

SPECIAL ISSUE ARTICLE

Are nonwords really superior to real words to train L2 sounds? Insights from an HVPT study on L2 English front vowels

Celia Gorba¹ , Miquel Llompart²  and Pilar Prieto^{2,3}

¹Filologia Anglesa i Germanística, Universitat Autònoma de Barcelona, Spain, ²Translation and Language Sciences, Universitat Pompeu Fabra, Spain and ³Institució Catalana de Recerca i Estudis Avançats, Spain
Corresponding author: Celia Gorba; Email: celia.gorba@uab.cat

(Received 20 November 2025 UTC; revised 29 April 2026 UTC; accepted 1 June 2026 UTC)

Abstract

This study investigated the effects of a high-variability phonetic training (HVPT) regime using nonwords (NW), compared to a comparable regime using real words (RW), on the perception of English front vowels by L1-Catalan/Spanish learners. Participants completed pre- and posttest tasks targeting both pre-lexical (vowel identification of NW and RW) and phono-lexical processing (word-picture matching, WPM, and lexical decision, LD). Training consisted of six sessions of phoneme identification with either NW or RW stimuli. Results revealed that the RW group consistently outperformed the NW group, with significant differences in RW identification and WPM. Importantly, learning generalized to new speakers and items, particularly in the RW condition. Lexical tasks proved more challenging than pre-lexical ones, with lower overall scores and no significant gains in LD. These findings indicate that RW training produces stronger perceptual and phono-lexical gains than NW training, challenging findings from prior studies that favored NW-based regimes.

Keywords: English front vowels; high-variability phonetic training; perception training; phono-lexical processing; real word vs nonword training

Introduction

Second language (L2) learners often face challenges in mastering L2 phonology, as their perceptual and articulatory systems are shaped by the phonological categories of their first language (L1), which can interfere with accurate perception and production of L2 sounds. Prominent L2 speech models, such as the Speech Learning Model (SLM, Flege & Bohn, 2021) and PAM-L2 (Best & Tyler, 2007), posit that pinpointing the phonetic differences with L1-similar sounds is necessary for accurate category creation. Specifically, the SLM proposes that greater perceived

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-ShareAlike licence (<https://creativecommons.org/licenses/by-sa/4.0/>), which permits re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited.

dissimilarity between L1 and L2 sounds promotes accurate category formation, while similar sounds often merge into the same L1 category, causing errors in perception and production. Moreover, PAM-L2 predicts L2 perceptual accuracy based on how L2 sounds map onto L1 categories: discrimination is strongest when an L2 contrast maps to two distinct L1 categories (Two-Category Assimilation). In contrast, discrimination is poorer when both L2 sounds fall into one L1 category, either equally well (Single Category Assimilation) or with one sound as a better fit (Category Goodness Difference).

The models also claim that accurate category formation is based on exposure to authentic (i.e., native-like) input (e.g., Flege & Bohn, 2021). Previous studies have found that access to authentic input in an immersion setting leads to more accurate perception and production skills, as greater L2 use results in greater L2 input in this type of context (e.g., Gorba, 2019; Gorba & Cebrian, 2021). However, access to authentic input is often limited in instructional settings (e.g., Muñoz, 2008), and pronunciation is seldom prioritized in the classroom (e.g., Levis, 2005). Moreover, the input received in foreign language learning contexts may be L1-accented, where some L2 contrasts are not realized or realized in a non-native manner, as learners are exposed to non-native speech that is likely to present errors (Llompart & Reinisch, 2019). In this context, phonetic training can be useful to compensate for this limitation by providing focused input that addresses particularly difficult L2 phonetic contrasts. Specifically, high-variability phonetic training (HVPT) has proven its effectiveness in improving L2 phonetic learning (Thomson, 2018). This training paradigm is a form of perceptual training that exposes learners to L2 target phones through different stimuli and voices. The lexical status of the stimuli has been debated, particularly the use of nonwords (NW) versus real words (RW), with NW being more common.

The current study aims to determine (i) whether there are differences in the effectiveness of NW and RW stimuli as training materials in HVPT, and (ii) whether training gains derived from a task focusing on sound identification transfer to learners' ability to recognize spoken words (i.e., link auditory forms to meanings) that differ in the critical sounds. To do this, we focus on how bilingual L1-Catalan/Spanish learners of English perceive the front vowels of English, particularly the distinction between /i/ and /I/, and on their perceptual gains after completing either a training regime with RWs or a comparable one using NWs. In the following subsections, we discuss (i) the acquisition of the English front vowels by L1 Catalan/Spanish speakers, (ii) methodological considerations in HVPT, especially the role of stimulus type (words vs. nonwords), and (iii) the transfer (or lack thereof) of pre-lexical perceptual training gains to lexical processing.

The acquisition of English front vowels by L1-Catalan/Spanish learners

The current study investigates the acquisition of the front vowel section of English vowels (i.e., /i:/, /I/, /e/, and /æ/), by L1-Catalan/Spanish L2 learners. In order to predict the challenges that L2 learners may face when acquiring this set of vowels, crosslinguistic perceptual data needs to be considered. English has four front vowels (/i:/, /I/, /e/, /æ/), while Spanish has only two (/i/, /e/), and Catalan has three (/i/, /e/, /ɛ/). Previous studies have investigated the equivalences and degree of

perceived similarity between L2-English and L1-Catalan and Spanish vowels. Overlapping assimilation patterns—i.e., assimilation to the same L1 phone—have been observed for English front vowels on the part of L1-Spanish and L1-Catalan/Spanish learners, which points to reduced discriminability (Baigorri et al. 2019; Cebrian, 2019, 2021; Cebrian et al., 2021; Escudero & Chládkova, 2010).

When it comes to both L1 Spanish and L1 Catalan learners of English, previous studies have found that English vowel /i:/ is almost unequivocally perceived as L1 vowel /i/ (Baigorri et al., 2019; Cebrian, 2019; Escudero & Chládkova, 2010; Morrison, 2012). Regarding /I/, different assimilation patterns have been described in the literature. Although some studies have found that English /I/ is assimilated to both Spanish and Catalan /i/ and /e/ (Cebrian, 2019; Baigorri, et al., 2019; Escudero & Chládkova, for Spanish; Cebrian 2019; Rallo Fabra & Tyler, 2023, for Spanish), other studies suggest that this sound is mostly perceived as L1 /e/ (Escudero & Chládkova, 2010; Morrison, 2012). As for /ε/, previous studies report it is mostly assimilated to Spanish /e/ or Catalan /ε/, whereas /æ/ is perceived as L1 /a/ (Baigorri et al., 2019; Cebrian, 2019; Escudero & Chládkova, 2010 (for Southern British English)).

These results, thus, point to a difficulty in accurate categorization of English front vowels by L1-Catalan/Spanish speakers. This is true particularly regarding the /i:/-/I/ contrast, due to their overlapping L1 mappings into L1 /i/ and overreliance on duration cues, and, to a lesser extent for /I/-/ε/, as both may be categorized as Spanish /e/, according to some studies. Empirical data support the models' predictions and show that acquiring a new L2 phone may lead to confusion with an existing L2 category in its vicinity. In the case of English front vowels, both L1-Spanish speakers (e.g., Escudero & Boersma, 2004; Flege et al. 1997; Kondaurouva & Francis, 2010; Morrison, 2008, 2009; Smith & Hayes-Harb, 2016) and L1-Spanish-Catalan speakers (e.g., Cebrian, Gorba & Gavalda, 2021; Fullana Rivera & Mackay, 2003; Rallo Fabra & Romero, 2012) have been found to struggle with accurate perception and production, particularly for /i:/-/I/. Even among advanced learners, these two vowels often show an acoustic overlap in production.

