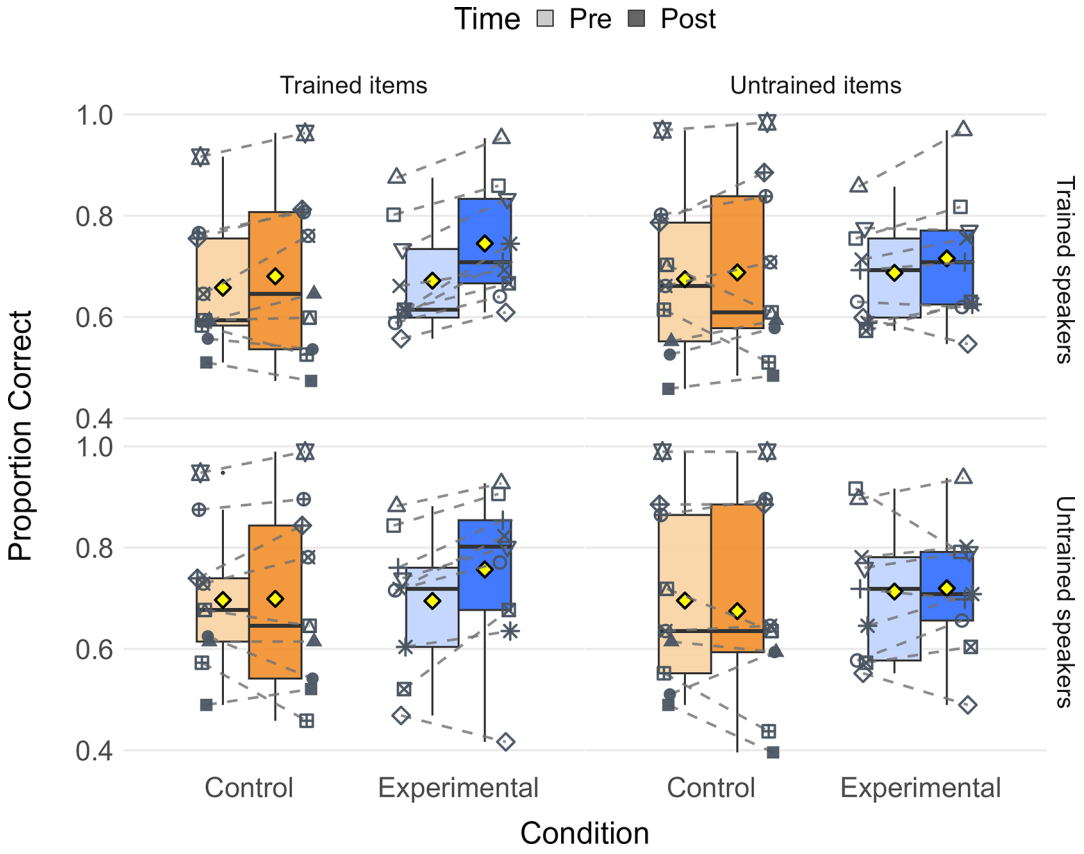

Figure 3. Pre- and post-test accuracy by group and familiarity. Pre- and post-test accuracy in the 2AFC identification task for the experimental and control groups as a function of item and speaker familiarity. Results are shown separately for trained and untrained speakers (rows) and items (columns), averaged across listening conditions and phoneme contrasts. Boxplots and participant symbols are as in Figure 2.

