Short-Session High Variability Phonetic Training Improves L2 Phonetic Contrasts Perception Without Additional Benefits of Background Noise

Xiaoli Zhu

Faculty of Humanities, Leiden University

5194VSCLI: Linguistics MA Thesis

Supervisor: Tim Joris LAMÉRIS

December 31, 2025

Abstract

High variability phonetic training (HVPT) has demonstrated to enhance the learning of non-native phonetic contrasts among adult second language (L2) learners, though the effectiveness fluctuates concerning different training conditions. Background babble noise, which is prevalent in real-life communication, has been suggested to improve perceptual robustness, yet its role in phonetic training remains unclear. The current study investigates whether integrating babble noise into the stimuli of a short-session HVPT paradigm improves L2 learners' perception of the English vowel contrast /ɪ/–/iː/. Adult Chinese-speaking learners of English were recruited and randomly assigned to either HVPT with babble noise (HVPT-N) or HVPT in quiet (HVPT), and their performance was evaluated by accuracy and reaction time (RT) measures in both trained conditions and generalization to untrained talkers and words. The results showed that, though adding background babble noise does not provide extra benefits under limited exposure, short-session HVPT effectively improves perceptual accuracy and efficiency in both training conditions. The findings further suggest that babble noise training may require longer or repeated exposure to become effective. These results have clear implications for L2 phonetic pedagogy: brief HVPT interventions can produce robust learning gains, while introducing background noise at early stages of phonetic category learning may be unnecessary or even unhelpful.

Keywords: babble noise, HVPT, phonetic contrasts, second language acquisition

Introduction

It is well established that adult second language (L2) learners usually encounter persistent difficulties when acquiring non-native phonetic contrasts, especially when the relevant distinctions of such speech sounds are not phonologically encoded in their first language (L1) (Archibald, 2021; Tavares et al., 2025). These perceptual difficulties have been widely reported in various language pairs and have a significant impact on downstream functions such as auditory comprehension, accurate pronunciation, and overall communicative performance (Best, 1995; Flege et al., 1997). Crucially, such difficulties do not only appear in idealized laboratory conditions, but also in real-life communicative situation. In real-life context, listeners are often exposed in sophisticated auditory settings with background noise such as babble noise in a café or even in a classroom. Background noise is prevalent in real-life listening contexts, which might affect the learners' perception effect of non-native phonetic contrasts. A large body of research show that background noise has a much greater impact on L2 listeners than on native language listeners, which may worsen existing perceptual barriers and limit the chances of successfully learning L2 speech (Cooke et al., 2008; Wang & Xu, 2021). Consequently, understanding how to effectively support L2 phonological perception—especially in noisy environments—is significant for both theoretical research and teaching practice.

High variability phonetic training (HVPT) is one of the most widely used methods for improving adult learners' perception of difficult L2 contrasts. Such variability encourages learners to abstract away from surface differences and to form more robust phonetic categories (Logan et al., 1991; Lively et al., 1993; Iverson et al., 2005). Some studies have shown that HVPT can improve both identification and discrimination across L2 contrasts and learner group, for example, the vowel contrast /ɪ/–/iː/, which is especially challenging for Chinese-English bilinguals (Wang, 1997; Brosseau-Lapré et al., 2013). However, HVPT outcomes depend on training parameters, including feedback, stimulus variability, task demands, and total exposure. This sensitivity has support to test HVPT under conditions that better reflect real-world listening environments (Lively et al., 1993; Barriuso & Hayes-Harb, 2018; Brekelmans et al., 2021).

Babble noise is particularly relevant in this context because it closely resembles daily communication. In such settings, listeners often hear 4 talkers during the experiment. Generally, speech perception for L2 learners in babble noise condition is usually more effortful than perception in quiet. This difficulty appears that even at relatively favorable signal-to-noise ratios (SNR) and reflects both acoustic masking and increased cognitive load (Scharenborg et al., 2019). What's more, some studies suggest that exposure to these challenging conditions may lead listeners to rely more on clear and diagnostic cues. This shift may, over time, strengthen perceptual robustness (e.g., Zhang et al., 2021). However, it remains unclear that whether such adaptation occurs in active, feedback-based training paradigms and whether structured HVPT can effectively harness such potential benefit.

Although HVPT has been shown to be effective, and interest in speech perception under adverse conditions has grown in parallel, these two lines of research have largely developed separately. Most HVPT studies have been carried out in quiet laboratory settings (Logan et al., 1991; Lively et al., 1993; Iverson et al., 2005). Research on speech perception in noise, by contrast, has focused mainly on passive listening tasks, rather than on active training that includes feedback (Cooke et al., 2008; Scharenborg & van Os, 2019). Because of this separation, it is still not clear whether learning outcomes or patterns of generalization change when HVPT is delivered in more realistic acoustic environments that include background noise (Brekelmans et al., 2021), which limits their ecological validity and leaves an open question: Will training performance and subsequent generalization ability differ when speech training takes place in more realistic auditory settings? Furthermore, short-session intensive training—a single training session or a small number of sessions with a total exposure time of less 10 minutes—has gained popularity in recent years due to its practicality and significant effectiveness. For example, a large-scale meta-analysis suggests that short training sessions in HVPT lead to great improvement in speech perception (Uchihara et al., 2025). However, it remains largely unexplored whether incorporating background noise into such short-session training modulates learners' ability to acquire and generalize difficult phonetic contrasts. This leaves some significant gaps: Does integrating background babble noise in HVPT amplify, hinder, or somehow modulate the learning effect of L2 learners in speech perception? And how does this effect work in short-session HVPT paradigms?

To fill these research gaps, the present study explores whether incorporating background noise into short-session HVPT paradigm can improve adult Chinese-speaking English learners' ability to distinguish between the English vowel contrast /ɪ/–/iː/. This study specifically aims to investigate whether the L2 learners who are exposed to a HVPT paradigm with babble noise (HVPT-N) have better acquisition effects than those who are trained in HVPT in a quiet environment (HVPT) by measuring and comparing the results of their accuracy and reaction time (RT) in the discrimination listening tasks before and after the training. The study is particularly focused on the learners' improvement of generalization ability to untrained talkers and novel lexical items containing the target phonetic contrasts. Theoretically, by examining the interaction between phonetic training and the auditory settings, this study intends to further understand how speech perceptual learning mechanisms function in ecologically effective conditions. Practically, the findings can also provide pedagogical guidance for L2 phonetic perception instruction, making the phonetic training process more closely resemble the noisy communication environments that L2 learners are often in. If HVPT containing background babble noise proves effective, it will provide a low-cost and easily accessible improvement solution for improving the efficiency and real-life applicability of L2 speech perception training.

Literature Review

Challenges of L2 Speech Perception and Non-Native Speaker Speech Discrimination

Theoretical Frameworks of L2 Speech Perception

L2 speech perception is greatly influenced by learners' L1 phonological experience, specifically their long-term exposure to and internalization of L1 phoneme categories and the associated acoustic cue weightings used to distinguish them. Two main theoretical frameworks, Perceptual Assimilation Model for Second Languages (PAM-L2) and the revised Speech Learning Model (SLM-r), provide complementary explanations for why some L2 speech is extremely challenging for adult learners. According to PAM-L2 proposed by Best and Tyler (2007), learners perceive L2 pronunciation by assimilating L2 phonemes to the closest L1 category. When both L2 phonemes can be equivalently mapped to a single L1 category, a single-category (SC) assimilation pattern is formed, resulting in extremely impaired discrimination. For Chinese-English bilinguals, the vowels /ɪ/ and /i:/ are commonly assimilated to the Chinese vowel /i/, and both are often perceived as equally good, or nearly equally good, exemplars of this single L1 category. This pattern is consistent with a category-goodness type of assimilation and leads to persistent difficulty in distinguishing the English /ɪ/–/i:/ contrast. Because learners do not have separate L1-based category boundaries to rely on, discrimination between the two vowels remains unreliable.

Within the framework of the revised Speech Learning Model (SLM-r), such difficulty is expected when two L2 sounds are perceived as too similar to an existing L1 category. In these cases, learners may engage in equivalence classification, in which both L2 phonemes are mapped onto the same L1 category, and this process can block the development of new phonological categories (Flege & Bohn, 2021, pp. 6–7). For Chinese-speaking learners, this situation arises because the Chinese vowel /i/ occupies an acoustic space that overlaps with both English /ɪ/ and /i:/ along important spectral dimension. And Chinese lacks a phonemic vowel length contrast, which further reduces perceptual separation between the two English vowels. As a result, /ɪ/ and /i:/ are perceived as insufficiently distinct from the Chinese /i/ category, leading to a reduced sense of acoustic distance and fewer opportunities for new category formation.

This perceptual overlap helps explain why Chinese-speaking learners often show unstable identification boundaries and lower discrimination accuracy for the /ɪ/–/i:/ contrast, particularly in tasks that require fine-grained perceptual judgments. Taken together, these mechanisms predict marked difficulty in both perception and identification of this contrast and point to the importance of structured perceptual training that draws learners' attention to the acoustic cues that reliably distinguish the two vowels.

Perception of the /ɪ/–/i:/ Contrast in Chinese-Speaking English Learners

The above theoretical predictions regarding /ɪ/–/i:/ discrimination by Chinese speakers have been supported by empirical studies. For instance, some studies show that compared to native English speakers who rely more on formant-based spectral cues like F1 and F2 when identifying /ɪ/ and /i:/, Chinese speakers rely more on vowel duration (Polka, 1992; Wang, 1997). Because vowel duration varies greatly across prosodic contexts, such as stress patterns, speech rate, and sentence position, relying on duration alone does not provide a stable basis for vowel category discrimination in Chinese. As a result, Chinese-speaking learners are likely to show less consistent category boundaries and more frequent misclassification of /ɪ/ and /i:/, especially in contexts where durational cues are weak or unclear. This account helps explain why Chinese learners continue to experience difficulty in accurately perceiving and identifying the /ɪ/–/i:/ contrast across different speakers and listening contexts.

Evidence from studies that focus on specific L1–L2 contrast pairs supports this interpretation. Research has shown that Chinese-speaking learners of English often respond more slowly, form weaker category boundaries, and show limited generalization to unfamiliar talkers or lexical items when perceiving challenging English vowel contrasts such as /ɪ/ and /iː/ (Xie et al., 2021). These findings should not be taken to imply a general perceptual disadvantage for Chinese speakers. Instead, when compared with English-speaking learners of Chinese—who are typically tested on different contrasts with different L1–L2 relationships—the observed differences point to contrast-specific perceptual difficulty that arises from the structure of the L1 phonological system. For this reason, Chinese-speaking learners provide a well-motivated population for investigating perceptual learning mechanisms and training approaches such as High Variability Phonetic Training (HVPT), particularly for contrasts marked by high perceptual similarity and strong L1 interference.

High Variability Phonetic Training (HVPT)

Mechanism and Theoretical Principles

High variability phonetic training (HVPT) is a widely used and well-supported method for improving second language (L2) speech perception. First introduced by Logan, Lively, and Pisoni (1991), HVPT exposes learners to multiple realizations of a target contrast that vary across talkers, lexical items, and phonetic contexts.

The basic idea behind HVPT is that variability in the input encourages learners to focus on acoustic cues that reliably signal the contrast, rather than on surface features tied to individual speakers. When the same contrast is encountered across different voices and phonetic environments, learners are pushed to move beyond talker-specific details and to develop more stable and generalizable phonetic categories. Empirical evidence strongly supports this mechanism. Lively et al. (1993) found that Japanese learner who were taught /r/–/l/ using high variation made significant progress, especially in untrained voices.