HVPT: some methodological considerations

HVPT is a type of perceptual training, usually in the form of a forced-choice identification task, that introduces variability by using multiple voices (Barriuso & Hayes-Harb, 2018; Thomson, 2018), which has been found to be more effective than exposure to a single speaker, as it facilitates generalization to new talkers (e.g., Lively et al., 1993; Wong, 2014). Over the past two decades, HVPT has gained popularity and has been widely used, demonstrating its effectiveness in L2 phonetic acquisition (see Thomson, 2018, for a review). Studies have shown significant perceptual gains from pretest to posttest, with evidence of transfer to production for vowels (e.g., Aliaga-García & Mora, 2009; Carlet & Cebrian, 2019; Iverson & Evans, 2009; Lambacher, et al., 2005; Mora et al., 2022; Mora-Plaza et al., 2022) as well as consonants (e.g., Aliaga-García & Mora, 2009; Carlet, 2017; Logan, Lively, & Pisoni, 1991). For example, regarding L2 vowel acquisition, Carlet and Cebrian (2019) found that Catalan/Spanish learners of English improved their perception of /i: I æ ʌ ɜ:/ after five 30-minute training sessions.

When it comes to the methodological implementation of HVPT, some aspects have been established as best practice, including the use of a forced-choice identification task, providing trainees with feedback on the accuracy of their responses, and the use of multiple talkers rather than a single talker (Barriuso & Hayes-Harb, 2018; Thomson, 2018; Uchihara *et al.*, 2025). However, variability has been observed regarding other aspects. Specifically, differences between studies involve the length of the training and sessions, the number of speakers used and their distribution, the nature of the feedback, and the type of materials used—particularly their modality (audio-only vs. audiovisual) and lexicality (RW vs. nonword stimuli) (Aliaga-Garcia, 2017; Barriuso & Hayes-Harb, 2018; Carlet, 2017; Thomson, 2018). Given the wide methodological variability in its implementation, Thomson's (2018) review concluded that, while the effectiveness of HVPT is demonstrated, there is a lack of any identifiable best approach. With regard to modality, while most HVPT studies rely on audio-only training, we chose to use audiovisual input, consistent with multimodal theories of enriched learning (Mathias & von Kriegstein, 2023, for a review). Some studies have shown that audiovisual input can enhance perception and production through visible articulatory cues (Hardison, 2003; Inceoglu, 2016), whereas others have found no clear advantage (Reyes & Hazan, 2021). Uchihara *et al.*'s (2025) meta-analysis likewise reported that audiovisual input was not a significant moderator of perceptual learning, but the authors cautioned against interpreting this as evidence of ineffectiveness and recommended directly comparing audiovisual and audio-only conditions. Therefore, although the superiority of audiovisual input over audio-only training remains uncertain, it may still offer benefits for both the experimental design (e.g., greater naturalness, inclusion of articulatory movements) and the L2 learners (e.g., increased motivation; see Muñoz, 2023).

As far as lexicality of the stimuli, an important methodological difference between HVPT training studies that warrants special attention in the current paper is the nature of the stimuli, specifically whether RW or NW are used. While a recent meta-analysis on HVPT studies shows that stimuli lexicality, operationalized as the inclusion of NW or lack thereof, was not a significant moderator of training gains (Uchihara *et al.*, 2025), some studies that have directly compared NW and RW suggest that training with NW leads to greater improvement in perceptual abilities than training with RW, as NW allow learners to focus on distinctive phonetic features without lexical interference (e.g., Mora-Plaza *et al.*, 2022; Thomson & Derwing, 2016). For example, Mora-Plaza *et al.* (2022) investigated the effects of HVPT on the perception of /æ/-/ʌ/ under different conditions, which included comparisons across training stimuli type (NW vs. RW) and training administration (with or without noise and/or visual monitoring). Given the fact that several conditions were tested, the sample size per group was of 11–14 participants. Results indicated that participants trained with NW improved to a significantly greater extent in their perceptual abilities from pretest to posttest than those trained with RW. Importantly, however, little improvement was obtained in the lexical encoding of the contrast, as revealed by lexical decision and delayed sentence repetition tasks at pre- and posttest. Moreover, in a study including the same tasks and a greater sample size (Mora *et al.*, 2021), no differences were reported between RW and NW training. Thomson and Derwing's (2016) exploratory study (with 9 to 12 participants

per group) also showed that L2 learners (with mixed language backgrounds) who had undergone training with NW were perceived to have improved their pronunciation of English vowels to a greater extent than those trained using RW. This study, however, did not assess perception abilities.

On the other hand, research also suggests that L2 vowels appear more difficult to categorize when embedded in NW than in RW, suggesting that lexico-phonological categorization may be more precise than pre-lexical categorization (Rato & Carlet, 2020). This finding supports the idea that lexical categories and phonological categories may be established at different times and that the sound categories may be abstracted from lexical items at a later stage of L2 learning. Furthermore, Carlet (2017) trained participants exclusively with NW, yet learners clearly expressed they would have preferred to be trained with RW in a post-training survey, highlighting potential considerations for training design. All in all, the limited research comparing RW and NW stimuli in training L2 phonological perception, along with findings suggesting an advantage of RW in phonological categorization, highlights the need for further investigation into the conditions under which each stimulus type may be more effective.

Lexical vs. pre-lexical processing

Real-world communication involves recognizing words—that is, matching particular acoustic signals to pairings of phonological forms and meanings—rather than merely identifying sounds or strings of sounds in a vacuum. While accurate sound-level categorization is a necessary foundation for L2 phonological learning, it does not always ensure successful lexical processing (see Darcy et al., 2025, for a recent theoretical overview). A growing number of studies have shown that learners may distinguish L2 sounds in isolation yet struggle to use those same contrasts when recognizing words—whether rejecting NW or retrieving the intended lexical items (e.g., Amengual, 2016; Darcy et al., 2013; Llompарт, 2021; Llompарт & Reinisch, 2019) and that accuracy in the former only weakly predicts outcomes in the latter (e.g., Díaz et al., 2012; Llompарт et al., 2025; Rocca et al., 2025). These mismatches raise doubts about whether gains from training regimes focused mostly on sound identification may have a direct impact on learners' spoken word recognition ability. In particular, it is difficult to see how improvements at the phonetic level alone derived from training with only non-lexical material (i.e., NW) would support more accurate or efficient word recognition. By contrast, with training regimes using RW stimuli, one may at least expect some level of engagement of the lexicon, which may help bridge the gap between pre-lexical gains and improvements at the lexical level.

Importantly, the issues outlined above are compounded by the fact that most phonetic training studies have assessed progress using tasks closely aligned with those used during training—typically involving sound-level-focused categorization or discrimination tasks. As a result, the field has relatively little evidence regarding whether training enhances the ability to recognize spoken words when the focus is not on the critical sounds. To date, only Melnik and Peperkamp (2021) have reported evidence that phonetic training can enhance spoken word recognition. And crucially for the questions at hand, their training involved RW stimuli. By contrast, a very recent study (Rebuschat et al., 2025) failed to find evidence that training—with stimuli that were functionally NW for the learner population—led

to improvements in L2 novel word learning outcomes, even despite observable gains in categorization and discrimination accuracy for the target contrasts. Thus, despite the growing interest in the distinction between pre-lexical and lexical processing in L2 speech acquisition, little is known about how different training types shape learners' ability to activate L2 word forms and meanings during comprehension. This gap is central to the present study, which not only compares the efficacy of word-based and nonword-based training for conventional identification tasks but also examines whether and how benefits derived from these training types transfer to the lexical domain.

The current study

The current study's aim is twofold, as it examines whether (1) phonetic trainings involving NW or RW differ in their effectiveness to improve the perception of English front vowels and (2) if HVPT training regimes may result in gains not only in pre-lexical tasks (which focus on phonetic form) but also in lexical tasks (which focus on meaning). To investigate these issues, L1-Catalan/Spanish upper-intermediate learners of English underwent a phonetic training regime. Two comparable regimes were used, with half the participants assigned to each condition: a program with NW and an identical one with RW. Progress was tested with a pretest/posttest design including four perception tasks, varying in lexical activation. Although the scope of the present study focuses on perceptual tasks, participants also completed a battery of production tasks in the same session, preceding the perception tasks. These included a sentence reading task, a word reading task, and two repetition tasks (one NW, one RW). Due to word limitations, production results will be presented in a subsequent paper ("Gorba, Llompart & Prieto (in preparation-a). Transfer effects from speech perception to production: Perceptual training with real words boosts generalization to (L2) production abilities." Gorba, Llompart & Prieto, in preparation-a).¹

Regarding expectations, some previous research suggests that the NW group may show greater improvement than the RW group (Thomson & Derwing, 2016; Mora-Plaza et al., 2022), although the evidence remains inconclusive. In fact, Mora et al. (2021) did not observe a significant difference between RW and NW training. Moreover, RW are easier to categorize and may provide additional benefits: they can make the training procedure more engaging, reduce uncertainty through lexical support, and connect more directly to the representations involved in speech comprehension. Consequently, RW may yield gains comparable to, or greater than, those observed with NW. With respect to lexical tasks, only modest improvements are expected; if gains occur, they are more likely with the RW group, as these learners trained phonetic contrasts embedded in lexical items.