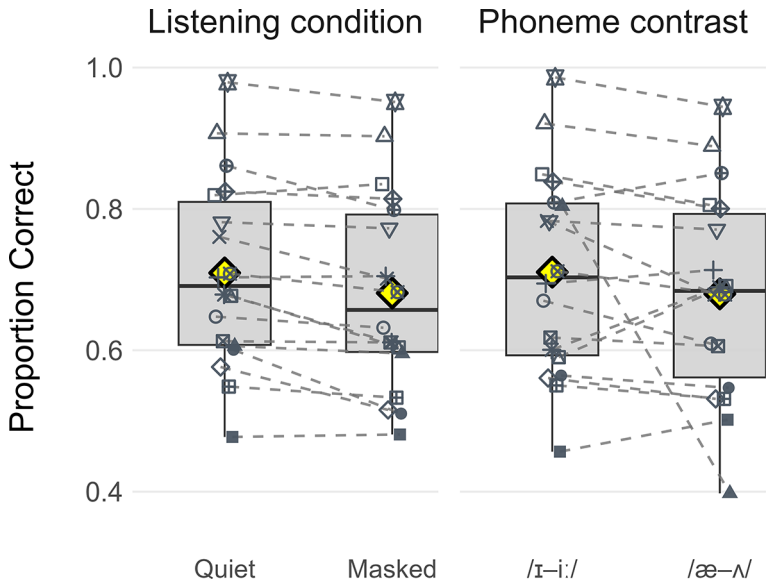

Figure 4. Accuracy by listening condition and vowel contrast. Accuracy in the 2AFC test as a function of listening condition and phoneme contrast, averaged across groups and time. Boxplots and participant symbols as before.

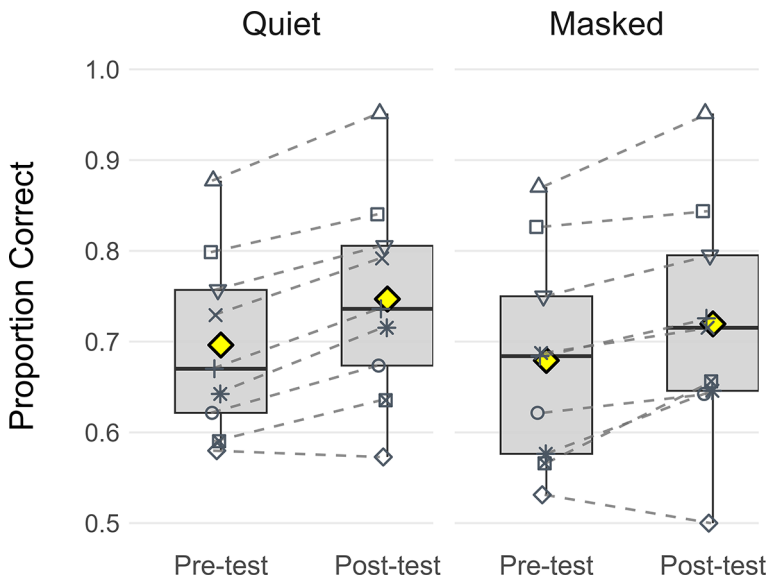

Figure 5. Pre- and post-test accuracy across listening conditions in the experimental group. Pre- and post-test accuracy for the experimental group in quiet and masked conditions. Boxplots and participant symbols as before.