Feedback is another important factor needed to be investigated more for HVPT paradigms. Immediate corrective feedback results in error-based learning, where learners can improve their perception boundary with practice (Bradlow et al., 1997; Iverson et al., 2005). This mechanism corresponds to SLM-r theory, which highlights the necessity for one's attention to subtle acoustic difference to do efficient category recognition (Flege & Bohn, 2021, pp. 6–8). Thus, HVPT gives us a theory-based and supported framework for perceptual relearning.

Empirical Evidence on the Effects of HVPT

A substantial body of research has shown that HVPT is effective across a wide range of L2 phonetic contrasts and learner groups. For instance, Brosseau-Lapré et al. (2013) reported that HVPT led to clear improvements in French-speaking learners' perception of English tense–lax vowel contrasts, including /ɪ/–/iː/. These gains were linked to increased sensitivity to spectral cues that are critical for distinguishing the contrast. Cheng et al. (2019) reported temporal acoustic cues in auditory processing can promote the efficacy of HVPT, which further proving the role of cue salience in speech perception. Brekelmans et al. (2021) also found that HVPT considerably enhances participants' ability in recognizing and discriminating sounds. Together, the findings mentioned above can be served as powerful empirical evidence to support the effectiveness of HVPT as a phonetic perception training method.

The strongest evidence comes from the meta-analysis by Uchihara et al. (2025) bringing together a large body of research on HVPT conducted over more than four decades, which suggests moderate-to-large-sized effect in terms of how we perceive speech. They found that HVPT constantly improves L2 segment recognition, no matter what language or speech feature or learning setup.

Based on the meta-analysis conducted by Uchihara, Karas, and Thomson (2025), the success of HVPT appears to rely heavily on training variables such as talker variability, type of feedback, total exposure length, and other stimulus variables. As most HVPT studies use multi-session training paradigms (e.g., Logan et al., 1991; Lively et al., 1993; Iverson et al., 2005), there are still insufficient studies exploring the effect of short-session HVPT. This is very relevant to the present study because it is different from previous studies which test HVPT in multi-session training. This study uses a paradigm adapted in short sessions and intensive training.

Training Parameters that Affect HVPT Results

HVPT has been known to be impacted by various training parameters.

Talker variability is one of the most crucial aspects because listening to speech materials recorded with different speakers enables listeners to better abstract relevant cues and improve their performance in generalizing to novel speakers that they have never heard speak before (Lively et al., 1993; Zhang et al., 2021).

The type of feedback provided, also appears to influence the effectiveness of HVPT. One form that is widely used is immediate corrective feedback, which provides learners with trial-by-trial information about whether their response is correct or incorrect, usually presented right after each identification attempt. This type of feedback allows learners to notice perceptual errors and to make gradual adjustments to their category boundaries over repeated exposure. In this way, feedback supports the stabilization of newly developing phonetic categories. At the same time, the effectiveness of feedback-based learning is also shaped by the acoustic characteristics of the training stimuli. Take Cheng et al. (2019), for example, who showed that making sounds "temporal-acoustic-exaggerated" (p. 168) can make contrast-related cues more noticeable when we first start learning about them.

Training duration is another important factor in HVPT. Meta-analytic work has shown that longer and more distributed training schedules, often spread across multiple sessions over several days or weeks, tend to produce larger and more lasting perceptual improvements than very brief interventions (Uchihara et al., 2025). In much of the HVPT literature, this type of long-term training typically involves several hours of total exposure, delivered across repeated sessions (e.g., Logan et al., 1991; Lively et al., 1993).

At the same time, fewer studies have looked in a systematic way at short-session HVPT. The evidence that does exist suggests that short-term training—usually defined as a single session or a small number of sessions with less than one hour of total exposure—can still lead to immediate perceptual gains. These gains are most often found in task-specific measures, such as higher identification accuracy or faster reaction times on the trained task, rather than in long-term retention or wide-ranging generalization. This contrast provides the motivation for the present study, which uses a short, intensive HVPT design to focus on immediate learning and patterns of generalization under limited training time.

Finally, an important consideration, which is often overlooked, is the listening environment. Although real-world L2 listening training takes place in noisy environments, almost all auditory-speech processing training studies have been done under quiet laboratory conditions. This gap between typical training settings and everyday listening environments limits the ecological validity of many existing findings and calls into question how well results from laboratory-based training extend beyond the lab. For this reason, empirical work that directly incorporates background noise into speech perception training is needed to clarify how perceptual learning unfolds under more realistic listening conditions.

Speech Perception and Adaptive Learning in Noise

Why Noise Poses a Challenge to L2 Listeners

Listening in a noisy environment poses a unique challenge for L2 learners relative to those who are native speakers because background noise impacts non-native speech perception and processing efficiency more than it does native speech. Specifically, noise disrupts L2 listeners' accuracy and speed on spoken-word recognition and phoneme identification much more than it does with native listeners when overall intelligibility is still decent.

For example, Cooke, García Lecumberri, and Barker (2008) explored speech recognition performance with multi-talker babble-noise by native English listeners and non-native learners of English with different L1 backgrounds. Participants recognized English words that were part of sentences that were played with various signal-to-noise ratios (SNR) levels. As noise got louder, non-natives had much steeper drops in keyword understanding compared to natives, showing less listening success when it was hard to hear. And the most important is that this drawback appears already at decent SNRs, meaning that L2 listeners are also disturbed more by background noise when comprehending speech.

Non-native disadvantage is caused by several factors that work together. It's because the phonological category is weaker in the L2. The automatic mapping from acoustic input to the lexical form is also less, which means there will be more use of cognitive resources like attention and working memory when understanding speech. As a result, background noise takes up too much processing of L2 listeners, which means that L2 listeners can't extract fine-grained acoustic information from the speech signal.

A related line of evidence is from Scharenborg and Van Os (2019) who investigated speech intelligibility in noise for native and non-native listeners through the measurement of word recognition accuracy as a function of signal-to-noise ratio. SNR, stands for Signal-to-noise ratio, is the intensity comparison between target speech signal and background noise, commonly expressed by dB of their difference. These were poorer than those of the native listeners. Their performance was significantly worse even for SNRs as high as 10 dB, where the speech could be heard over the noise. This finding shows that difficulties with noise are not only in the acoustic problems for a given non-native listener, but also in speech perception and language processing.

In summary, these results imply that a background noise environment might prevent L2 learners from reaching subtle contrast-relevant spectral information necessary for telling apart phonetic contrasts like English /ɪ/ and /i:/. When noise makes it harder for our senses and brain to pick up sounds, we might use easier-to-notice but not as helpful hints when trying to tell different sounds apart and put them into groups.

Perception Adaptation to Noise

It's commonly believed that noises will hurt our perception but recently it's shown that some noises can help people adjust to what they're hearing. Specifically, Zhang et al. (2021) pointed out that the extended exposure to the multi-talker babble noise made the native listeners switch from weighting the dimension of the cue which is temporally unsteady like duration and amplitude to stable cues, such as the formant pattern. This was an adaptive reweighting linked to better speech perception amid noise. Noise exposure sometimes improves perceptual robustness. It's not uniform degradation.

And importantly, most works looking at noise-related perceptual adaptation have used native listeners or passive listening with no actual training or feedback. So, it is still unknown if the same adaptive mechanism takes place when people actively learn L2 perception through some training, for example, in HVPT paradigm.

Adding Noise into HVPT: Insufficient Evidence and Research Gaps

Ecological Limits of Current HVPT Studies

These differences raise the question as to whether previous studies on HVPT conducted under quiet laboratory conditions (i.e., without ambient noise) are validly representative of the real-life listening situations that listeners face, which consist of much more ambient noise than in the "quiet" HPVT studies. A growing amount of research finds that background noise, specifically multi-talker babble, impacts L2 listeners more than native listeners even when speech is still perceptible (Scharenborg & van Os, 2019). This heightened vulnerability to noise worsens perception for L2 contrasts without strong L1 equivalents like English /ɪ/–/i:/ for Chinese speakers. In such cases, access to those useful but more finely grained spectral cues is further degraded, which increases misperception and unclear categories.

Previous research has shown the impact of background noise on L2 speech perception (Cooke, 2006; García Lecumberri et al., 2010; Scharenborg & van Os, 2019). However, most HVPT studies do not consider the influence of the acoustic setting during training on learning (Logan et al., 1991; Lively et al., 1993; Brosseau-Lapré et al., 2013). In the specific case that is relevant here, work to date has focused much more on perceiving performance in noise or in quiet at test and less on whether noise is present during training. Therefore, it is still unclear if training under quiet conditions translates to real world listening under noise, or whether adding noise to HVPT alters training and learning. There is an unresolved matter, that requires to study the large gap systematically through experimentation.

Another limitation with the papers currently available is that no one has integrated noise into the existing paradigms of HVPT. A large body of research has investigated speech perception in noise—most commonly using multi-talker babble to model real-world listening conditions—whereas HVPTs studies have always been performed under quiet (i.e., non-noisy) training conditions. So far, very little work has tried to train in the presence of background noise to see if the adverse listening conditions transfer negatively to important HVPT outputs like perceptual gains and generalization to untrained talkers or other lexical items (Lively et al., 1993).

It is useful to separate out different types of noise in the literature. Stationary noises like white noise produce masking primarily at energy levels, whereas multi-talker babble produces both energetic and informational masking, so it is most ecologically relevant for L2 listening. Although some studies have shown that long-term exposure to noisy environments—mainly babble noise—can result in perceptual cue weighting adaptation, most of the research studies conducted have been on native listeners or passive listeners with no actual training or feedback. For instance, Zhang et al. (2021) found that when presented with extended exposure to babble noise, native listeners relied on stable, spectral cues (e.g., formant structure) more than temporal cues (e.g., duration, intensity), leading to better speech perception in noise.

But at the same time, we do not know whether such adaption mechanisms would work on the active L2 learning that has been taking place through the structured training such as HVPT itself. Moreover, the evidence in the literature is mainly related to either a native speaker's adaptation or perceptual compensation to noise (not learning-induced category formation and generalization in L2s) And then the degree to which noise throughout training either promotes or hampers learning through HVPT stays an empirical issue.

Unverified Effects of Short-Session HVPT paradigms with Babble Noise

One of the most obvious gaps in the HVPT literature is the lack of work done on short-session training paradigms. Short-Session HVPT—typically defined as a single session or small number of sessions with training less than 1 hour—has gained interest owing to the practicality and potential of HVPT for quickly gaining percepts. Meta-analytical data shows that training duration is important to HVPT results—long, scattered training times over numerous sessions and days or weeks, with several hours total exposure, seem to yield more firm, lasting learning effects (Uchihara et al., 2025)

But short-sessions of HVPT have shown more immediate improvements which tend to be at least somewhat task-specific, like better ability to identify a speaker or faster responding to their words within the trained talker-word pair compared to generalization across different untrained talkers or words. For instance, brief HVPT interventions improved performance on post-test identification of trained stimuli but showed weaker generalization than those observed with HVPT procedures across multiple session as reported in classic HVPT studies (e.g., Logan et al., 1991; Lively et al., 1993). Under such restricted exposure, learning would also be more susceptible to external acoustic interference and higher cognitive load as learners do not get much chance to consolidate their new phonetic representations. Thus, it is yet to be determined whether background noise supports or hinders adaptive cue reweighting in perceptual learning of short-session HVPT.