Methodology

Participants

A total of 49 participants ($N = 42$ females) in their early twenties (mean age: 20.3 (SD: 4)) completed the experiment. They were first- and second-year undergraduate students majoring in Linguistics or Translation Sciences at Universitat Pompeu

Fabra and first year students majoring in English Studies at Universitat Autònoma de Barcelona. Participants were L1-Catalan and Spanish speakers who had learnt English as an L2 and had an upper-intermediate to advanced level as well as a very limited knowledge of phonetics and phonology, as they still had not received training in the field. Moreover, they had not lived in an English-speaking country (only visited) and had not participated in other phonetic training programs. Although no proficiency or vocabulary size test was conducted on participants, English functions as a vehicular language in a substantial portion of coursework across all three majors, and especially in English Studies. Thus, a comparable level of proficiency is expected within and across groups.

Materials and procedure

Stimuli

In order to create the materials for the training study, 7 speakers of General American English (4 females and 3 males) were video recorded producing a set of English RW as well as NWs (which matched the phonological and phonotactic system of English). Specifically, a total of 120 RW and 120 NW were recorded, out of which 30 each were used for the pre-and posttests and 42 for each type were used for the training sessions² (see Supplementary Materials). An effort was made to use RW that were picturable, as was required by the WPM task (see below). It should also be noted that are RW stimuli were common English words that participants were expected to be familiar with given their degree of proficiency. In fact, none reported unfamiliarity with any of the items, based on informal conversations with the experimenter in which participants were asked about their impressions of the training, including what they found difficult, easy, or entertaining. Additionally, Speaker 2 (male) was audio-recorded reading a set of RW that was used in a lexical decision task.

Both testing and training stimuli were monosyllabic CVC, CCVC, CVCC RW, and NW that included the target phones, that is the English front vowels, i.e., /i, I, e, æ/. All stimuli were part of a minimal pair with another stimulus from the set, including minimal pairs that involved contiguous categories (i.e., /i:/-/I/, /I:/-/e/ and /e:/-/æ/, 5 each) and non-contiguous categories (/i:/-/æ/, /I:/-/æ/ and /i:/-/e/, 2 each). A subset of the testing words and NWs (40%) were also included in the training, whereas 60% were only used in testing. It should be noted, though, that none of the stimuli were included in both training and testing tasks, as different speakers were included at each stage.

The LD presented at pretest included 120 stimuli, including 60 RW and 60 NW. The RWs included 10 stimuli with each target vowels plus 20 fillers. They were monosyllabic and disyllabic common words which presented the target vowel in the stressed syllable. Crucially, the NW stimuli were similar to an actual RW except they presented an unexpected target vowel. For instance, the stimulus /'nIdəl/ is similar to the word “needle,” but vowel /I/ is used instead of /i:/. Ten NW stimuli were used for each target vowel, plus 20 fillers involving the consonants /k/ and /m/ (ex. the non-word /'lounməl/ based on “local”). These fillers were selected because they yielded very high accuracy in previous studies (Llompert, 2021; Llompert & Reinisch, 2019), allowing us to include easy-to-reject nonwords as control stimuli, which served as an internal check that the task functioned as intended.

Table 1. Perceptual tasks (phono-lexical and pre-lexical) in their order of completion. Abbreviations in brackets

Phono-lexical tasks	Pre-lexical tasks
1. Lexical decision task (LD)—audio-only stimuli	3. Real word identification task (RW ID)—audiovisual stimuli
2. Word-picture matching task (WPM)—audiovisual stimuli	4. Non-word identification task (NW ID)—audiovisual stimuli

Tasks and procedure

Participants took part in a total of eight sessions, including a pretest, six training sessions and a posttest. Each participant only participated in one training modality (either with RW or with NW), as explained below, and all completed the same pre- and posttest tasks.

Testing tasks

The pretest and posttest consisted of 4 perception tasks (see Table 1) that were completed in person on campus in a soundproof room. A laptop was used to implement the tasks, which were built in the platform Alchemer. Participants wore soundproof headsets and received brief instructions at the beginning of the testing session. Before the completion of the experiment, a consent form was handed out,³ and, before each task, detailed written instructions were also provided. The experimenter was the first author of the paper, namely a native speaker of Catalan and Spanish with a near-native level of English. She was present during the entire testing sessions and spoke English with the participants at all times.

The rationale for order of completion was to begin with tasks that involve the lexicon to a greater extent and focus less on form, in order to foster metalinguistic awareness of the sounds being tested. For the perception tasks, audio-only stimuli from Speaker 2 were used in the lexical decision task (LD), and video stimuli from Speakers 2 and 4 were used in the remaining tasks. For every task, participants could replay the stimulus presented at each trial one more time.

In the LD, participants were asked to determine whether the stimuli they were presented with in its exact form was an existing English word. The word-picture matching task (WPM) and the RW identification (RW ID) task involved the same RW stimuli, and both included audiovisual stimuli. The WPM consisted of the presentation of the audiovisual stimuli followed by a picture in the following screen and participants were instructed to determine whether the word they heard matched the concept depicted in the picture. The RW ID task was a forced-choice phoneme identification task in which participants had to select the target vowel that the stimulus presented via a mouse click. The auditory stimuli could be replayed once, and response options involved the phonemic symbol representing the target phones (i.e.,/i/, I, ε, æ/). Crucially, participants received a short familiarization training right before completing the ID tasks in which they were instructed on the sound to IPA symbol correspondence. The NW ID (NW ID) task was identical to the RW ID task—in terms of speaker, number of trials, response options, syllable structure of the stimuli, and display—except that, in this case, the stimuli involved NWs.

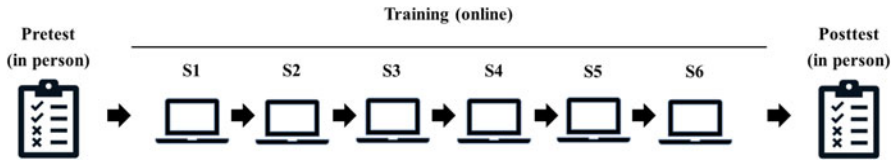


Figure 1. Diagram illustrating the sessions involved in the training, from pretest to training sessions (S1–S6) and posttest.

Training tasks

The training consisted of six short online sessions involving an audiovisual phoneme identification task (see Figure 1 for a depiction of the phases of the experiment). Participants received an e-mail with instructions to begin the training the day after the completion of the pretest and were allowed two to three weeks to complete all sessions, although most participants used up to three weeks. They were instructed to complete at least two sessions per week on separate days, at their time of convenience, with each session lasting about 10–15 minutes depending on the participant's own pace. Links to the platform Alchemer were provided for each session, which participants were instructed to complete without interruptions, using a computer or laptop and wearing headsets at all times. Weekly reminders were sent, and their progress was monitored by the experimenter on the Alchemer website. Each session presented detailed instructions and a reminder of words containing the target vowels and their corresponding response options, as already introduced in the familiarization phase at pretest. After completing all 6 sessions, participants were scheduled to complete the posttest in person at their time of convenience as close to the completion date as possible.

The task varied according to the group to which the participants were assigned, but speaker distribution across sessions and training conditions were the same. Specifically, the group assigned to the NW training condition was trained using NW stimuli and a forced-choice phoneme identification task, the RW group was trained using RWs and a comparable forced-choice phoneme identification task. Just as in the ID testing tasks, all stimuli were audiovisual, and IPA symbols for each target vowel were provided as response options. Participants had to click on the selected response option and then on a “Next” button to move on from trial to trial, ensuring they proceeded at their own pace. Participants had to click on the selected response option and then on a “Next” button to move on from trial to trial, ensuring they proceeded at their own pace. There was no fixation cross or response time limit. (see Figure 2 for an example of a trial display). They received feedback both immediately after each trial, indicating whether the response was correct or incorrect as well as showing the correct answer (see Figure 3), and at the end of session with a score (i.e., cumulative feedback). Each trial could be replayed one more time upon participant's request.