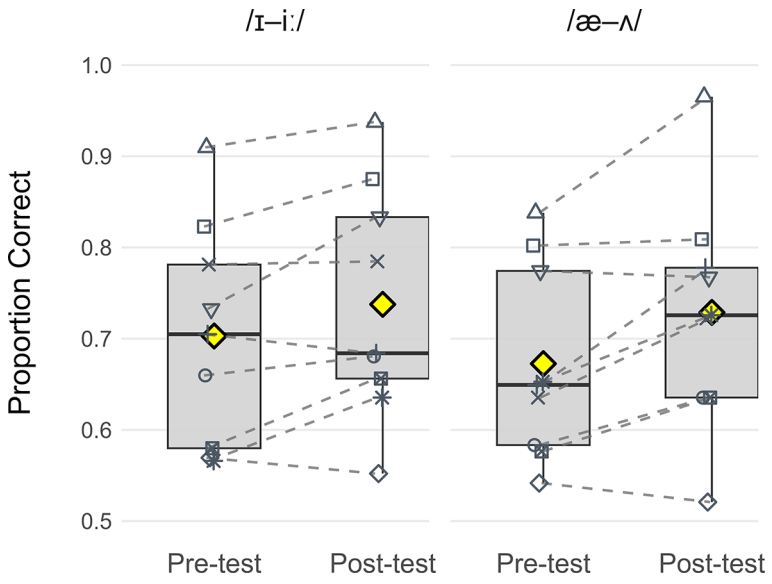

Figure 6. Pre- and post-test accuracy by vowel contrast in the experimental group. Pre- and post-test accuracy in the 2AFC test for the experimental group across vowel contrasts (/ɪ–iː/ and /æ–ʌ/). Boxplots and participant symbols as before.

To evaluate training-related improvement, a linear mixed-effects model was fitted in lme4 (Bates et al., 2015) predicting mean level as a function of Session (1–3) and Trial Bin (1–3), and their interaction, and random intercepts for participants:

```
mean_level ~ Session * Bin + (1|subj)
```

Fixed effects were evaluated using Type III F-tests with Satterthwaite degrees of freedom, and pairwise comparisons were obtained using estimated marginal means (emmeans; Lenth, 2024). There were significant main effects of Session, $F(2, 64) = 10.68$, $p < .001$, and Bin, $F(1, 64) = 10.52$, $p = .002$, and no Session × Bin interaction, $F(2, 64) = 0.06$, $p = .94$. Pairwise comparisons (Tukey-adjusted) confirmed increases from Session 1 to 2 ($p = .007$) and from Session 1 to 3 ($p < .001$), with no difference between Sessions 2 and 3 ($p = .32$).

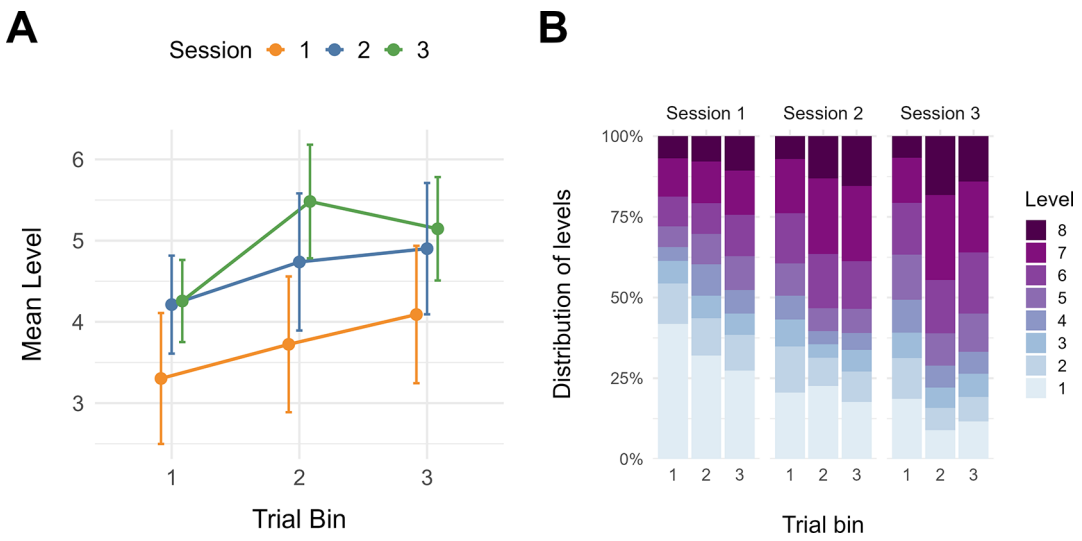

Figure 7. Training performance across sessions and difficulty levels. Training performance across sessions. (A) Mean difficulty level per session and trial bin (80 trials per bin). Error bars represent standard errors across participants. (B) Distribution of difficulty levels across trial bins for each training session (data combined across participants). Each stacked bar represents the proportion of trials completed at each level (1–8), with darker shades indicating higher difficulty levels.