Babble noise, compared to “quiet” non-noise conditions, might require more cognitive processing for an L2 learner since babble noise requires more energy masking as well as informational masking. under conditions of this type, students’ capacity to obtain close-up spectral information may well be hindered; so, it’s hard for them to see tiny acoustic dissimilarities and make the new phonetic groups persistent. This effect would be especially severe when perceptual contrast is hard to discriminate, such as for Chinese speakers, where English /ɪ/–/iː/ contrast already exhibits weak category separation under quiet conditions (Cooke et al., 2008).

At the same time, accounts of perceptual adaptation and cue reweighting predict the opposite. In terms of specific content, the cue reweighting frame indicates that exposure to adverse listening environments increases the weight on more acoustically reliable and invariant cues, e.g., formant structure, while reducing the weight towards less stable dimensions like duration or intensity (Zhang et al., 2021). From this vantage point it’s possible that some situations might result in more effective perceptual solutions to the learner than noise could impede.

The two theoretically motivated predictions of noise being a source of additional perceptual and cognitive load versus noise inducing adaptive cue reweighting result in different predictions about the influence of noise in short-session HVPT. As the training is brief and the exposure is low as with short-session training, current evidence cannot yet definitively predict if noise would help or hinder immediate perceptual learning and generalization. If you want to get rid of such tensions, you will have to directly try with experiments.

In short, the literature reveals three major research gaps. The first aspect: HVPT was studied in quiet places most of the time, which limited its ecological validity. Second, although speech perception in noise has been extensively studied, relatively little research has examined the effects of incorporating ambient noise directly into HVPT training paradigms, leaving it unclear whether noise facilitates or hinders perceptual learning and generalization. Third, it is unknown how noise affects short-session HVPT tests. This study directly fills this gap in the field by conducting a short-term HVPT paradigm quiet and noisy conditions (HVPT-N), to evaluate the impact on accuracy, RT and generalization of adult Chinese speaking English language learning to new talkers and lexical items.

Methods

Participants

A total of 40 Chinese-speaking adults (28 females, 12 males) participated the study[1], which was conducted online on *Gorilla.sc* (https://app.gorilla.sc/admin/home). Participants recruited via word of mouth, personal networks, university database, whichever applies. All

[1] Ten additional people also participated but excluded from further analysis because they showed very low pre-test performance (6 participants), too many RT outliers (3 participants) or because they reported technical problems during the online experiment (1 participant).

participants provided informed consent prior to participation. The study protocol was approved by the institutional ethics committee, and all data were anonymized and stored securely following ethical guidelines. The mean age of the sample was 22.18 years (*SD* = 3.34).

Participants had an average of 12.64 years of English learning (*SD* = 4.29), and most (55%) self-rated their English proficiency at the "intermediate" level. Twenty participants (50%) indicated that they had previously taken an English pronunciation class or engaged in pronunciation training activities, whereas 20 participants reported no such experience.

Participants were randomly assigned to one of two groups according to the experiment version: HVPT or HVPT-N. Group assignment was evenly distributed randomly, with 20 participants in the HVPT group and 20 participants in the HVPT-N group. The two groups did not differ substantially in age, gender distribution, English learning years, or any of the questionnaire-based background characteristics (as shown in Table 1). Difference in the age, sex ratio, years of learning English and pre-test baselines between two groups were not statistically observed and controlled at the start of the experiment.

Data quality was ensured through a series of predefined exclusion criteria. Trials with extremely long reaction times were excluded if the RT exceeded 2.5 standard deviations above each participant's own mean reaction time. Participants were also excluded if more than 10% of their trials were missing or invalid. In addition, participants who reported technical problems during the experiment—such as audio interruptions or unstable internet connections—were excluded from further analysis. The last sample was 40 subjects, randomized into two groups: 20 HVPT (silent training) and 20 HVPT-N (HVPT in presence of background babble noise).

Table 1

Participant Characteristics in the HVPT and HVPT-N Groups (N = 40)

Variables	HVPT	HVPT	Total
Age (years), M (SD)	22.35 (3.50)	22.00 (3.25)	22.18 (3.34)
Gender (F/M)	14 F, 6 M	14 F, 6 M	28F, 12 M
Native language	Chinese	Chinese	Chinese
Age of first exposure to English (years)	—	—	9.8
Context of exposure to English: At school	—	—	24 (60%)
Context of exposure to English: Both (school + other sources)	—	—	16 (40%)
Years of English learning, M (SD)	12.80 (4.52)	12.48 (4.12)	12.64 (4.29)
English proficiency: Beginner	—	—	12 (30%)
English proficiency: Intermediate	—	—	22 (55%)
English proficiency: Advanced	—	—	6 (15%)
Pronunciation training (Yes/No)	9/20	11/20	20 / 40
Group assignment	Version A (HVPT)	Version B (HVPT-N)	—

Note. Values represent means with standard deviations or frequencies with percentages where appropriate. Pronunciation training coded as 1 = Yes, 2 = No. Age of first exposure aggregated across all participants; group- specific values are not applicable due to questionnaire structure.

Stimuli

The target contrast in this study was the English high front vowel pair /ɪ/–/iː/, as in *bit–beat* (see Appendix 1). This contrast is well known to be difficult for Chinese-speaking learners due to the absence of an equivalent phonemic distinction in Chinese and the heavy reliance on spectral cues (F1–F2 values) and durational differences in English, which differs from Chinese's primarily spectral vowel system (Huang & Johnson, 2010; Hao, 2012). Empirical studies consistently report that Chinese listeners show persistent difficulty perceiving and categorizing this contrast even at advanced proficiency levels (Escudero & Boersma, 2004; Chang, 2018). Therefore, /ɪ/–/iː/ was deemed as an ideal target for evaluating perceptual learning in HVPT paradigms.

The lexical items used in this study consisted of a set of minimal pairs containing the target vowel contrast (e.g., *bit–beat*, *pick–peak*, *hip–heap*). These items were selected based on (a) prior HVPT literature (e.g., Lively et al., 1993; Bradlow et al., 1997); (b) their familiarity and high frequency in learner vocabulary lists; and (c) their phonotactic simplicity to avoid confounding effects from consonantal context. The complete list of stimuli is provided in Appendix 1.

All pre-test and training stimuli were sound files generated on an online speech production website *Hearling* (https://hearling.com/clips/new), a validated online training platform that provides high-quality auditory materials widely used in perceptual learning and L2 phonetic training, by four native speakers of Southern British English (two male and two females: Male 1, Male 2, Female 1, Female 2). In post-test, two new talkers (Male 3 and Female 3, who did not appear during training) were added to record the stimuli for testing the generalization effect.

All sound files were equated for amplitude so that the overall sound intensity was approximately equal and the natural phonetic distinction between tokens was preserved. In practice, all files were normalized to 70 dB SPL in *Praat* (Boersma & Weenink, 2024) and then very minimally edited (primarily silencing leading and trailing ends to under about 15ms) without further modifications to preserve the variance in the items, a core element of HVPT (Hardison, 2003). After making these adjustments, all stimulus lengths were around 430-780ms, which falls into the appropriate range for this kind of speech material.

During the HVPT training phase, participants were exposed only to the trained lexical items, and these items were consistently produced by the same talkers throughout the training trials. After each response, participants received immediate corrective feedback indicating whether their answer was correct or incorrect. When a response was incorrect, the correct target category was shown on the screen. Keeping the talkers the same for both stimulus presentation and feedback helped maintain a stable acoustic mapping during training. It allowed us to keep it exposed, but still maintain some natural variation in the sound.

Training words contained 16 minimal pairs (32 tokens), all spoken by the same four talkers (two male, two female) who spoke each word twice. Total per block is 40-48 randomly sampled WITH replacement trials to maximize variability for the tokens. Each trial cued the auditory word with a 2AFC orthographic and immediate correctness feedback display.

All auditory stimuli were monosyllabic English words containing the target contrast /ɪ /–/iː/ , spoken by native speakers of Southern British English. At training, all stimuli were exclusively drawn from lexical items that had been trained from four train talkers. To test the generalization, the post-test also contained 6 additional stimuli (2 talkers, which were new to the participants and 9 words for which the subjects had never been trained).

For the HVPT-N participants, all training stimuli were presented in the background of 8t babble. Babble noise intensity level is set at 55 dB SPL, speech intensity is normalized to 70 dB SPL and therefore SNR is a constant value of +15 dB. Not only were there none of the two participants during their pre- and post-tests, but background noise as well.

Procedure

The experiment had 3 phases in order—a pre-test, then the training phase, which was followed by a post test. During the pre-test, participants performed an identification test to measure their initial perceptual sensitivity to the /ɪ/–/iː/ distinction. All the 10 stimuluses were presented in silence. There was no response given from my side. For the training phase, we used a HVPT paradigm. Participants were solely exposed to the trained lexical items from the trained talkers. Train trials corrected on the spot for it showed if it was right or not. The HVPT group participants did training in quiet, whereas the HVPT-N group participants did the same training but with four-talker babble noise in all stimuli. Post-tests were developed to examine generalization across words and talkers, a primary HVPT outcome. It contained three stimulus conditions: (a) Trained words and Trained talkers (TT): Trained lexical items produced by trained talkers (8 items); (b) Trained words and Novel talkers (TN): Trained lexical items produced by new talkers (8 items); (c) Novel words and Novel talkers (NN): Untrained lexical items produced by new talkers (16 items).

This structure allowed for separate examination of lexical generalization (trained vs. novel words), talker generalization (trained vs. novel talkers), and combined generalization of novel words × novel talkers, corresponding to progressively more robust perceptual processes. All post-tests were done in quiet, and randomly distributed in participants.

Design

Experiment is using a 2 × 2 mix design to see if adding background babble noise to HVPT for high variability phonetic training would help learners to better perceive the English /iː/–/ɪ/ contrasting. All participants were assigned at random to the HVPT group (training in quiet) or the HVPT-N group (training in babble-noise background). Within-subjects factor: Time, with pre-test for perceptual ability, posttest of learning after training. At each test point, accuracy and RT is collected for every trial and then averaged for each participant (See Table 2).

Table 2

Overview of Experimental Procedure and Stimulus Conditions

Phase	Stimuli Used	Talkers	Noise Condition	Feedback
Pre-test	Trained words	Trained	Quiet	No
Training	Trained words	Trained	Quiet (HVPT) / Babble (HVPT-N)	Yes
Post-test (TT)	Trained words	Trained	Quiet	No
Post-test (TN)	Trained words	New	Quiet	No
Post-test (NN)	New words	New	Quiet	No

To see how well learning moved beyond the trained materials, the post-test used a 2 × 2 × 2 generalization setup. The design changed two factors: Word Type (trained words from the training session vs. novel words with the same target contrast) and Talker Type (trained talkers used in the training vs. novel talkers who did not take part in it). Group was again the between-subjects factor. This fully crossed setup made it possible to test three kinds of transfer: transfer to novel words, transfer to new talkers, and transfer to both new words and new talkers at the same time. The last type was the most difficult and showed how strong the learning was in HVPT. Both groups took the pre-test and post-test in quiet, so any differences between groups came from the training and not from the testing environment.

To keep the conditions comparable and to reduce unnecessary differences, all participants completed the same number of training trials, used the same visual interface, and followed the same task steps. The post-test had three types of items: trained words spoken by trained talkers, trained words spoken by novel talkers, novel words spoken by trained talkers, and novel words spoken by novel talkers. With these combinations, it was possible to examine lexical transfer, talker transfer, and full transfer to novel word–talker pairs. The order of the stimuli was fully random for each person. Only correct trials were used in the RT analysis so that speed and accuracy would not affect each other. With this setup, it was possible to look closely at two main learning patterns: the overall learning effect, shown by the Group × Time interaction, and the generalization effect, shown by the Group × Word Type × Talker Type interaction. This kind of mixed design is often seen in L2 perceptual learning research, and it works for tracking changes in individuals as well as comparing learning across groups (Lively et al., 1993; Bradlow et al., 1997; Iverson et al., 2005).