An attempt was made to assign participants to groups homogeneously by looking at the pretest perception results. Both overall accuracy across perceptual tasks and individual task accuracy were taken into account to distribute participants in groups, resulting in similar mean scores, as no significant differences were observed between groups at pretest (see section Data analysis and results below for results).

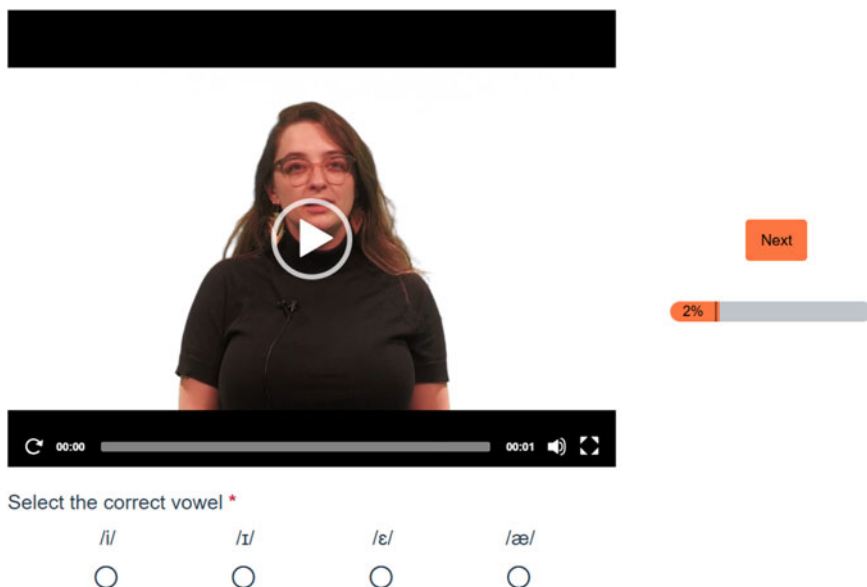


Figure 2. Example training trial display (RW condition, Session 1). All sessions and conditions shared the same layout.

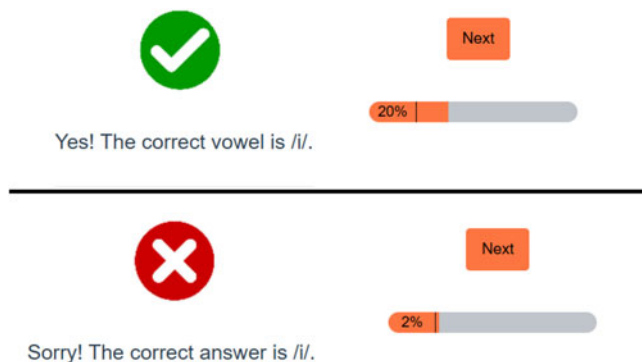


Figure 3. Feedback displays for correct (top) and incorrect (bottom) responses. Correct responses were accompanied by a tick icon and the selected answer and incorrect responses by a cross icon and the correct answer.

Data analysis and results

The training paradigm of the current study included all four English front vowels, /i:/, /ɪ/, /e/ and /æ/, as greater benefits have been observed for approaches that simultaneously train a greater portion of the vowel set (Nishi & Kewley-Port, 2007). However, we only discuss the potential training gains for the /i:/-/ɪ/ contrast. This is because /e/ and /æ/, as expected (see e.g., Aliaga-Garcia, 2017), were already easy to distinguish by our learner population in the pretest. Scores for this contrast in the identification tasks (NW ID and RW ID) ranged between 85% and 93% and the two groups were overall similarly accurate (NW training group = 87.4%; RW training

Table 2. Average percentage scores for each training group (NW training—RW training) in the study's 4 tasks (NW ID, RW ID, WPM, and LD). Standard deviations are in parentheses

Group Task	NW training		RW training	
	Pretest	Posttest	Pretest	Posttest
NW ID	72.4 (15.8)	83 (13.5)	71.1 (22.7)	87.4 (13)
RW ID	75.3 (17.6)	86.2 (12.9)	77.4 (20.1)	94.4 (9.2)
WPM	76 (15.3)	79.3 (17.3)	73.3 (15.6)	86.4 (12.2)
LD	63.4 (13.3)	64.4 (14.4)	61.5 (12.2)	64.4 (10.6)

group: 91.4%). In addition, combined accuracies in the two lexical tasks (WPM and LD) were also around 80–85%, which suggests that learners were quite able to use the mid-low vowel contrast to recognize words even prior to any training. For this reason, and also to be able to provide a thorough yet sufficiently streamlined account of the statistical analysis of our multi-task dataset, henceforth we only consider trials in which the target was either /i:/ or /I/. Importantly, for NW ID, RW ID, and LD, all trials with these targets are considered, while for the WPM, we only included in our analyses trials in which (i) the target was an /i:/- or an /I/-word and the picture matched that word (match condition), and (ii) the target was an /i:/- or an /I/-word and the picture represented the minimal pair with the other vowel (high-vowel mismatch condition). Trials where /i:/- and /I/-words were presented but the picture depicted a minimal pair with either /ε/ or /æ/ were excluded. This makes the analysis of the WPM more comparable to the other tasks, where there were no mismatch conditions (ID tasks) or the only possible mismatch was with the other high vowel (LD). For /i:/- /I/, Table 2 provides the mean accuracy by training group by session. Figures 4–7 are also presented below to showcase the distribution of individual values for each task.

/i:/- /I/ data from the four perception tasks (NW ID, RW ID, WPM, and LD) were analyzed by means of generalized linear mixed-effects regression models (lme4 package 1.1-31, Bates et al., 2015) in R (Version 4.2.2) run on each task's trial-by-trial data.⁴ All models had the same structure: Response (1 = correct, 0 = incorrect) was the binary dependent variable and Session (pretest-posttest) and training group (NW training-RW training), as well as their interaction, were the predictors of interest. Predictors were entered using the default “dummy” coding provided by lme4, such that posttest and NW training were mapped onto the intercept. Afterwards, given that test materials for the NW ID, RW ID tasks, and WPM contained items that had appeared in the training procedure of one of the groups (but not the other), follow-up analyses on only posttest data were run for these tasks in order to assess the extent to which posttest outcomes may have been influenced by this. Hence, generalized linear mixed-effects regression models (i.e., GLMER) on posttest data were run with response as dependent variable and training group and item type (trained-untrained) as predictors. In these models, NW training and Trained were mapped onto the intercept. Note that, for all models, model comparisons using the *anova()* function in R to elicit log-likelihood ratio tests were conducted to determine the structure of the random-effects structure. Random

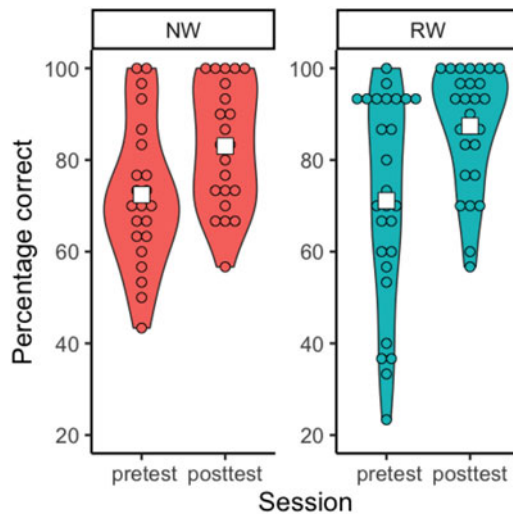


Figure 4. Distribution of accuracy scores in the NW ID task by session (pretest-posttest) and training group (NW training—left, in red; RW training—right, in green). Filled circles represent individual scores, and white squares indicate the group mean.

intercepts by Participants and Items were always included, but random slopes for a particular predictor over Participants or Items were only included if they improved the model's fit vs. a version of the model without the slopes in question.