4. Discussion

This study examined whether brief training with a videogame-based program could improve Greek learners' perception of two challenging American English vowel contrasts: /ɪ–iː/ and /æ-ʌ/. Participants completed about 1.5 hours of high-variability phonetic training, conducted in both quiet and multitalker babble conditions, across three sessions. At post-test, the experimental group showed clear improvement: accuracy increased by roughly 7% for trained items from trained speakers, with partial generalization to untrained voices. These gains were not observed in the control group. These findings are consistent with previous work showing that brief training can enhance nonnative segmental perception (e.g., Iverson et al., 2005; Lim and Holt, 2011; Vlahou et al., 2012; Vlahou et al., 2019) and offer initial support for the potential usefulness of the training program.

Exploratory analysis suggested that performance gains in the experimental group were not restricted to the specific voices used during training, indicating some degree of generalization. However, this transfer was only partial: listeners showed improvement for trained items spoken by untrained voices (+6%), whereas gains for untrained items spoken by trained voices were small (+3%), and no improvement was observed for untrained items produced by untrained speakers. The training incorporated high variability phonetic training, with four different voices and three word pairs per contrast. Such variability is thought to encourage learners to attend to stable, contrast-relevant cues rather than idiosyncratic features of individual voices or words (Lively et al., 1993; although some recent accounts suggest that learners may instead rely on simpler decision strategies, see Shinohara & Iverson, 2021). In this study, the amount of training was substantially smaller than in most high variability phonetic training studies, which often span several hours or multiple weeks (e.g., Lively et al., 1993; Pruitt et al., 2006; Shinohara and Iverson, 2021). With such limited exposure, the observed partial generalization is unsurprising. A potentially fruitful direction for future work is to introduce talker and item variability adaptively, gradually increasing diversity as performance on current words and talkers stabilizes (see e.g., Pruitt et al., 2006).

As expected, overall performance was poorer in the masked condition, confirming that multitalker babble substantially reduced performance. Importantly, however, there was learning for both contrasts across both listening conditions (see Figures 5 & 6). Figure 7 further shows that during training participants reached higher difficulty levels across sessions and maintained performance even as background babble increased and even dominated the signal (−8.5 dB TMR). The babble used here represents a demanding masker type, combining energetic and informational interference, both of which are known to affect non-native listeners more strongly than native listeners (García Lecumberri et al., 2010; Jafari et al., 2025; Leong et al., 2018). The −4 dB SNR used during masked testing is representative of many everyday listening situations, such as conversations in cafeterias, classrooms, or busy offices, where speech remains intelligible for native listeners but poses significant difficulty for nonnative listeners (García Lecumberri et al., 2010). Taken together, the

results suggest that training-related improvements can extend to acoustically challenging conditions similar to those encountered in everyday communication.

It should be noted that, in natural listening environments, normal-hearing listeners typically benefit from spatial release from masking, i.e., spatial separation between target and competing talkers (e.g., Freyman et al., 1999; Shinn-Cunningham et al., 2001). Because spatial cues were not simulated in the present study, participants were deprived of important cues that would facilitate speech segregation and intelligibility that normally supports speech-in-noise perception. Thus, the multitalker babble used here likely represents an extra demanding listening scenario, suggesting that the observed training-related gains might be more conservative estimates of potential real-world benefits.

A key feature of this study was the use of synthesized speech generated with text-to-speech systems. Synthetic speech offers several practical advantages: it removes constraints such as limited speaker availability, enables precise control over acoustic and prosodic parameters, and allows rapid creation of large stimulus sets. These features can be particularly beneficial for high variability phonetic training research. Although there are concerns about naturalness of synthetic voices, recent text-to-speech systems have improved substantially and can perform comparably to natural speech in perception tasks (Calandruccio et al., 2025; Nuesse et al., 2019). In this study, an initial validation with a native speaker confirmed that the phonetic stimuli were intelligible, and post-training results suggest that they induced learning. At the same time, the two untrained voices were already more intelligible at pre-test than the trained voices. All voices were obtained from the same text-to-speech platform and selected using the same procedure, and during our initial inspection of the speech materials we did not find these voices more hyperarticulated or perceptually easier to understand than the others. Their higher intelligibility may therefore reflect normal variability among the available synthetic voices, similar to what is observed among natural voices. There is some evidence that there is more variability in recognition scores between human and synthetic voices (Calandruccio et al., 2025), although additional work is needed to corroborate this finding and assess its implications for the selection and design of text-to-speech voices in phonetic training research. Overall, synthetic speech is a promising tool, but further work is needed to determine when and how it supports phonetic learning.