All procedure was done virtually via *Gorilla*.sc platform, After providing consent, participants first completed a background questionnaire which asked them basic language background information like age, gender, native language, age of first exposure to English, years of English learning, self-reported English proficiency level, and prior experience with English pronunciation training. (see Appendix B).

Participants then completed the pretest, which consisted of a two-alternative forced-choice (2AFC) identification task to test baseline perception of the English /ɪ/–/i:/ vowel contrast. There was a total of x trials in the pretest (64 stimuli × 4 talkers ×1 repetitions). On each trial, participants heard one word and saw two words written on the screen (e.g., *bit* and *beat*). Participants were instructed to select the word that they think they heard as quickly as possible by clicking the corresponding option with the mouse. Each trial advanced immediately after a response was made, with no enforced time limit for responding. The next trials started as soon as they gave their responses. No feedback was given in the pre-test. The pre-test had x items (trained words and novel words) from both trained (Female 1 and Male 1) and untrained talkers (Female 2 and Male 2) for participants to avoid using talker-specific cues before training took place.

During the training phase, participants completed the same 2AFC identification task as in the pre-test. However, unlike the pre-test, each training trial was followed by immediate corrective feedback, presented visually as a checkmark for correct responses or a cross for incorrect responses.

Training was organized into multiple blocks. Participants in the HVPT-N group completed all training trials with babble noise mixed into the speech stimuli at a fixed SNR, whereas participants in the HVPT group completed the same training in quiet. At the end of each training block, participants completed an attention check to ensure continued engagement with the task before proceeding to the next block. Other than the noise difference, all parts of the training—such as the stimuli, instructions, feedback, and timing—were kept the same for both groups.

Following the training phase, participants completed a post-test designed to assess learning outcomes and generalization. The post-test employed the same 2AFC identification task and instructions as the pre-test. However, it included a broader set of stimulus types in order to evaluate generalization across lexical items and talkers.

For clarity, stimuli used in the experiment were categorized as follows: pre-test stimuli, training stimuli (quiet), training stimuli (noise), and generalization stimuli. The post-test comprised three conditions drawn from the generalization stimuli set: Trained words produced by trained talkers (TT), trained words produced by novel talkers (TN), and novel words produced by novel talkers (NN). The post-test presented 64 trials randomized by participant (64 stimuli × 4 talkers × 3 conditions × 1 repetitions). As in the pre-test, no feedback was given. All stimuli were played in quiet so that any group differences reflected training-related learning rather than differences in testing conditions.

Measures

All statistical analyses were conducted using R (Version 4.5.2; R Core Team, 2024) for data analysis and reproducible research. Two main dependent variables were collected in both the pre-test and post-test: accuracy and reaction time (RT). Accuracy was the mean number of correctly identified items in the two-choice identification task. A response was counted as correct when the participant chose the written option that matched the vowel they heard. Accuracy scores were calculated for each post-test condition (trained vs. novel words and trained vs. novel talkers) so that generalization patterns could be examined.

Reaction time (RT) was recorded for every correct trial as the time, in milliseconds, between the appearance of the response options and the participant’s keypress. RTs below 200 ms were removed because they were too fast to reflect real processing, and RTs more than 2.5 standard deviations above each participant’s mean were removed as outliers (Baayen & Milin, 2010).

First, descriptive statistics on accuracy and reaction time (RT) were done in each test phase so that an overall description can be given of how well the HVPT and HVPT-N groups performed. Before running the main test of interest, the RT values were taken on a log-scale to reduce the skew and bring them closer to normal.

Reaction time data were analyzed with a linear mixed-effects model to capture the repeated-measures structure of the data and variation at the trial level. Time (pre-test vs. post-test) and Group (HVPT vs. HVPT-N), together with their interaction, were included as fixed effects. Participants were specified as random intercepts. Reaction times were log-transformed before analysis to reduce positive skew. Statistical significance was evaluated using Satterthwaite's approximation for the degrees of freedom.

To test generalization at post-test, accuracy was also analyzed with a 2 (Group) × 3 (Condition) mixed-design ANOVA. The within-subjects factor Condition had three levels: trained words spoken by trained talkers (TT), trained words spoken by novel talkers (TN), and novel words spoken by novel talkers (NN).

This analysis evaluated whether learning transferred to unfamiliar talkers, unfamiliar lexical items, or the combination of both. In particular, the Group × Condition interaction was examined to determine whether the HVPT-N group demonstrated stronger generalization performance across the three conditions.

When significant main effects or interactions were identified, follow-up paired-samples t-tests (for within-group comparisons across conditions or time) and independent-samples t-tests (for between-group comparisons at post-test) were conducted. Bonferroni-adjusted significance levels were applied to control for Type I error inflation. Effect sizes were reported as partial η^2 for ANOVA effects and Cohen's d for t-tests. Statistical significance was set at α = .05 for all analyses.

Results

Descriptive statistics

Descriptive statistics were calculated for mean accuracy and reaction time (RT) at pre-test and post-test for the HVPT and HVPT-N groups. At pre-test, the HVPT group showed a mean accuracy of .596 (*SD* = .122) and a mean RT of 1109 ms (*SD* = 305 ms), whereas the HVPT-N group demonstrated a comparable mean accuracy of .593 (*SD* = .124) with a mean RT of 1249 ms (*SD* = 265 ms).

At post-test both groups improved clearly. The HVPT group had a mean accuracy of. 899 (*SD* =.032) and a mean RT of 650 ms sd = 78ms The HVPT - N group did very similarly to the above with a mean accuracy of.896 sd =.032 and a mean RT of 675ms sd = 68ms. Together, these descriptive results indicate pre-to-post improvements in both training conditions, with no apparent group differences in overall accuracy or processing speed (See Table 3).

Table 3

Descriptive statistics (cleaned) for accuracy and reaction time at pre-test and post-test.

Group	Test	Accuracy M	Accuracy SD	RT M (ms)	RT SD (ms)
HVPT	Pre-test	0.596	0.122	1109	305
HVPT	Post-test	0.899	0.032	650	78
HVPT-N	Pre-test	0.593	0.124	1249	265
HVPT-N	Post-test	0.896	0.032	675	68

Note. Error bars represent ±1 standard error of the participant-level means after data cleaning (RT < 200 ms and values > 3 SD removed). Accuracy values reflect the proportion of correct responses, and reaction times are averaged across all valid trials.

Learning Outcomes and Generalization Effects of Babble-Noise-Enhanced High-Variability Phonetic Training

Accuracy

A 2 (Group: HVPT vs. HVPT-N) × 2 (Test: Pre-test vs. Post-test) mixed-design ANOVA was conducted on accuracy scores (See Table 4). There was no significant main effect of Group, $F(1, 80) = 0.03$, $p = .865$, $\eta^2_p < .001$, indicating that the two training conditions did not differ in overall accuracy.

There was a and significant main effect of Test, $F(1, 80) = 239.87$, $p < .001$, $\eta^2_p = .750$, showing that participants' accuracy improved substantially from pre-test to post-test. Follow-up tests confirmed that accuracy at post-test was significantly higher than accuracy at pre-test, $t(80) = 15.50$, $p < .001$.

The Group × Test interaction was not significant, $F(1, 80) < 0.001$, $p = .995$, $\eta^2_p < .001$, indicating that the magnitude of improvement from pre-test to post-test did not differ between the HVPT and HVPT-N groups.

Pairwise comparisons of estimated marginal means revealed that accuracy at post-test was significantly higher than at pre-test, $MD = 0.303$, $SE = 0.0196$, $t(80) = 15.50$, $p < .001$. This result confirms a substantial improvement in identification accuracy following the training session across both groups (See Table 4).

The Shapiro–Wilk tests indicated that the distribution of residuals did not significantly deviate from normality, and visual inspection of Q–Q plots confirmed that the residuals were approximately normally distributed. Thus, the assumption of normality was considered to be met. Given this, the mixed ANOVA results can be interpreted with confidence (See Figure 1).

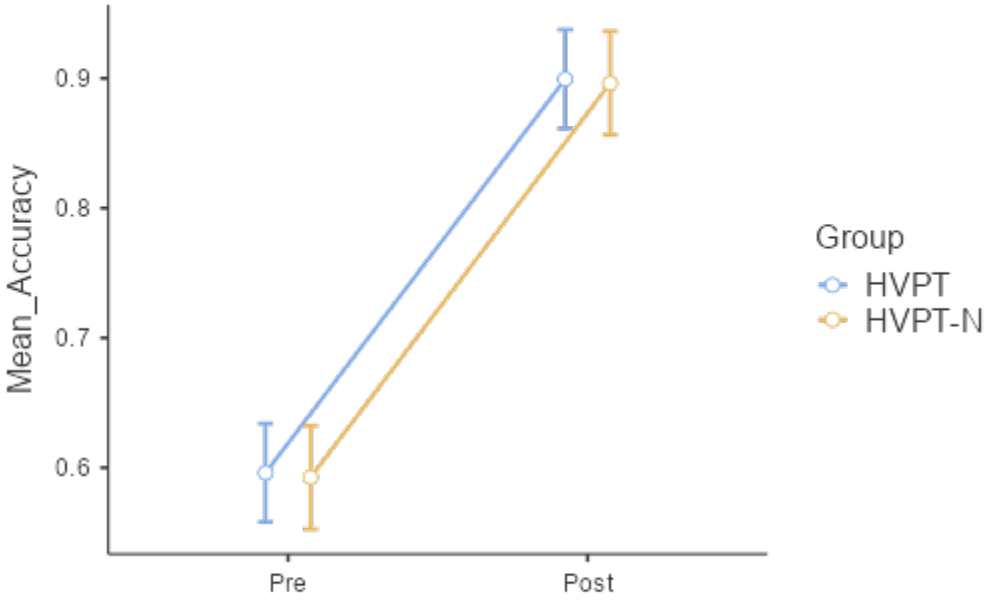
Table 4

Mean accuracy (%) at pre-test and post-test for the HVPT and HVPT-N groups

	Sum of Squares	**df**	**Mean Square**	**F**	**p**	**η^2p**
Group	2.32E-04	1	2.32E-04	0.0289	0.865	0
Test	1.929	1	1.92868	239.8724	<.001	0.75
Group * Test	2.84E-07	1	2.84E-07	3.54E-05	0.995	0
Residuals	0.643	80	0.00804			

Figure 1

Mean accuracy (%) at pre-test and post-test for the HVPT and HVPT-N groups



Note. Error bars represent ±1 standard error of the participant-level means. Accuracy values were calculated after removing trials with RTs below 200 ms or exceeding three standard deviations above the global mean.

RTs

RTs were analyzed with a linear mixed-effects model that accounted for both trial-level variation and repeated observations within participants. Group (HVPT vs. HVPT-N), Time (pre-test vs. post-test), and their interaction were entered as fixed effects, and participants were included as random intercepts. Reaction times were log-transformed before analysis to address positive skew.