Figure 4 showcases the distribution of individual scores in the NW ID task by Session by training group. The GLMER model for the NW ID task included random intercepts by Participants and Items and a random slope for Session over Participants. The model revealed a significant effect of Session ($b = -0.82$; $SE = 0.24$; $z = -3.37$; $p < .001$), while the effect of training group ($b = 0.54$; $SE = 0.43$; $z = 1.25$; $p = .21$) and its interaction with Session ($b = -0.53$; $SE = 0.33$; $z = -1.62$; $p = .11$) were not significant. This indicates that accuracies were overall higher in the posttest than in the pretest, suggesting that there were generalized training gains for this task, but no difference between the two groups was found. The lack of an interaction additionally suggests that the effect of session was not significantly modulated by group.

Subsequently, we assessed whether there were any differences between the posttest accuracies obtained for the items that were part of the NW training condition (i.e., *freet*, *deast*, *jeeb*, *frit*, *pllick*, *piff*) and those that were not (i.e., *deege*, *pleeg*, *preet*, *speech*, *didge*, *miv*, *plit*, *prit*, *sig*), and whether these differed by training group, since only the NW group that practiced with them during training. A GLMER with only random intercepts by Participants and Items rendered no effect of training group ($b = -0.02$; $SE = 0.49$; $z = -0.04$; $p = .97$) or item type ($b = -0.65$; $SE = 0.43$; $z = -1.49$; $p = .14$), but showed a significant interaction between the two ($b = 0.71$; $SE = 0.35$; $z = 2.04$; $p < .05$). Follow-up analyses with the data split by group showed that item type was not significant for the RW group ($b = 0.13$; $SE = 0.50$; $z = 0.26$; $p = .80$), as expected, but it was marginally significant for the NW training group ($b = -0.69$; $SE = 0.39$; $z = -1.75$; $p = .08$), suggesting

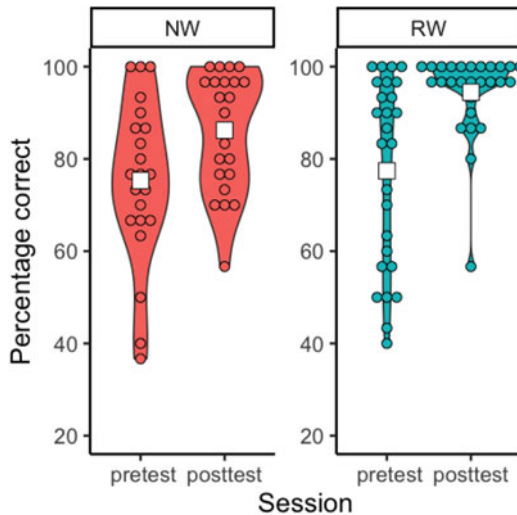


Figure 5. Distribution of accuracy scores in the RW ID task by session (pretest-posttest) and training group (NW training—left, in red; RW training—right, in green). Filled circles represent individual scores, and white squares indicate the group means.

that this group may have indeed performed to some extent better with the trained items than the untrained items.

Figure 5 provides the distributions of scores in the RW ID task by Session by training group. The GLMER model for the RW ID task only included random intercepts by Participants and Items because no random slopes improved its fit. The model revealed significant effects of Session ($b = -0.86$; $SE = 0.16$; $z = -5.48$; $p < .001$) and training group ($b = 1.35$; $SE = 0.44$; $z = 3.04$; $p < .01$), as well as a significant interaction between the two ($b = -1.08$; $SE = 0.25$; $z = -4.40$; $p < .001$). This indicates that scores were overall higher in the posttest than the pretest and also higher for the RW group than the NW group, but differences between groups were likely of different magnitudes at the two time points. Given that there was a main effect of Session pointing to overall training gains, in order to follow up on this interaction, we split the data by Session and ran two GLMERs, one for pretest data and one for posttest data, with Response as dependent variable and only training group as predictor to assess whether the two groups differed in each session. Results showed that the effect of training group was significant at posttest ($b = 1.37$; $SE = 0.50$; $z = 2.75$; $p < .01$) but not at pretest ($b = 0.27$; $SE = 0.40$; $z = 0.67$; $p = .50$). Therefore, gains from training on RW ID performance seem to have been larger for the RW training group, as they show similar accuracies to those in the NW training group in the pretest but end up with significantly higher scores in the posttest.

As with the previous task, then we assessed the potential effects of item type (trained vs. untrained) for posttest accuracies and how these may have interacted with training group, as only the RW training group had been exposed to items for this task during training (trained items: *cheap*, *heat*, *read*, *chip*, *lid*, *fit*; untrained items: *league*, *peak*, *seat*, *seed*, *bill*, *lift*, *pick*, *sick*, *sit*). Thus, a GLMER with random intercepts by Participants and Items was run on the RW ID posttest data, rendering

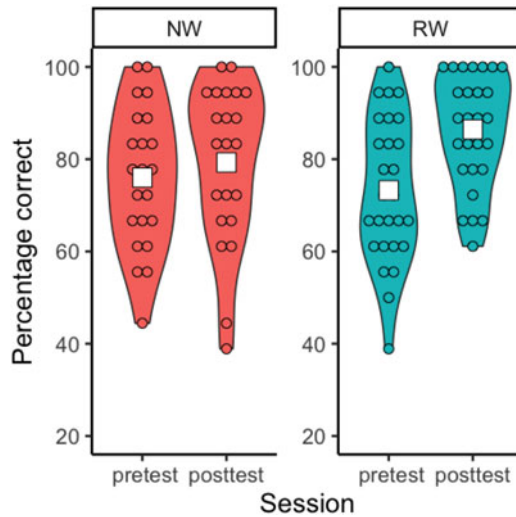


Figure 6. Distribution of accuracy scores in the WPM task by session (pretest-posttest) and training group (NW training—left, in red; RW training—right, in green). Filled circles represent individual scores, and white squares indicate the group means.

no effect of item type ($b = 0.69$; $SE = 0.56$; $z = 1.23$; $p = .22$) nor an interaction between item type and training group ($b = -0.18$; $SE = 0.43$; $z = -0.419$; $p = .68$). Only training group was significant ($b = 1.46$; $SE = 0.54$; $z = 2.68$; $p < .01$), which is in accordance with the results of the previous analysis. Therefore, this analysis did not provide any evidence that accuracies in RW ID in the posttest session were influenced by the training status of the individual items.

Figure 6 visualizes the individual scores in the WPM by session by training group. The GLMER model for the WPM only included random intercepts by Participants and Items because no random slopes improved its fit. The model showed a marginally significant effect of training group ($b = 0.57$; $SE = 0.32$; $z = 1.76$; $p = .08$) and no effect of Session ($b = -0.23$; $SE = 0.19$; $z = -1.24$; $p = .21$). However, the interaction between the two was significant ($b = -0.77$; $SE = 0.26$; $z = -2.96$; $p < .01$). This means that there was no significant difference between pretest and posttest scores overall, and that the RW training group was marginally more accurate than the NW training group. The significant interaction, together with the lack of an effect of Session, suggests that the effect of Session may differ between the two groups. In the light of this, we followed up on this interaction by splitting the dataset by training group and running two GLMERs, one by group, with response as dependent variable and only session as predictor. Results showed a significant effect of session for the RW training group ($b = -1.10$; $SE = 0.24$; $z = -4.60$; $p < .001$) but not for the NW training group ($b = -0.36$; $SE = 0.24$; $z = -1.47$; $p = .14$). This can be taken as evidence that only the RW training group significantly improved in their WPM performance from the pretest to the posttest session. Finally, as in the RW ID task, no evidence was found of accuracies in WPM in the posttest session having been influenced by the training status of the individual items. A GLMER with random intercepts by participants and items run on the WPM posttest failed to find

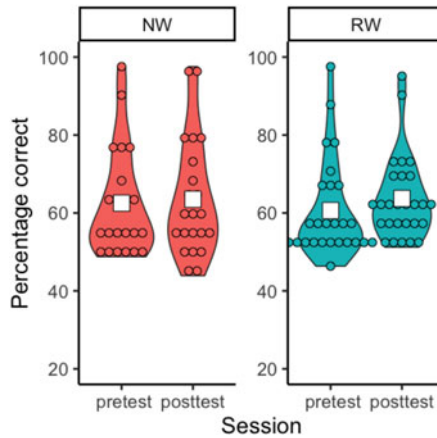


Figure 7. Distribution of accuracy scores in the LD by session (pretest-posttest) and training group (NW training—left, in red; RW training—right, in green). Filled circles represent individual scores, and white squares indicate the group means.

any effects of item type, training group, and, importantly, their interaction (all $p > 0.15$).