The paired format of the task may have facilitated learning. Presenting minimal pairs in immediate succession (e.g., *bit–beet*) likely promoted contrastive processing by drawing attention to the critical phonemic differences. Future work could compare this format with a single-token identification task, which is already implemented in the game, to determine whether the contrastive presentation offers any advantage. An additional manipulation of interest would be to vary speaker identity within a trial. In the current version, both words of the minimal pair are produced by the same speaker, which may reduce irrelevant acoustic variability and support early contrast formation. In future versions, using different speakers for the two tokens could further enhance learning by encouraging abstraction over talker-specific characteristics, thereby directing attention toward phonemic rather than idiosyncratic voice-related cues. At the same time, such a manipulation would likely increase task difficulty by introducing additional acoustic variability within trials, which might not be ideal during early stages of learning. Future work could examine how within-trial speaker variability affects learning.

Informal brief self-reports after training suggested moderate increases in perceived phonemic awareness and high engagement with the game. These subjective ratings should be interpreted with caution, but they suggest that the game-based format may support motivation and sustained participation.

The present study was performed with adult Greek speakers with upper- to high English proficiency. Even in this relatively experienced group, the short, 1.5-hr training led to measurable improvement. Larger effects might be expected in less experienced learners with lower proficiency or younger populations, where phonological categories remain more plastic (Flege, 2003; Flege & Bohn, 2021; Shinohara & Iverson, 2021). Analysis of the training data showed that most participants improved their performance over the course of the game (Figure 7). However, interpretation of the level distributions (Figure 7B) requires some caution, as difficulty was capped at level 8, such that participants who reached this level could not progress further, reducing across-subject variability in later trial bins. Despite this ceiling effect, the shift toward reaching higher levels earlier and staying there longer across sessions suggests an improved ability to handle increased task difficulty.

Several limitations should be acknowledged. First, the sample size was small, which reduces statistical power and limits the generalizability of the findings. Second, the study did not include follow-up testing, so it remains unclear whether the observed improvements were retained over time. Third, the number of items per contrast was modest (only three during training), which may have restricted the range of phonetic variation available to learners. In addition, the adaptive procedure used −8.5 dB TMR as the highest difficulty level. While some level of background babble is useful for

approximating adverse real-world listening conditions, very low TMRs risk shifting the task away from phonetic discrimination, as the relevant acoustic cues become difficult to access. This also created a practical limitation during training: two higher-performing participants reached the ceiling early, leaving little room for additional improvement once the task became dominated by noise. Importantly, the primary goal of the training was to promote robust phonetic learning and generalization, rather than performance in extremely adverse listening conditions. From this perspective, future versions of the game would benefit from using a finer TMR progression and a narrower noise range (e.g., down to −6 dB), which would better preserve access to phonetic cues while maintaining ecological validity. Increases in task difficulty could then be achieved more effectively by adding more voices and tokens, thereby increasing phonetic and talker variability—an approach more directly aligned with mechanisms known to support generalization in phonetic learning—rather than relying on increasingly extreme babble levels. Finally, training was brief—approximately 1.5 hours across three sessions. Although the observed improvements are encouraging, longer and more distributed training schedules typically yield stronger and more stable learning.

Despite these limitations, the study provides initial evidence that a videogame-based high-variability phonetic training with synthetic speech, delivered in both quiet and babble conditions, can support improvement in difficult English vowel contrasts. Because the program is open-source, it can be adapted to other languages and learner groups, including those whose native language has small vowel inventories, such as Spanish or Italian. A game-based format may help sustain motivation, making the platform potentially useful for classroom and experimental applications.