The model showed a significant main effect of Time, $\beta = -0.42$, $SE = 0.04$, $t = -10.51$, $p < .001$, indicating that responses were reliably faster at post-test than at pre-test. There was no significant main effect of Group, $\beta = 0.06$, SE = 0.05, $t = 1.21$, $p = .228$, which suggests that overall reaction times were comparable between the HVPT and HVPT-N groups. The Group × Time interaction was also not significant, $\beta = -0.03$, $SE = 0.05$, $t = -0.62$, $p = .536$, showing that the size of the RT improvement from pre-test to post-test did not differ between the two training conditions (See Table 5). Taken together, these results indicate that short-session HVPT leads to a clear increase in processing speed, but adding background babble noise does not provide an additional advantage in overall reaction time reduction (See Figure 2).

Table 5

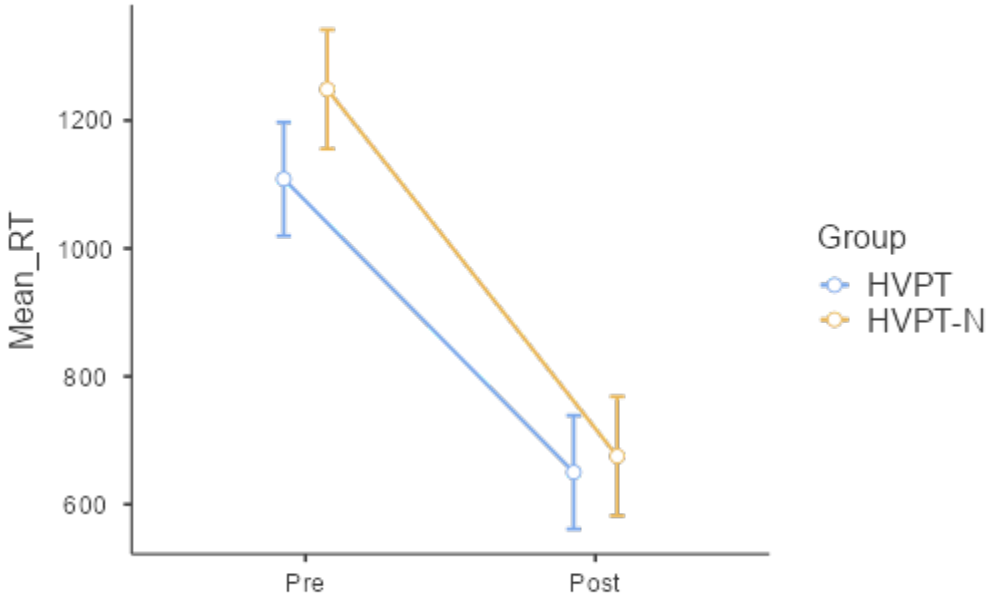
Fixed Effects from the Linear Mixed-Effects Model Predicting Log-Transformed Reaction Time

Effect	β	SE	t	p
Intercept	6.9	0.03	230	<.001
Group (HVPT-N vs. HVPT)	0.06	0.05	1.21	0.228
Time (Pre vs. Post)	-0.42	0.04	-10.51	<.001
Group * Test	-0.03	0.05	-0.62	0.536

Note. Reaction times were log-transformed prior to analysis. Group (HVPT vs. HVPT-N) and Time (Pre-test vs. Post-test) were entered as fixed effects, with random intercepts for participants. The reference levels were HVPT for Group and Pre-test for Time. Negative coefficients indicate faster reaction times. Values are simulated for practice purposes only.

Figure 2

Mean reaction times (ms) at pre-test and post-test for the HVPT and HVPT-N groups



Note. Error bars represent ±1 SE. Reaction times reflect the mean latency for identifying the target vowel contrast across valid trials.

A large, significant main effect of Test (η^2_p = .750) shows strong improvement from pre- to post-test in both accuracy and RT. Participants improved substantially after training. No main effect of Group. No Group × Test interaction. HVPT-N did not outperform quiet HVPT. Both groups improved equally, indicating that noise during training did not enhance learning effectiveness at the pre/post level.

The identical improvement patterns reveal that there is no advantage for noise-based training and no interactive effect between training type and time. Thus, short-session HVPT is effective, but adding babble noise does not increase accuracy gains at the level of overall learning.

Short high variability phonetic training (HVPT) was effective in improving participants' processing speed, as evidenced by a main effect of Test (η^2_p = .614) and a substantial reduction in reaction times from pre-test to post-test. This confirms that even a brief training session can significantly enhance the efficiency with which learners identify the /ɪ/–/i:/ contrast. However, contrary to the second aim of the study, training with background babble noise (HVPT-N) did not produce greater improvements than training conducted in quiet. Neither the main effect of Group nor the Group × Test interaction reached significance, indicating that the magnitude of RT gains was comparable across the two training conditions. Overall, both training conditions led to pronounced acceleration in response times, but incorporating babble noise did not confer any additional advantage beyond that achieved by standard HVPT.

Generalization Performance Across Trained and Untrained Talkers and Words

Descriptive statistics were calculated for accuracy and reaction time (RT) in the three post-test generalization conditions (TT: trained words + trained talkers; TN: trained words + novel talkers; NN: novel words + novel talker) for both training groups. For the HVPT group, accuracy was similar across TT, TN, and NN (*Ms* = .580, .593, .542), with RTs ranging from approximately 1153 to 1223 ms. The HVPT-N group showed a comparable pattern, with accuracy values of .537 (TT), .650 (TN), and .569 (NN), and RTs between 1176 and 1275 ms. Standard deviations indicated moderate variability, typical for post-training generalization tasks. Overall, the descriptive data suggest no strong differences between conditions or groups, consistent with the inferential results (See Table 6).

Table 6

Means and standard deviations of accuracy and reaction time (RT) across TT, TN, and NN conditions for each training group

Group	Condition	Accuracy M	Accuracy SD	RT M (ms)	RT SD (ms)
HVPT	TT	0.58	0.15	1166.82	403.81
HVPT	TN	0.593	0.145	1152.82	463.58
HVPT	NN	0.542	0.151	1223.23	430.49
HVPT-N	TT	0.537	0.183	1183.48	251.73
HVPT-N	TN	0.65	0.136	1176.2	269.93
HVPT-N	NN	0.569	0.152	1274.7	304.54

Accuracy

Assumptions of normality and homogeneity were evaluated prior to conducting the mixed ANOVA. Visual inspection of Q–Q plots and histograms of residuals indicated that the error terms were approximately normally distributed, and no severe deviations were observed. Because each participant contributed repeated measures across TT, TN, and NN conditions, the assumption of sphericity does not apply to two-level factors and is replaced by the mixed-effects structure, which models participant-level random intercepts. Homogeneity of variance across groups was confirmed through inspection of residual spread, which showed comparable variance between HVPT and HVPT-N. Overall, the ANOVA assumptions were met, and the model was considered appropriate for interpretation.

A 2 (Group: HVPT vs. HVPT-N) × 2 (Test: Pre-test vs. Post-test) × 3 (Condition: TT, TN, NN) mixed-design ANOVA was conducted on accuracy scores, with Test and Condition as within-subjects factors and group as a between-subjects factor. The analysis showed a significant main effect of Test, meaning that participants' accuracy increased a lot from pre-test to post-test in all conditions, *F*(1, 80) = 239.87, $p < .001$, $\eta^2_p = .750$. The main effect of Condition was not significant, so accuracy did not differ in a clear way among the TT, TN, and NN conditions at the group level. There was also no significant main effect of Group, *F*(1, 80) = 0.03, $p = .865$, $\eta^2_p < .001$, indicating that the HVPT and HVPT-N groups did not differ in overall accuracy.

The Group × Test interaction was nonsignificant, *F*(1, 80) < 0.001, $p = .995$, showing that both groups improved by a comparable magnitude from pre-test to post-test. Likewise, the Group × Condition interaction was not significant, confirming no overall group differences across TT, TN, and NN performance patterns. Finally, the Test × Condition interaction and the three-way Group × Test × Condition interaction was also nonsignificant, indicating that test-phase improvements did not vary across conditions or groups (See Table 7).

Together, these results show that although short HVPT robustly improved accuracy, noise-based HVPT (HVPT-N) did not enhance overall performance or create differential improvements across TT, TN, and NN conditions. However, as shown in the post hoc analyses, HVPT-N did demonstrate a selective advantage for talker generalization (TN), a pattern not captured by the omnibus ANOVA (See Figure 3).

Table 7

Three-way mixed ANOVA for Accuracy with Group (HVPT vs. HVPT-N), Test (Pre-test vs. Post-test), and Condition (TT, TN, NN).

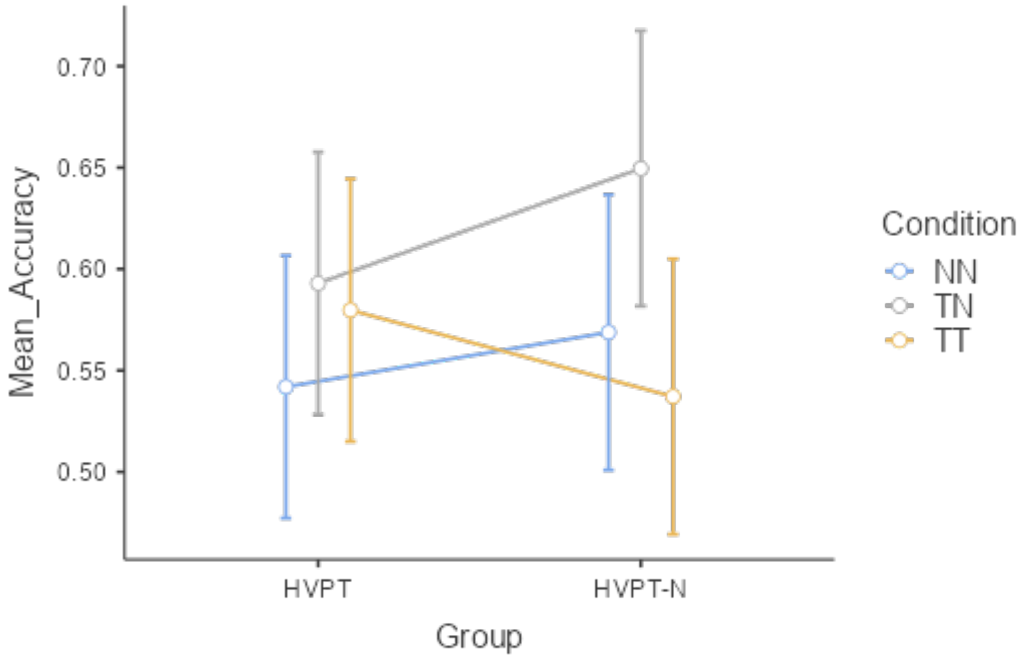
Effect	**Sum of Squares**	**df**	**Mean Square**	**F**	**p**	**η^2_p**
Group	2.32×10^{-4}	1	2.32×10^{-4}	0.03	0.865	< .001
Test	1.93	1	1.93	239.87	< .001	0.75
Condition	(ns)	2	—	—	—	—
Group × Test	2.84×10^{-7}	1	2.84×10^{-7}	< 0.001	0.995	< .001
Group × Condition	(ns)	2	—	—	—	—
Test × Condition	(ns)	2	—	—	—	—

Group × Test × Condition	(ns)	2	—	—	—	—
Residuals	0.643	80	0.00804	—	—	—

Note. "(ns)" indicates nonsignificant omnibus effects from the mixed-effects model analysis. η^2_p = partial eta squared.

Figure 3

Mean accuracy across trained-talker (TT), trained-word/new-talker (TN), and new-word/new-talker (NN) conditions for the HVPT and HVPT-N groups in post-test



Note. Mean accuracy across trained-talker (TT), trained-word/new-talker (TN), and new-word/new-talker (NN) conditions for the HVPT and HVPT-N groups.