Finally, Figure 7 illustrates the individual variation found in the LD by session and by training group. The GLMER model for the LD included random intercepts by participants and items and random slopes for training group over items. The effect of session was not significant ($b = -0.10$; $SE = 0.15$; $z = -0.69$; $p = .49$) and neither were the effect training group ($b = 0.31$; $SE = 0.41$; $z = 0.76$; $p = .45$) and the two-way interaction ($b = -0.25$; $SE = 0.20$; $z = -1.22$; $p = .22$). Therefore, in this task, no significant improvement after training was observed, and the two groups did not perform significantly differently.⁵ As none of the stimuli in the LD had been used during training, no analysis testing potential training effects was conducted.

Discussion

The main goal of the present study was to determine whether an HVPT regime involving NWs was superior to a similar training regime using RW at improving perceptual accuracy of English front vowels. Moreover, it was assessed whether perceptual gains applied only to pre-lexical tasks or also extended to phono-lexical tasks. To this end, two comparable audiovisual training regimes were designed with careful attention to methodological detail. A battery of tasks, ranging from those requiring more attention to phonetic form to those requiring more attention to meaning, was used at pretest and posttest to assess gains in tests varying in the degree of (pre-)lexical processing they demand. Overall, the results of the current study generally show an improvement in vowel perception across tasks and training modalities. However, some differences were observed. Even though both groups showed significant gains from pretest to posttest, the RW group outperformed the NW group, particularly with the WPM and RW ID task, whereas no significant differences between groups were observed for NW ID.

When it comes to the results obtained in the identification tasks (pre-lexical), different outcomes were observed for each group and task. Specifically, results showed that, although both groups identified RW stimuli significantly more accurately at posttest, the RW group improved to a significantly greater extent (Gains: NW Group: 11%; RW Group: 17%), as indicated by the Group*Test interaction, and reached a very high score at posttest (94%). This finding is further supported by the generalization analysis, as there was no significant difference between how learners performed with trained and untrained items at posttest, even when the RW group had been trained with a part of the items in the task. As for the identification of NW, both groups improved to a similar extent. One would expect the NW group to outperform the RW group, since they were trained with this type of stimuli and a subset of testing items was included in the training. However, this was not the case; in fact, the opposite trend appeared in the descriptive results, as the RW group showed greater numeric gains than the NW group (16% vs. 11%). Regarding item type, the NW group also seemed to improve more with familiar (i.e., trained) items than with new ones. Needless to say, all items were new for the RW group, once again supporting the effectiveness of the training modality with RW.

On the whole, the aforementioned results indicate that RW training was more effective than NW training for improving the identification of the /i:/-/I/contrast, even with NW (cf. Mora, 2021; Mora-Plaza *et al.*, 2022). Consequently, our findings challenge the claims of some previous works that NW are superior—or equal to RW—for L2 phonetic training. Instead, they suggest that RW training can not only lead to larger improvements but also to broader generalization, while NW training appears to yield smaller and more item-specific gains. A likely explanation is that lexical activation helps integrate perceptual improvements into the learners' linguistic system, which goes against previous claims that accessing the lexicon may in fact be detrimental to phonetic learning (Mora-Plaza *et al.*, 2022; Thomson & Derwing, 2016). In addition to this, the use of audiovisual stimuli as well as the higher motivation associated with RW training (Carlet, 2017) may also have contributed: videos provide a more naturalistic learning context, and increased motivation can enhance focus and training outcomes. Note, however, that motivation and engagement were not directly measured in the current study and therefore remain proposed explanatory factors rather than empirically established ones. Another possible explanation for these discrepancies may be related to the training tasks. Although Mora (2021) and Mora-Plaza *et al.* (2022) combined discrimination, identification, and immediate repetition, the present study used only phoneme identification. Moreover, this study focused on the /i:/-/I/contrast, a vowel pair with high functional load, whereas Mora (2021) and Mora-Plaza *et al.* (2022)'s, trained /æ/-/ʌ/. Exposure to multiple productions of the high vowels during and outside the training, together with increased awareness in the first sessions, may have facilitated form–meaning mappings.

The results of the WPM align with those from the ID tasks. As explained above, the WPM used the same stimuli as the RW ID task but assessed how learners link these forms to specific meanings in the L2. To our knowledge, no previous studies have tested phono-lexical learning in this manner, and only a few have examined the effects of phonetic training on lexical tasks (Melnik & Peperkamp, 2021; Rebuschat

et al., 2025). Our results show that the RW group performed significantly better at posttest than at pretest (gain: 13%), while the NW group showed only minor, non-significant gains (3%). The relevance of this finding is amplified by the absence of differences between trained and untrained stimuli for the RW group, as this suggests that gains were not just a by-product of extended practice with a reduced set of items. The NW group's limited improvement may in part relate to the fact that none of the items from this task were included in their training, and this would align with patterns for the ID tasks, as it could again be considered evidence that training gains with NW are more task-and context-specific.

Overall, findings from the WPM are consistent with Melnik and Peperkamp (2021), who found that phonetic training enhanced spoken word recognition of known words when training used RW. By contrast, Rebuschat et al. (2025), who used novel word stimuli (i.e., functional nonwords), reported no improvement. This comparison should be interpreted cautiously, however, since Rebuschat et al. examined cross-situational novel word learning rather than recognition of known words, and the task therefore imposed different processing demands from both Melnik and Peperkamp's study and the present one. Accordingly, although the difference in findings is suggestive, more research is needed before firm conclusions can be drawn about how training may differentially affect known word recognition and novel word learning. Nevertheless, the two studies focusing on known word recognition support the view that learners can benefit from improved pre-lexical identification when retrieving form-meaning mappings during L2 lexical processing. In addition, the lack of a trained vs. untrained effect and the observed generalization suggest that learners in the RW group did not simply improve at identifying the trained phonemes in the trained items and this helped them link trained forms to the visual referents of the new task. Instead, the results are better explained by an overall improvement in their ability to identify the vowels across stimuli, which translated into a more reliable sub-lexical perception of the contrast when tasked with accessing their lexicon and long-term word representations in the WPM. This sheds new light on the generalizability of L2 phonetic training, as it helps mitigate concerns raised from previous studies that report mismatches between L2 pre-lexical and lexical processing (e.g., Amengual, 2016; Llompарт, 2021; Llompарт & Reinisch, 2019; Llompарт et al., 2025). While such mismatches do occur, the present findings suggest that they do not prevent that training at the pre-lexical level yields improvements at the lexical level, at least provided that the right conditions are met.

Finally, regarding the LD, only minor gains were observed: the NW groups barely improved by 1% from pretest to posttest, whereas the RW did so by 3%.⁶ No significant differences from pre-to posttest were thus found for this task. The null result of the current study can be explained by the difficulty of the task itself as well as methodological aspects. A LD involves the rejection of a lexical form that contains a competitor vowel. Two main issues may arise from this: (i) the fact that identifying such phonetic detail in an untrained task is challenging (ii) the listener may be able to reconstruct the lexical form from the remaining phonological information and thus tend to accept it as an existent word. The latter may in fact be reinforced by the fact that L2 speakers are often exposed to accented L2 speech (including their own), which may in fact present similar errors to the ones displayed

in the items we expected them to reject. Previous research has indeed pointed out the difficulty of this task and showed how lower scores tend to be obtained in it than in tasks that focus on form (e.g., Amengual, 2016; Darcy *et al.*, 2013; Llompарт, 2021; Llompарт & Reinisch, 2019; Llompарт *et al.*, 2025).