Ethical approval
The study was approved by the Research Ethics Committee of the University of Thessaly (protocol no.72/10.09.2021).

Informed consent
All participants provided written informed consent prior to participation. Participation was voluntary, and participants were free to withdraw from the study at any time without penalty.

Use of AI
The first author (E.V.) used OpenAI's ChatGPT (GPT-4 series) to assist with language editing, refinement of R code used for statistical analyses and figure generation, and various aspects of videogame development, including programming support and debugging. All content was reviewed and approved by the authors.

Data availability statement
Software availability
The PhonemeQuest software and associated materials (audio stimuli and images) are available on Zenodo: https://doi.org/10.5281/zenodo.15384223 (Vlahou et al., 2025). Materials are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Underlying data
The anonymized behavioral data and R scripts used to reproduce the analyses are available at Zenodo: https://doi.org/10.5281/zenodo.18729701 (Vlahou, 2026). Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Extended data
Appendix A1 from the manuscript is available in the same Zenodo repository as the anonymized data and analysis scripts: https://doi.org/10.5281/zenodo.18729701 (Vlahou, 2026). Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgments
We thank Maruša Laure for creating and providing the 12-talker multitalker babble used in the training and testing sessions.

References

Arvaniti A: **Greek phonetics: The state of the art.** *Journal of Greek Linguistics.* 2007; **8**(1): 97–208.
Publisher Full Text

Baayen R, Davidson D, Bates D: **Mixed-effects modeling with crossed random effects for subjects and items.** *J. Mem. Lang.* 2008; **59**(4): 390–412.
Publisher Full Text

Bates D, Martin M, Bolker B, *et al.*: **Fitting linear mixed-effects models using lme4.** *J. Stat. Softw.* 2015; **67**(1): 1–48.
Publisher Full Text

Bench J, Kowal Å, Bamford J: **The Bkb (Bamford-Kowal-Bench) sentence lists for partially-hearing children.** *Br. J. Audiol.* 1979; **13**(3): 108–112.
PubMed Abstract | **Publisher Full Text**

Best CT: **The emergence of native-language phonological influences in infants: A perceptual assimilation model.** Goodman JC, Nusbaum HC, editors. *The development of speech perception: The transition from speech sounds to spoken words.* MIT Press; 1994; pp. 167–224.
Publisher Full Text

Best CT, Tyler MD: **Nonnative and second-language speech perception.** Bohn O-S, Munro MJ, editors. *Second language speech learning: The role of language experience in speech perception and production.* John Benjamins; 2007; pp. 13–34.
Publisher Full Text

Bradlow AR: **Training non-native language sound patterns: Lessons from training Japanese adults on the English/r/–/l/contrast.** Hansen Edwards JG, Zampini ML, editors. *Phonology and second language acquisition.* John Benjamins; 2008; **36**: 287–308.
Publisher Full Text

Calandruccio L, Weidman D, Leatherwood A, *et al.*: **Testing sentence-in-noise recognition with synthetic speech and automatic speech recognition.** *J. Speech Lang. Hear. Res.* 2025; **68**: 6114–6128.
PubMed Abstract | **Publisher Full Text**

Deterding S, Dixon D, Khaled R, *et al.*: **From game design elements to gamefulness.** *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments.* 2011; 9–15.
Publisher Full Text

EF Education First: **EF English Proficiency Index (EF EPI) – 2023 edition.** 2023. Retrieved February 21, 2026.
Reference Source

Flege JE: **Assessing constraints on second-language segmental production and perception.** Meyer AS, Wheeldon LR, Crocker MW, editors. *Phonetics and phonology in language comprehension and production: Differences and similarities.* Mouton de Gruyter; 2003; Vol. **6**: pp. 319–355.
Publisher Full Text