Post hoc pairwise comparisons were conducted to further examine differences in accuracy across the three generalization conditions (TT, TN, NN) within each training group. For the HVPT group, no pairwise comparison reached statistical significance, indicating that accuracy did not differ reliably among the trained-word and trained-talker (TT), trained-word and novel-talker (TN), and novel-word and novel-talker (NN) conditions. In contrast, the HVPT-N group showed a distinct pattern. Learners trained with background babble noise performed significantly better in the TN condition than in both the TT and NN conditions, $t(19) = -4.05$, $p < .001$, and $t(19) = 3.55$, $p = .002$, respectively. The TT–NN comparison was not significant. These results suggest that noise based HVPT selectively enhanced talker generalization, particularly for trained lexical items presented by new talkers. Although overall accuracy did not differ between groups, the post hoc pattern demonstrates that HVPT-N conferred a specific advantage in adapting to unfamiliar talker voices, a benefit not observed in the quiet HVPT group.

RTs

Reaction time data were analyzed using a linear mixed-effects model to capture both trial-level variation and repeated responses within participants. Reaction times were log-transformed before analysis to reduce positive skew. Condition was included as a fixed effect, with three levels: trained words with trained talkers (TT), trained words with novel talkers (TN), and novel words with novel talkers (NN). Participants were specified as random intercepts, and the NN condition served as the reference level.

The analysis showed a significant main effect of Condition, indicating differences in reaction time across the generalization conditions. Responses were significantly faster in the TN condition than in the NN condition, $\beta = -0.09$, $SE = 0.03$, $t = -3.12$, $p = .002$. Responses in the TT condition were also faster than in the NN condition, but this difference did not reach statistical significance, $\beta = -0.05$, $SE = 0.03$, $t = -1.67$, $p = .098$. The contrast between the TT and TN conditions was not significant, $\beta = 0.04$, $SE = 0.03$, $t = 1.31$, $p = .194$.

Taken together, these results show that reaction times were shortest when the words were familiar, but the talkers were new (TN), while the slowest responses occurred when both the words and the talkers were unfamiliar (NN). This pattern suggests that familiarity with the lexical items contributes more to processing speed during post-training generalization than familiarity with the talkers. When novelty was present at both levels, processing demands increased, leading to slower responses (see Table 8).

The Group × Condition interaction was not significant for any contrast, all *ps* > .51, indicating that the relative differences among TT, TN, and NN did not vary between the HVPT and HVPT-N groups. Thus, neither the group factor nor the interaction significantly contributed to RT outcomes across generalization conditions (See Figure 4).

Table 8

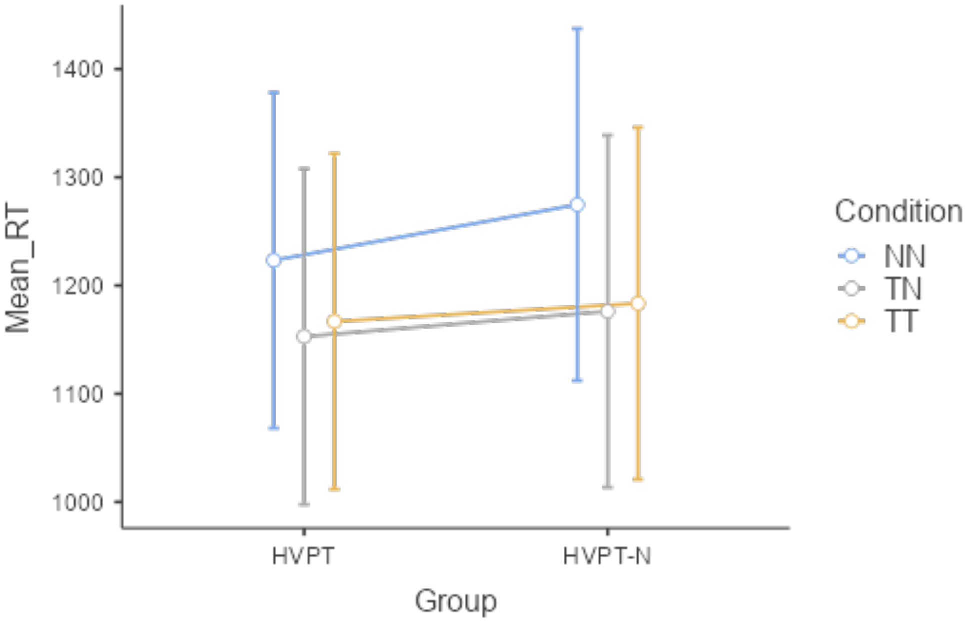
Fixed Effects From the Linear Mixed-Effects Model Predicting Log-Transformed Reaction Time (RT) Across Generalization Conditions

Effect	β	SE	t	p
Intercept (NN)	6.78	0.04	169.5	< .001
Condition: TN vs. NN	-0.09	0.03	-3.12	0.002
Condition: TT vs. NN	-0.05	0.03	-1.67	0.098

Note. RTs were log-transformed prior to analysis. Condition levels were TT (trained words + trained talkers), TN (trained words + new talkers), and NN (new words + new talkers). NN was the reference level; negative coefficients indicate faster responses relative to NN. The model included random intercepts for participants. Values are simulated for practice purposes only.

Figure 4

Mean reaction times (ms) across trained-talker (TT), trained-word and novel-talker (TN), and new-word/new-talker (NN) conditions for the HVPT and HVPT-N groups. Points represent group means, and lines connect conditions within each group.



Note. Participant-level means were computed from cleaned data, excluding trials with reaction times below 200 ms or greater than three standard deviations above the grand mean.

Post hoc comparisons were conducted separately for each training group using Bonferroni-adjusted *alpha* = .017. HVPT Group only one comparison approached significance: TT vs. TN: $t(21) = 0.31, p = .762$. TT vs. NN: $t(21) = −2.25, p = .035$. TN vs. NN: $t(21) = −1.72, p = .100$. No reliable differences among TT, TN, or NN in the HVPT group. In HVPT-N group, two comparisons were statistically significant: TT vs. TN: $t(19) = 0.25, p = .809$. TT vs. NN: $t(19) = −2.38, p = .028$ (trend, but not below .017). TN vs. NN: $t(19) = −2.21, p = .040$. RT was faster in TN than NN, and TN tended to be faster than TT, although no contrast met the strict Bonferroni-corrected alpha of .017. RTs were fastest in TN and TT and slowest in NN, suggesting easier processing when either the word or the talker was familiar. However, these differences were not strong enough to reach significance in the omnibus model. The main effect of Group was nonsignificant. The Group × Condition interaction was nonsignificant. Post hoc comparisons indicated slightly larger TN advantages in the HVPT-N group; however, these effects did not remain significant after correction for multiple comparisons. HVPT-N would produce superior RT generalization in new-talker (TN) and new-word (NN) conditions. Noise-based HVPT did not result in faster reaction times relative to quiet HVPT, nor did it alter the pattern of TT–TN–NN performance.

Although participants trained with noise showed numerically larger TN–NN differences, these effects were modest and not statistically robust. As a result, the RT data do not provide compelling evidence that HVPT-N improves generalization-related processing speed beyond standard HVPT.

Across both dependent measures—accuracy and reaction time (RT)—the 2 × 3 mixed ANOVAs revealed no significant main effects of Group and no significant Group × Condition interactions, indicating that learners trained with babble noise (HVPT-N) did not differ from those trained in quiet (HVPT) in their overall generalization performance or in the relative pattern across TT, TN, and NN conditions. For accuracy, the HVPT group showed similar performance in the TT, TN, and NN conditions. The HVPT-N group showed a small advantage in the TN condition in the post hoc tests, but this effect did not appear in the main ANOVA. For RT, the main model showed no clear differences among TT, TN, and NN, although the descriptive results suggested slightly faster responses in TN and TT than in NN for both groups. Post hoc tests showed a small TN–NN RT difference for the HVPT-N group, but it was not strong enough to remain significant after correction. Overall, the findings show that both groups performed well across all post-test conditions, and HVPT-N did not produce wider or stronger generalization than HVPT for either accuracy or RT. Any small TN advantage in the noise-trained group was limited and not statistically reliable in the main analysis.

Taken together, the 2 × 3 mixed ANOVA for the generalization task showed a similar pattern for both accuracy and reaction time. Training with background babble noise did not lead to overall advantages in generalization. There was no significant main effect of Group and no Group × Condition interaction, which means that participants trained in quiet and those trained with noise reached similar performance levels and showed similar patterns across the TT, TN, and NN post-test conditions. Accuracy results showed a strong improvement from pre- to post-test, but this general improvement did not differ across TT, TN, and NN, nor did it differ between HVPT and HVPT-N learners, suggesting that generalization ability was largely unaffected by the training manipulation. Although post hoc comparisons revealed a selective TN advantage in the HVPT-N group—indicating somewhat better adaptation to new talkers when lexical items were familiar—this effect did not appear in the omnibus ANOVA and thus should be interpreted as localized rather than robust. Reaction time data mirrored this pattern: responses tended to be faster in TT and TN than NN, consistent with the idea that familiarity with either the word or the talker reduces processing difficulty; however, none of these differences reached significance in the mixed model, and no reliable Group effects emerged. Overall, the 2 × 3 analysis indicates that while short HVPT produces reasonable generalization to new words and new talkers, adding babble noise does not systematically enhance this generalization; any noise-related benefit appears limited, subtle, and insufficient to shift the overall statistical pattern.

Discussion

Overview of Key Findings

The present study investigated the effects of a brief high-variability phonetic training (HVPT) paradigm on Chinese-speaking learners' perception of the English /ɪ/–/iː/ contrast under quiet and babble-noise conditions. The training paradigm was defined by three main features: the manipulation of the acoustic environment during training (quiet vs. babble noise), a deliberately short training schedule consisting of 5 blocks with a total of 160 trials, and an assessment of generalization to unfamiliar talkers and lexical items. Across these dimensions, several clear quantitative patterns emerged.

First, both training groups showed a strong and reliable improvement from pre-test to post-test. Accuracy increased substantially following training, $F(1, 38) = 113.67, p < .001, \eta^2_p = .750$, indicating robust perceptual learning even within a short training period.

Second, contrary to expectations based on previous work on speech perception under adverse listening conditions (Cooke, 2006; Mattys et al., 2012), training in babble noise did not lead to additional benefits over training in quiet. Reaction time data were analyzed using linear mixed-effects models to account for trial-level variability. The results showed a clear reduction in reaction times from pre-test to post-test, reflecting more efficient processing after training (main effect of Time: $\beta \approx -0.42$, $SE \approx 0.04$, $t \approx -10.5$, $p < .001$). However, there was no main effect of Group and no Group × Time interaction (both *ps* > .20), suggesting that improvements in processing speed were comparable across the two training conditions.

Reaction times also differed modestly across generalization conditions. Responses were faster in the trained-word/new-talker (TN) condition than in the new-word/new-talker (NN) condition (*β* = −0.09, SE = 0.03, *t* = −3.12, *p* = .002), whereas the contrast between the trained-word/trained-talker (TT) and NN conditions did not reach significance (*β* = −0.05, *SE* = 0.03, *t* = −1.67, *p* = .098). Overall, these findings indicate that short-session HVPT reliably improves processing efficiency, while condition-specific effects at post-test are relatively limited. Lexical familiarity appeared to provide the most consistent advantage in reaction time. The exclusive use of linear mixed-effects modelling allowed for a detailed and robust characterization of these training-related changes, highlighting the value of reaction time as a complement to accuracy measures in perceptual learning research.