The null result we observed is interesting, nonetheless, as Melnik and Peperkamp (2021) also used an LD to test phono-lexical transfer and did find improvement. This difference may stem from the fact that their study examined the perception of English word-initial/h/by French learners of English, which already showed fairly high accuracy at pretest. The present study, however, focused on /i:/-/I/, a difficult vowel contrast that showed low pretest accuracy (62–63%). In fact, previous studies have reported that even advanced L2 learners obtain low accuracy rates in this task when it involves vowel contrasts, especially when rejecting nonwords (Llompарт, 2021; Llompарт *et al.*, 2025; Barrios & Hayes-Harb, 2021). This may have to do with the fact that vowels are produced less categorically than consonants and, therefore, learners are exposed to greater variability in their production, which may impact their performance when making categorical judgments of lexicality. In addition to this, a second methodological difference with Melnik and Peperkamp (2021) is training duration: their study included eight sessions of 30 minutes each, while ours involved 6 shorter sessions of 10–15 minutes. Thus, the combination of a demanding task from a processing point of view, a potentially more difficult L2 contrast and shorter training may have resulted in a challenge too great for the present training to yield measurable effects.

Limitations and further research

The current study is not without limitations. First, although the ID tasks were fully comparable—identical in speaker, speaker distribution, number of sessions, items, trials, and phonotactic structure—the lexical tasks are different in nature and thus harder to compare to the ID tasks and to each other. Regarding the WPM, although the same items as in RW identification were used, differences arise from task typology; there is a delay in the presentation of the response options: in the ID tasks the IPA symbols appeared on the same screen as the video, whereas in the WPM the video was shown first and the response options appeared only on the next screen, without the possibility of replaying it in the second screen. As for the LD, all stimuli were new, and although generalization was found in the ID task and WPM, this may have influenced LD results. Finally, the LD is different in nature, as it requires acceptance or rejection of words and pseudowords with a competing phoneme, while the ID and WPM involve matching form with a particular response option—a phonetic symbol in the case of the former, a word referent in the latter.

Finally, the current study has proven the efficacy of using lexical items (i.e., RW) to train both form and, to a lesser extent, meaning, and thus points to further research involving the lexicon in phonetic training. The fact that the possible activation of the lexicon was positive for phonetic learning and was transferred to a phono-lexical task (i.e., WPM)—but not to the (LD)—opens the door to a type of perceptual training that engages the lexicon even to a greater extent. Thus, further research will involve a similar perceptual training regime, this time using a lexical

task, such as a WPM (“Gorba, Llompart & Prieto (in preparation-b). Phonetic vs. Phono-lexical High Variability Training: effects on sound identification and word recognition.” Gorba et al., in preparation-b). Moreover, the fact that real concepts are trained is expected to enhance motivation to a greater extent. Another interesting avenue for further research would be a combined training regime in which a more pre-lexical approach is taken first (for instance, an ID task like the one presented in the current study) followed by a more lexical approach (with a WPM, for instance) that would help learners update their phono-lexical mappings efficiently.

Summary and conclusions

The current study has assessed the effects of an audiovisual HVPT regime involving NW, as opposed to a comparable training regime with RW, on the perception of English front vowels by L1-Catalan/Spanish speakers. Perceptual tasks focusing on form (i.e., identification tasks for NW and RW) and phono-lexical tasks (a WPM and a LD) were administered at pretest and posttest. The training regime included 6 sessions involving a phoneme identification task, either with NW or RW stimuli.

Results revealed that the RW group overall outperformed the NW group, with key significant differences observed in RW identification and WMPT. Moreover, generalization results confirmed that learning was robust enough to transfer to new speakers, stimuli, and even to a phono-lexical task (WPM) when the training was administered in the RW condition. It should be noted here, however, that these benefits are likely modulated by factors such as learner proficiency, phonetic contrast, and task demands, which may account for some of the variability observed across the broader phonetic training literature. Furthermore, the study also builds on and extends previous research highlighting the greater difficulty of lexical tasks compared to pre-lexical ones (Amengual, 2016; Llompart, 2021; Llompart et al., 2025), as evidenced by lower scores in the phono-lexical tasks relative to the identification tasks.

In summary, the current study provides clear evidence that RW training leads to stronger perceptual and phono-lexical gains for difficult vowel contrasts—such as /i:/-/I/—than NW training, even when methodological factors and phonetic focus are controlled. This contrasts with some earlier studies that found NW training more effective and the view that a stronger focus on form—reducing lexical influence—yields larger gains (Mora-Plaza et al., 2022; Thomson & Derwing, 2016). Instead, the present results align with studies showing that RW training is effective and, crucially, may promote transfer to word recognition (Melnik & Peperkamp, 2021).

Replication package. The dataset analyzed in this article and the code to reproduce the analyses reported are available at <https://osf.io/y94z6> (Open Science Framework).

Acknowledgements. The authors would like to acknowledge the financial support awarded by the Generalitat de Catalunya (2021 SGR 00922) and the Spanish Ministry of Science and Innovation (PID2021-123823NB-I00). The first author’s contribution was additionally supported by the Spanish Ministry of Universities (Margarita Salas Fellowship, ID-706211), as was the second author’s (Beatriz Galindo Junior Research Fellowship, BG22/00161). We would also like to thank the editor and two anonymous reviewers for their helpful suggestions.

Notes

- 1 Preliminary production results show smaller improvement overall, with significant gains only for the RW group—specifically for tasks involving RW stimuli—while the NW group did not improve significantly for any task or stimulus type.
- 2 Forty-one stimuli were used in Training Session 3 because one stimulus from Speaker 7 was invalid. A high-quality recording for the RW *tech* could not be obtained, so it was excluded. To maintain comparability across training modalities, the stimulus *tedge* was also removed from Session 3 in the NW training.
- 3 The current study was assessed and approved by the Institutional Committee for Ethical Review of Projects at Universitat Pompeu Fabra (reference number: 283).
- 4 The datasets analyzed in this article and the code to reproduce the analyses reported are available at <https://osf.io/y94z6>, together with the study's test and training materials (Open Science Framework).
- 5 As mentioned above, the datasets from the WPM and LD tasks included both match-real word trials and high-vowel mismatch and nonword (i.e., high-vowel substitution) trials. Previous studies, including our own (Amengual, 2016; Darcy & Thomas, 2019; Llompарт, 2021; Llompарт et al., 2025), have shown a preference for analyzing only mismatching and nonword trials, respectively, as these provide clearer evidence of the phonological encoding of the L2 contrasts in the lexicon (see Llompарт et al., 2025). While here we opted for keeping both trial types to be better able to compare performances across all 4 tasks, we ran the same analyses on the effects of Session and Training Group that are reported in the body of the article but on WPM's mismatch trials only and LD nonword trials only to ensure that results were consistent when we restricted our scope to these smaller datasets. Indeed, the same interaction between Session and Training group reported in the main text was also present for the WPM, while no significant effects were found for the LD task.
- 6 An important avenue for further exploration across all tasks in the present study, but especially in light of the small gains observed in the LD task, is the role of individual differences. More specifically, it would be valuable to examine whether learning gains depend on learners' initial pre-lexical identification acuity and whether improvement during training is itself modulated by those initial baselines. Such analyses could offer a more nuanced understanding of the relation between pre-lexical and lexical recognition. However, given the scope of the present study and the need to keep the group-level analyses clearly contextualized, we leave this question for future work.

References

- Aliaga-García, C. (2017). The effect of auditory and articulatory phonetic training on the perception and production of L2 vowels by Catalan-Spanish learners of English [Doctoral dissertation, Universitat de Barcelona].
- Aliaga-García, C., & Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. Recent research in second language phonetics/phonology: Perception and production, 231.
- Amengual, M. (2016). The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics*, 37(5), 1221–1251.
- Baigorri, M., Campanelli, L., & Levy, E. S. (2019). Perception of American–English vowels by early and late Spanish–English bilinguals. *Language and speech*, 62(4), 681–700.
- Barrios, S., & Hayes-Harb, R. (2021). L2 processing of words containing English/æ/-/e/and/l/-/ɪ/contrasts, and the uses and limits of the auditory lexical decision task for understanding the locus of difficulty. *Frontiers in Communication*, 6, 689470.
- Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal*, 30(1), 177–194.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., . . . & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, 12(1), 2.
- Best, C. T., Tyler, M., Bohn, O., & Munro, M. (2007). Nonnative and second-language speech perception. *Language Experience in Second Language Speech Learning*, 17, 13–34.
- Carlet, A. (2017). L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study. [Doctoral dissertation, Universitat Autònoma de Barcelona].