Flege JE, Bohn O-S: **The revised speech learning model (SLM-r).** Wayland R, editor. *Second language speech learning: Theoretical and empirical progress.* Cambridge University Press; 2021; pp. 3–83.
Publisher Full Text

Freyman RL, Helfer KS, McCall DD, *et al.*: **The role of perceived spatial separation in the unmasking of speech.** *J. Acoust. Soc. Am.* 1999; **106**(6): 3578–3588.
PubMed Abstract | **Publisher Full Text**

García Lecumberri ML, Cooke M, Cutler A: **Non-native speech perception in adverse conditions: A review.** *Speech Comm.* 2010; **52**(11–12): 864–886.
Publisher Full Text

Georgiou GP: **Bit and beat are heard as the same: Mapping the vowel perceptual patterns of Greek-English bilingual children.** *Lang. Sci.* 2019; **72**: 1–12.
Publisher Full Text

Golestani N, Zatorre RJ: **Learning new sounds of speech: Reallocation of neural substrates.** *NeuroImage.* 2004; **21**(2): 494–506.
PubMed Abstract | **Publisher Full Text**

Goto H: **Auditory perception by normal Japanese adults of the sounds "l" and "r".** *Neuropsychologia.* 1971; **9**(3): 317–323.
PubMed Abstract | **Publisher Full Text**

Guevara-Rukoz A, Martin A, Yamauchi Y, *et al.*: **Prototyping a web-based phonetic training game to improve/r/–/l/identification by Japanese learners of English.** *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019).* 2019; 116–120.
Publisher Full Text

Hamari J, Koivisto J, Sarsa H: **Does gamification work? A literature review of empirical studies on gamification.** *2014 47th Hawaii International Conference on System Sciences.* 2014; 3025–3034.
Publisher Full Text

Iverson P, Evans BG: **Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration.** *J. Acoust. Soc. Am.* 2007; **122**(5): 2842–2854.
PubMed Abstract | **Publisher Full Text**

Iverson P, Hazan V, Bannister K: **Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/–/l/ to Japanese adults.** *J. Acoust. Soc. Am.* 2005; **118**(5): 3267–3278.
PubMed Abstract | **Publisher Full Text**

Jafari Z, Ryait H, Habibnezhad M, *et al.*: **Reaction time, speech recognition, and verbal memory performance: Nonnative versus native English speakers.** *J. Speech Lang. Hear. Res.* 2025; **68**: 3056–3068.
PubMed Abstract | **Publisher Full Text**

King S: **Measuring a decade of progress in text-to-speech.** *Loquens.* 2014; **1**(1): e006.
Publisher Full Text

Kuhl PK, Conboy BT, Coffey-Corina S, *et al.*: **Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e).** *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2008; **363**(1493): 979–1000.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ladefoged P, Johnson K: *A course in phonetics.* Wadsworth;6th ed 2011.

Lengeris A: **Perceptual assimilation and l2 learning: Evidence from the perception of southern British English vowels by native speakers of Greek and Japanese.** *Phonetica.* 2009; **66**(3): 169–187.
PubMed Abstract | **Publisher Full Text**

Lengeris A, Hazan V: **The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek.** *J. Acoust. Soc. Am.* 2010; **128**(6): 3757–3768.
PubMed Abstract | **Publisher Full Text**

Lenth RV: **Emmeans: Estimated marginal means, aka least-squares means [R package version 1.10.3].** 2024.
Reference Source

Leong CXR, Price JM, Pitchford NJ, *et al.*: **High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained.** *PLOS ONE.* 2018; **13**(10): e0204888.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lim S-J, Fiez JA, Holt LL: **Role of the striatum in incidental learning of sound categories.** *Proc. Natl. Acad. Sci.* 2019; **116**(10): 4671–4680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lim S-J, Holt LL: **Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization.** *Cogn. Sci.* 2011; **35**(7): 1390–1405.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lively SE, Logan JS, Pisoni DB: **Training Japanese listeners to identify English/r/and/l/.ii: The role of phonetic environment and talker variability in learning new perceptual categories.** *J. Acoust. Soc. Am.* 1993; **94**(3): 1242–1255.
Publisher Full Text