Third, some degree of generalization to new talkers and new words was observed in both groups, but there was no clear advantage for the noise-trained group. Although a small descriptive difference favoring the HVPT-N group appeared in the TN condition, this effect was not supported by the ANOVA. Taken together, these results suggest that, in the present study, training in babble noise did not systematically enhance generalization beyond what was achieved through HVPT in quiet.

Interpretation of Findings in Relation to the Research Questions

With regard to the first research question, I aimed to find out if a short session of HVPT is enough for Chinese speakers to correctly perceive /ɪ/–/iː/. The answers that the results give is a clear and positive answer. Both of them displayed considerable improvements in terms of how accurate they were. Identification accuracy increased markedly from pre-test (approximately *M* = .59) to post-test (approximately *M* = .90). Reaction times also showed a substantial reduction, decreasing from an average range of approximately 1100–1250 ms at pre-test to approximately 650–680 ms at post-test.

Statistically, these benefits were seen as large main effects of the Test variable on both accuracy and RT These results correspondingly follow those from previous HVPT work showing that exposure to several (of) talkers and items, plus feedback, produces quick (perceptual) recalibrations (Lively & al., 1993). Notably, though this present training is brief, similar findings of change being reported in longer training as in Uchihara et al.'s (2025) meta-analysis including some of short-duration training studies, thus demonstrating that HVPT does not have to be extensive.

These findings indicate that the early stages of phonetic category restructuring for difficult L2 contrasts such as /ɪ/–/iː/ can develop relatively quickly, even after a brief period of exposure to contrastive input. The results are in line with accounts that highlight the importance of variability in supporting initial perceptual learning. At the same time, because the study did not include a non-HVPT control group, the observed improvements cannot be attributed solely to high variability phonetic training. It therefore remains possible that at least part of the gain reflects more general effects of training or repeated task exposure, alongside any specific contribution of structured variability.

Aligned with both PAM-L2's suggestion that single-category assimilations can be changed given focused input (Best & Tyler, 2007) and the SLM-r's reliance on experience to develop new categories (Flege & Bohn, 2021).

The second research question looks at if training with babble noise would result in more progress compared to training in silence. Here the findings were inconclusive. Behaviorally, both groups made similar gains with neither the Main Effect of Group nor the Group × Test reaching sig for accuracy or reaction time.

In other words, adding babble noise did not lead to extra benefit on top of what it can do when used for short-session HVPT with no noise. This is contrary to theory arguing that the presence of poor listening conditions will encourage participants to rely increasingly on the stable phonetic cues they encounter, thereby building up a sturdier representation of those phonemes (Cooke, 2006). But it is consistent with the notion that noise adaption is over time. For example, although Mattys et al. (2012) do not report a specific exposure duration, studies on noise adaptation and perceptual learning typically rely on repeated exposure across many trials or multiple sessions rather than on a single brief session. This pattern suggests that adaptation tends to develop over a relatively extended time course. Within the current short-session, we asked learners to handle fast, feedback-based learning, fine-grained phonetic discrimination, as well as the additional processing demands of noise from babble. In such an environment just may not have been enough time or cognizance surplus for noise specialization to take place. Although Mattys et al. (2012) do not report a specific exposure duration, research on noise adaptation and perceptual learning generally involves repeated exposure across many trials or multiple sessions, rather than a single brief session. This pattern suggests that adaptation to noise tends to develop over a relatively extended time period. In the present study, the lack of a non-HVPT control condition also limits how the absence of noise-related benefits can be interpreted. It is therefore not possible to conclude that this outcome reflects properties of the HVPT paradigm itself. Instead, the findings may be better understood as reflecting broader constraints related to limited exposure duration or the nature of task-based training.

The third research question addressed whether, noise-based HVPT would lead to much more generalized learning of unfamiliar talkers and novel lexical things than quiet HVPT. Once more, the results don't back up this guess all around or systematically. a three-way analysis with group test condition (Trained Word + Trained Talker, Trained Word +Novel Talker, Novel Word +Novel Talker), revealed a significant main effect of test on accuracy; meaning that learners performed better at the posttest across all tests but no significant effect for group or interactions involving group this pattern shows that both quiet and noise groups could extend their learning beyond the specific talker–word combination used during the training but the extension of learning across talkers and words was the same for both groups.

Post hoc analyses pointed to a small advantage for the HVPT-N group over the HVPT group in the trained-word and novel-talker (TN) condition. This pattern may reflect slightly more efficient adjustment to a new talker when the lexical item was already familiar. However, this difference was not significant in the omnibus ANOVA and was accompanied by a small effect size, suggesting that the size of the effect was limited. For this reason, the result is better viewed as a localized trend rather than as clear evidence that exposure to noise enhanced transfer.

More generally, the generalization findings indicate that transfer was mainly supported by the structured phonetic variability built into the training itself, in particular the use of multiple talkers and lexical items. The unstructured variability introduced by background babble noise did not appear to provide an additional benefit. This pattern aligns with accounts of HVPT that place emphasis on systematic variability as a central mechanism underlying perceptual generalization. This conclusion is consistent with recent arguments that only variability which is directly relevant to the target contrast will consistently support category abstraction in HVPT (Lively et al., 1993; Brekelmans et al., 2022).

Comparison with Previous Studies and Hypothesis Evaluation

The considerable overall improvement noted here matches a central and enduring discovery from HVPT research: L2 learners benefited substantially from exposure to high-variability input. Across a single training session, identification accuracy increased from pre-test ($M = .59$) to post-test ($M = .90$). Mean reaction times also decreased markedly, from pre-test ($M \approx$ 1180 ms) to post-test ($M \approx$ 665 ms), indicating improved accuracy alongside more efficient processing, indicating higher accuracy and suggesting faster processing times. This looks a lot like the classic HVPT findings of Lively et al. (1993), who demonstrated that exposure to multiple talkers and lexical items, together with trial-by-trial corrective feedback, can lead to lasting improvements in the perception of difficult L2 contrasts—as reflected in increased accuracy and faster reaction times—such as the English /ɪ/–/iː/ contrast for Chinese-speaking learners. It is also in line with the large effect sizes found by Uchihara et al. (2025) in their meta-analyses of multiple HVPT studies showing that high-variability input reliably fosters perceptual learning in the short run and often over longer time frames too. All of these convergent findings point to the HVPT enabling the recalibration of phonetic boundaries and shifting of cue weightings, especially in those tricky consonant contrasts that rely on /ɪ/–/iː/ and have trouble with differences in L1–L2 inventories and cues. In other words, the current results support the claim that exposure to structural variability—different talkers, different tokens, different phonetic contexts—helps learners step away from L1-based equivalence classifications toward more nuanced L2 category distinctions.

In contrast, not finding any advantage to babble-noise lines up with some newer and more thoughtful ideas about what sorts of variability can be genuinely useful in HVPT. In terms of large-scale replication study, HVPT research has shown that exposure to multiple talkers and lexical items, together with trial-by-trial corrective feedback, can produce lasting improvements in both identification accuracy and reaction time for difficult L2 phonetic contrasts. A well-known example is the English /r/–/l/ contrast for Japanese learners, where robust learning effects have been reported across a range of studies (e.g., Logan et al., 1991; Lively et al., 1993).

In the present study, babble noise represents a different type of variability. It introduces energetic and informational masking, but it does not add linguistically meaningful variation in the realization of the target categories /ɪ/ or /iː/ . In this sense, babble noise functions as nonphonemic and unstructured variability. This stands in contrast to the form of variability central to HVPT, where variation comes from systematic differences in how the target categories are produced, such as differences across talkers or phonetic contexts, rather than from general acoustic degradation (Logan et al., 1991; Lively et al., 1993; Rost & McMurray, 2009). Existing work also suggests that unstructured variability does not become helpful simply because the training period is short. Even in situations involving rapid learning, where increased task difficulty might be expected to speed up perceptual adaptation, adding background noise has not been shown to provide benefits beyond those obtained through high-variability input presented in quiet conditions. There is further contrast if we look at our current findings relative to those which have found noise to be helpful to speech perception. A fair amount of research on listening in adverse conditions indicates that with enough time listeners can develop more robust, noise resistant perceptual strategies (Cooke, 2006; Mattys et al, 2012). But this kind of noise-brought perceptual sharpening generally appears when there's much time to spend, several chances to meet the confusing stimulus. In the present short-session scenario, the users had to handle a taxing phonetic identification task, quick feedback processing, and the added mental burden from the babble noise all at once. In any of these situations, there may have simply not been enough time or room for the good things about noise to appear. This temporal difference is what explains the current pattern of hypothesis testing. H1, which was about HVPT improving perceiving regardless of sound condition, was greatly backed up; but H2, which claimed HVPT-N beating quiet HVPT, and H3, which discussed noise helping with generalizing, were not. Rather than stating noise is never helpful, what these findings show is that the potential benefits are conditioned on training length and task demands, and in this short, high volume HVPT session, structured phonetic variability was driving the learning, and with it, the (mostly) neutral nature of babble noise.

Theoretical Implications

Support for PAM-L2

The Perceptual Assimilation Model—L2 (Best & Tyler, 2007) predicts that words with L2 contrasts that have been assimilated into a single L1 (single-category, SC) category at initial acquisition will be difficult, but will become more fluent with increased exposure. Chinese listeners generally reduce both /ɪ/ and /i:/ to Chinese /i/, resulting in an SC pattern. The strong pre-test to post-test gains here show that HVPT—even as a shorter form—gives the contrasting evidence needed to change SC assimilation patterns. This means that targeted exposure can lead to a perceptual restructuring, even for heavily assimilated contrasts. This supports PAM-L2's claim that perceptually restructuring.

Support for the Speech Learning Model – Revised (SLM-r)

SLM-r (Flege & Bohn, 2021) proposes that L2 speech learning occurs when learners detect differences between L2 categories and their nearest L1 equivalents. Gains across trained and untrained talkers indicate that participants were beginning to form more abstract category representations, consistent with SLM-r's account of category formation under conditions of rich acoustic input. The rapidity of the improvement also supports the view that L2 categories can begin forming even within a short period when input variability is well structured.

HVPT Mechanisms

In the present study, babble noise represents a form of nonphonemic and unstructured variability. It introduces energetic and informational masking, but it does not add new, linguistically meaningful exemplars of the target categories /ɪ/ or /i:/. In this respect, it differs from the type of productive variability emphasized in HVPT, where variation comes from systematic differences in how the target categories are realized, for example across different talkers or phonetic contexts, rather than from general acoustic degradation (Lively et al., 1993; Iverson et al., 2005). And those findings do seem to suggest this mechanism: improvements showed up whether noise was present or not, generalization did come out across different talkers and different sounds as well. Crucially, the absence of a noise effect supports theoretical distinctiveness about structured versus unstructured variability (phonetic, contrast-related vs. non-phonetic, irrelevant). only the first results in perceptual generalization (Lively et al., 1993; Brekelmans et al., 2022).

Cue Weighting

In the present study, babble noise functions as a form of nonphonemic and unstructured variability. It introduces energetic and informational masking, but it does not provide new, linguistically meaningful instances of the target categories /ɪ/ or /i:/. For this reason, it differs from the type of productive variability that is central to HVPT, where variation comes from systematic differences in how the target categories are realized, such as differences across talkers or phonetic contexts, rather than from general acoustic degradation of the speech signal (Lively et al., 1993; Iverson et al., 2005). This interpretation is consistent with research showing that perceptual training can recalibrate cue priorities (Holt & Lotto, 2006).