- Carlet Foresti, A., & Cebrian, J. (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. A sound approach to language matters: In honor of Ocke-Schwen Bohn.
- Cebrian, J. (2019). Perceptual assimilation of British English vowels to Spanish monophthongs and diphthongs. *The Journal of the Acoustical Society of America*, **145**(1), EL52–EL58.
- Cebrian, J. (2021). Perception of English and Catalan vowels by English and Catalan listeners: A study of reciprocal cross-linguistic similarity. *The Journal of the Acoustical Society of America*, **149**(4), 2671–2685.
- Cebrian, J., Gorba, C., & Gavalda, N. (2021). When the easy becomes difficult: Factors affecting the acquisition of the English/i:/-/I/contrast. *Frontiers in Communication*, **6**, 660917.
- Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, **8**(3), 372–420.
- Darcy, I., Llompарт, M., Hayes-Harb, R., Mora, J. C., Adrian, M., Cook, S., & Ernestus, M. (2025). Phonological processing and the L2 mental lexicon: Looking back and moving forward. *Studies in Second Language Acquisition*, **47**(1), 361–387.
- Darcy, I., & Thomas, T. (2019). When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Language and Cognition*, **22**(5), 1141–1159. <https://doi.org/10.1017/S1366728918001050>
- Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, **22**(6), 680–689.
- Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, **26**(4), 551–585.
- Escudero, P., & Chládková, K. (2010). Spanish listeners' perception of American and Southern British English vowels. *The Journal of the Acoustical Society of America*, **128**(5), EL254–EL259.
- Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, **25**(4), 437–470.
- 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: Cambridge University Press.
- Fullana Rivera, N., & MacKay, I. R. (2003). Production of English sounds by EFL learners: The case of /i/ and /i/. In Proceedings of the 15th international congress of phonetic sciences (pp. 1525–1528).
- Gorba, C. (2019). Bidirectional influence on L1 Spanish and L2 English stop perception: The role of L2 experience. *The Journal of the Acoustical Society of America*, **145**(6), EL587–EL592.
- Gorba, C., & Cebrian, J. (2021). The role of L2 experience in L1 and L2 perception and production of voiceless stops by English learners of Spanish. *Journal of Phonetics*, **88**, 101094.
- Gorba, Llompарт & Prieto (in preparation-a). Transfer effects from speech perception to production: Perceptual training with real words boosts generalization to (L2) production abilities.
- Gorba, Llompарт & Prieto (in preparation-b). Phonetic vs. Phono-lexical High Variability Training: effects on sound identification and word recognition.
- Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, **24**(4), 495–522.
- Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, **37**(5), 1175–1199.
- Institut d'Estudis Catalans. Secció Filològica. (1999). *Proposta per a un estàndard oral de la llengua catalana: Morfologia* (Vol. 2). Institut d'estudis catalans.
- Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, **126**(2), 866–877.
- Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, **38**(4), 569–587.
- Ladefoged, P., & Johnson, K. (2010). *A Course in Phonetics*. Wadsworth Publishing Company.
- Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, **26**(2), 227–247.

- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, 94(3), 1242–1255.
- Llompарт, M. (2021). Lexical and phonetic influences on the phonolexical encoding of difficult second-language contrasts: Insights from nonword rejection. *Frontiers in Psychology*, 12, 659852.
- Llompарт, M., Gorba, C. & Prieto, P. (2025). Revisiting the link between second-language sound identification and word recognition with an eye on methodological similarity. *Studies in Second Language Acquisition*, 47(5), 1456–1468. <https://doi.org/10.1017/S0272263125101113>
- Llompарт, M., & Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech*, 62(3), 594–622.
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/t/and/l/: A first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886.
- Mathias, B., & von Kriegstein, K. (2023). Enriched learning: Behavior, brain, and computation. *Trends in Cognitive Sciences*, 27(1), 81–97.
- Melnik, G. A., & Peperkamp, S. (2021). High-variability phonetic training enhances second language lexical processing: Evidence from online training of French learners of English. *Bilingualism: Language and Cognition*, 24(3), 497–506.
- Mora, J. C., Ortega, M., Mora-Plaza, I. & Aliaga-García, C. (2022). Training the pronunciation of L2 vowels under different conditions: The use of non-lexical materials and masking noise. *Phonetica*, 79(1), 1–43.
- Mora, J. C., Ortega, M., Mora-Plaza, I. & Fouz-González, J. (2021) The Role of Attention Control and Stimuli Lexical Status on High Variability Perceptual Training. Paper presented at the 8th International Conference on Experimental Phonetics (CIFE 2021), 28–30 June 2021, Universitat de Girona, Girona, Spain.
- Mora-Plaza, I., Ortega, M., & Mora, J. C. (2022). High-variability phonetic training under different conditions: Individual differences in auditory attention control. In V. G. Sardegna & A. Jarosz (Eds.), *Theoretical and practical developments in English speech assessment, research, and training: Studies in honour of Ewa Waniek-Klimczak* (pp. 241–260). Springer International Publishing.
- Morrison, G. S. (2008). L1-Spanish speakers' acquisition of the English/i/-/I/contrast: Duration-based perception is not the initial developmental stage. *Lang Speech*, 51(4), 285–315.
- Morrison, G. S. (2009). L1-Spanish speakers' acquisition of the English/i/-/I/contrast II: Perception of vowel inherent spectral change. *Lang Speech*, 52(4), 437–462.
- Morrison, G. S. (2012). Perception of natural vowels by monolingual Canadian-English, Mexican-Spanish, and peninsular-Spanish listeners. *Can Acoust*, 40(4), 29–40.
- Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied linguistics*, 29(4), 578–596.
- Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*, 39(1), 13–37.
- Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research*, 50(6), 1496–1509.
- R Core Team (2017) R: A Language and Environment for Statistical Computing. <https://www.R-project.org/>.
- Rallo Fabra, L., and Romero, J. (2012). Native Catalan learners' perception and production of english vowels. *J Phon*, 40, 491–508.
- Rallo Fabra, L. R., & Tyler, M. D. (2023). Predicting discrimination difficulty of Californian English vowel contrasts from L2-to-L1 categorization. *Ampersand*, 10, 100109.
- Rato, A., & Carlet, A. (2020). Second language perception of English vowels by Portuguese learners: The effect of stimulus type. *Ilha do Desterro*, 73, 205–226.
- Rebuschat, P., Correia, S. P., Ge, C., Llompарт, M. & Rato, A. (2025). The effect of phonetic training on (statistical) word learning. New Sounds 2025. University of Toronto: Toronto, Canada.
- Reyes, Y. P., & Hazan, V. (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. *Onomázein*, 51, 112–136.
- Roach, P. (2004). British English: Received pronunciation. *Journal of the International Phonetic Association*, 34(2), 239–245.

- Rocca, B., Llompart, M., & Darcy, I.** (2025). Phonological neighborhood density, phonetic categorization, and vocabulary size differentially affect the phonolexical encoding of easy and difficult L2 segmental contrasts. *Bilingualism: Language and Cognition*, **28**(3), 662–675.
- Smith, B. L., & Hayes-Harb, R.** (2016). Non-native speakers' acoustic variability in producing American English tense and lax vowels. *Journal of the Acoustical Society of America*, **139**(4_Supplement), 2163–2164.
- Thomson, R. I.** (2018). High variability [pronunciation] training (HVPT) A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, **4**(2), 208–231.
- Thomson, R. I., & Derwing, T. M.** (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le, I. Lucic, E. Simpson & S. Vo (Eds.), *Proceedings of the 7th pronunciation in second language learning and teaching conference* (pp. 88–97). Iowa State University.
- Uchiyara, T., Karas, M., & Thomson, R. I.** (2025). High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. *Studies in Second Language Acquisition*, **47**(3), 794–827. <https://doi.org/10.1017/S0272263125100879>
- Wong, J. W. S.** (2014). The Effects of High and Low Variability Phonetic Training on the Perception and Production of English Vowels/e/-/æ/by Cantonese ESL Learners with High and Low L2 Proficiency Levels. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech 2014* (pp. 524–528).

Cite this article: Gorba, C., Llompart, M., & Prieto, P. (2026). Are nonwords really superior to real words to train L2 sounds? Insights from an HVPT study on L2 English front vowels. *Applied Psycholinguistics*. <https://doi.org/10.1017/S0142716426100836>