Logan JS, Lively SE, Pisoni DB: **Training Japanese listeners to identify english/r/and/l/: A first report.** *J. Acoust. Soc. Am.* 1991; **89**(2): 874–886.
Publisher Full Text

Matuschek H, Kliegl R, Vasishth S, *et al.*: **Balancing type I error and power in linear mixed models.** *J. Mem. Lang.* 2017; **94**: 305–315.
Publisher Full Text

McCandliss BD, Fiez JA, Protopapas A, *et al.*: **Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a hebbian model of plasticity and stabilization in spoken language perception.** *Cogn. Affect. Behav. Neurosci.* 2002; **2**(2): 89–108.
PubMed Abstract | **Publisher Full Text**

Miyawaki K, Jenkins JJ, Strange W, *et al.*: **An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English.** *Percept. Psychophys.* 1975; **18**(5): 331–340.
Publisher Full Text

Nuesse T, Wiercinski B, Brand T, *et al.*: **Measuring speech recognition with a matrix test using synthetic speech.** *Trends in Hearing.* 2019; **23**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pruitt JS, Jenkins JJ, Strange W: **Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese.** *J. Acoust. Soc. Am.* 2006; **119**(3): 1684–1696.
Publisher Full Text

Saito K, Hanzawa K, Petrova K, *et al.*: **Incidental and multimodal high variability phonetic training: Potential, limits, and future directions.** *Lang. Learn.* 2022; **72**(4): 1049–1091.
Publisher Full Text

Shinn-Cunningham BG, Schickler J, Kopčo N, *et al.*: **Spatial unmasking of nearby speech sources in a simulated anechoic environment.** *J. Acoust. Soc. Am.* 2001; **110**(2): 1118–1129.
PubMed Abstract | **Publisher Full Text**

Shinohara Y, Iverson P: **The effect of age on English/r/–/l/perceptual training outcomes for Japanese speakers.** *J. Phon.* 2021; **89**: 101108.
Publisher Full Text

Trubetzkoy NS: *Principles of phonology.* Baltaxe CAM, Trans). University of California Press; 1969. [Original work published 1939].

Uchihara T, Karas M, Thomson RI: **High variability phonetic training (HVPT): A meta-analysis of l2 perceptual training studies.** *Stud. Second. Lang. Acquis.* 2025; **47**(3): 794–827.
Publisher Full Text

Vlahou E: PhonemeQuest – Anonymized Data and Analysis Scripts. [Data set]. *Zenodo.* 2026.
Publisher Full Text

Vlahou E, Kopčo N, Seitz A: **PhonemeQuest: A Video Game for Training in English Phoneme Identification [Software].** *Zenodo.* 2025.
Publisher Full Text

Vlahou E, Protopapas A, Seitz AR: **Implicit training of nonnative speech stimuli.** *J. Exp. Psychol. Gen.* 2012; **141**(2): 363–381.
PubMed Abstract | **Publisher Full Text**

Vlahou E, Seitz AR, Kopčo N: **Nonnative implicit phonetic training in multiple reverberant environments.** *Atten. Percept. Psychophys.* 2019; **81**(4): 935–947.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wade T, Holt LL: **Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task.** *J. Acoust. Soc. Am.* 2005; **118**(4): 2618–2633.
PubMed Abstract | **Publisher Full Text**

Wang Y, Spence MM, Jongman A, *et al.*: **Training american listeners to perceive mandarin tones.** *J. Acoust. Soc. Am.* 1999; **106**(6): 3649–3658.
PubMed Abstract | **Publisher Full Text**

Werker JF, Tees RC: **Developmental changes across childhood in the perception of non-native speech sounds.** *Can. J. Psychol.* 1983; **37**(2): 278–286.
PubMed Abstract | **Publisher Full Text**