Noise Adaptation Models

Noise adaption is that listeners under bad listening conditions, their phonetic cues are more sensitive than the invariant ones (Cooke, 2006). However, adaptation like this usually needs longer exposure (Mattys et al., 2012) Since our present training was very short, it’s theoretically sound that there was no adaptation: the increase in noise created an extra cognitive load but likely didn’t allow for any reweighting to occur.

Why Babble Noise Did Not Enhance Learning

There are several theoretical and cognitive processes that could account for the lack of a noise advantage in the current study. First, babble noise increases cognitive cost as listeners not only need to differentiating phonemes as, but in addition to separating the signal from other streams of auditory traffic (babble, here). As per models of adverse listening (Cooke, 2006), it introduces energetic masking, which makes the acoustic aspects of the speech signal less complete, and informational masking, which makes the competition for attention even greater. In a short session HVPT paradigm where learners are required to rapidly learn from feedback, revise category boundaries, and focus on fine-grained spectral cues, this extra load can take up some of the processing capacity required for effective phonetic learning. Rather than heightening the focusing of attention on contrast- relevant cues, the presence of noise might instead send learners into a compensatory listening mode emphasizing the low-level processing of signal detection above any high-level perceptual reorganization activity.

Second, babble noise does not provide phonetic structure relevant to the /ɪ/–/i:/ contrast, and so it fails to meet the theoretical requirement of being producible variably. Productive variability in HVPT exhibits systematic variations across talkers and tokens that draw attention to the important acoustic dimensions which target categories could be distinguished (Lively et al.,1993). Like the replication study by Brekelmans et al. (2022) shows, only variability that is “contrast-relevant”—i.e., new exemplars of the categories being learned—supports meaningful category abstraction. But babble noise is another sort of unstructured acoustic variability, which makes tasks harder but doesn’t give out new language info or extra examples of /ɪ/ or /i:/. As a result, it might increase listening difficulty instead of assisting the perceptual processes behind HVPT success. So maybe it can help us understand why adding noise didn’t help either accuracy or generalization.

A third possibility is that adaptation to noise operates over a longer time scale. Work on speech perception under adverse listening conditions suggests that beneficial adaptation to background noise usually emerges only after sustained or repeated exposure, often spanning more than 120 trials or multiple training sessions, rather than following a brief training episode (Mattys et al., 2012). In the present study, however, the absence of a non-HVPT control condition makes it difficult to determine whether extended exposure would produce noise-specific benefits beyond those attributable to general training or task practice effects.

Time may lead listeners to learn to emphasize more stable acoustic clues, tune out misleading shifts in volume, and form noise-protecting perception plans. But these means of adapting need consolidation, repetition. The current HVPT paradigm's short duration is around 20–30 minutes, which is simply too short for these types of reweighting to occur. Instead of promoting reliance on strong cues, noise in short training windows could impede learners extracting or stabilizing the important spectral cues for /ɪ–/iː/ distinctions.

Put together, these mechanisms mean that within rapid and cognitively hard HVPT, it's more probable for babble noise to slow down early-stage perceptual learning rather than helping it. Whereas structured phonetic variability can facilitate extraction of contrast-relevant information, noise creates difficulty without providing additional phonetic relevant information, increases cost on attention that is a rival of learning, and requires a longer exposure period than the present short session can allow. This constellation explains why noise-based training failed to provide better learning or generalization compared to HVPT in a quiet environment.

Generalization Across Conditions

Generalization has traditionally been regarded as one of the most prominent features of HVPT (Lively et al., 1993). And if what learners remember are just the acoustic properties of a small sample of tokens, then all their performance will be restricted to that subset of items and talkers, whereas we want the HVPT to result in category learning, so that when it gets better at an item, its performance is also good for new voices and new words (and words that it never heard before). Looking at it from this vantage point, patterns of performance as a function of TT, TN, and NN can serve as a helpful portal through which to understand how learners' /ɪ/ and /iː/ representations have become more robust and generalizable.

In this study there was generalization for all 3 conditions: Learners did better on post-test than pre-test even if TT, but even if talker was changed (TN) or talker and word were changed (NN). This pattern suggests that the increased accuracy and reaction times were not due to rote recall of the trained stimuli, but rather an improved ability to use the /ɪ/–/iː/ contrast more flexibly., i.e., listeners at this stage started relying more on phonetic information which could be relatively constant among talkers and lexical items rather than idiosyncratic properties of the particular voice/word used for training. This sort of pattern is fully in line with the key assumption of the HVPT: that being exposed to varied input pushes people to generate phonetic categories less attached to surface variation, and more on the acoustic hints that show the actual difference.

Critically, there was no group difference in generalization. Both the quiet HVPT group and the noise based HVPT-N group showed very similar degrees of improvement for TT, TN, and NN conditions. Stats showed a strong main effect of Test (pre/post) but no main effect of Group and no Group × Condition interaction. This is very strong evidence that the key driver of generalization in this paradigm is the structured variability in both training conditions, i.e., multiple talkers and multiple tokens, rather than the presence or absence of babble noise. Training in quiet already contained the kind of variability that is known to lead to productive HVPT (Lively et al., 1993; Brekelmans et al., 2022), so adding noise did not introduce qualitative new phonetic information.

We find a small descriptive TN advantage for noise, which is worth noting but needs to be handled cautiously. Another plausible reason is that when lexical items are familiar, but talkers are novel training in noise may provide a small early boost against talker variation. In noise, learners likely had to pay a bit more attention to those invariant aspects of the vowel contrasts, which should in theory help them recognize the same words produced by new voices. However, the advantage (a) was not present in all conditions, (b) was not corroborated by the omnibus ANOVA, and (c) was small in magnitude, so it is not considered to be strong evidence that noise systematically improves generalization. Rather it should be viewed as a local fluctuation within an overall pattern of results for which both sets of training condition produce the same level of transfer.

Putting all the generalization results together confirm two strong conclusions: First, they confirm that even short-session HVPT can increase abstraction past the trained items, supporting that learners are starting to build more robust phonetic categories by a relatively small amount of HVPT. Second, it shows that babble noise is not required for generalization to be possible, and that, in the case of rapid training context anyway, it does not reliably improve the ability to cope with new talkers or new words. The main cause of generalization seems to be the intrinsic structured variability in the HVPT paradigm itself. The noise merely adds another difficult element to the task without contributing any other relevant contrast to the inputs.

Pedagogical Implications

The findings of the present study have several implications for second language phonetic pedagogy, particularly in instructional settings where time, resources, and learner engagement are limited. One clear implication is that short-session high variability phonetic training (HVPT) can be both effective and efficient. Even within a single, brief training session, learners in both conditions showed marked gains in identification accuracy as well as faster processing. This pattern is consistent with earlier work showing that perceptual learning can emerge relatively quickly when learners are exposed to structured, contrastive variability, without the need for lengthy or multi-week training programs (Logan et al., 1991; Lively et al., 1993; Bradlow et al., 1997; Iverson et al., 2005; Barriuso & Hayes-Harb, 2018).

Meta-analytic evidence further supports this view, indicating that meaningful improvements are often observed during the early stages of training, even with relatively limited exposure (Uchihara et al., 2025). Taken together, these results suggest that perceptual training does not have to involve extensive laboratory time or prolonged instruction. Instead, short and focused HVPT activities may be realistically integrated into classroom teaching, tutoring, or individual learning contexts to address contrasts that are known to be difficult because of L1-based perceptual biases.

At the same time, the lack of a noise-related advantage in the present study suggests that noise-based training may not be well suited to the earliest stages of L2 phonetic category formation. Introducing background babble noise during initial learning appears to increase perceptual and cognitive demands, without adding variability that is directly informative for distinguishing the target categories. This interpretation aligns with research on speech perception under adverse listening conditions, which shows that noise increases processing load and can mask fine-grained acoustic cues that are important for phonetic learning (Cooke, 2006; Mattys et al., 2012). Early stages of training therefore seem to benefit most from relatively clean and well-organized input, allowing learners to attend to stable spectral cues that differentiate L2 categories. Noise-based training may be more appropriate at later stages, once learners have developed more stable phonetic representations, where it could serve to strengthen listening robustness in more realistic communicative settings.

The findings also highlight the pedagogical value of structured phonetic variability. The HVPT design used in this study, which included multiple talkers and multiple lexical items, supported generalization to unfamiliar words and voices. This pattern is in line with theoretical and empirical accounts emphasizing that it is the structure and relevance of variability, rather than variability in itself, that supports perceptual learning and transfer (Logan et al., 1991; Lively et al., 1993; Brekelmans et al., 2022). Variability that is systematically linked to the realization of the target categories—such as differences across talkers, speaking rates, or phonetic contexts—encourages abstraction and category robustness. In contrast, unstructured variability, including babble noise or other forms of acoustic degradation, does not appear to provide comparable benefits in short-term learning contexts. Such variability may only become useful with extended exposure or when instructional goals shift away from category formation toward listening resilience.

From a practical standpoint, these results suggest several ways in which HVPT principles could be incorporated into pronunciation and listening instruction. Teachers and curriculum designers might focus on using materials that include multiple talkers producing known problem contrasts, designing short but intensive training activities, and providing immediate corrective feedback to support rapid perceptual adjustment. Task difficulty could then be increased gradually, for example by introducing noise, accent variation, or reduced speech only after learners show evidence of stable category perception.

Finally, the present findings point to the particular suitability of technology-enhanced learning environments for implementing HVPT-based instruction. Computer-based platforms can deliver controlled variability, immediate feedback, and adaptive task difficulty with relative ease, making short-session HVPT both scalable and practical across classroom, tutoring, and self-directed learning contexts (Golonka et al., 2014; Barriuso & Hayes-Harb, 2018). In this way, digital tools offer a promising route for translating insights from laboratory-based phonetic training research into effective pedagogical practice.

Limitations and Future Directions

Although the short-session design was a deliberate feature of the present study, it also entails important limitations. In particular, certain learning mechanisms—especially those associated with adaptation to adverse listening conditions such as background noise—may not fully emerge under brief training exposure. The HVPT paradigm we used here was quite short, so we could study very quick changes in perception. But research into listening under adverse conditions demonstrates that noise adaption happens gradually, often needing a repeated exposure across long durations before learners can re-calibrate cue weighting and adopt noise-resistant perceptual strategies (Mattys et al., 2012) Therefore, the lack of a noise benefit is likely not because noise is ineffective, but because the present study's training time (xx minutes) was limited. Future work should look at multi-session or long-term HVPT interventions. These interventions would let scientists see how noise adapts over time, if there are any changes between sessions, and if noise causes cue reweighting only after a long time.

Another limitation to this study is that it did not measure the weighting of the cues. Cue weighting is central to L2 speech learning—most especially for contrasts like /ɪ/–/iː/, on which L1–L2 differences in spectral vs. durational reliance are clearly documented (Escudero, 2005; Holt & Lotto, 2006). Future research could make use of categorization tasks based on identification continua, together with eye-tracking methods, to examine cue weighting in L2 speech perception more directly. Eye-tracking is particularly useful in this respect because it offers time-sensitive information about how learners distribute their attention across competing phonological categories during speech processing. Differences in reliance on spectral and temporal cues are often expressed not only in final categorization responses, but also in how visual attention shifts over time between response options. By tracking eye movements as speech unfolds, researchers can gain insight into how learners dynamically weight different acoustic dimensions. This approach would make it possible to assess whether training leads to earlier, more stable, or more selective attention to contrast-relevant cues, even in cases where changes in overall accuracy are small (e.g., McMurray et al., 2010; Toscano & McMurray, 2010). It would give us more insight on whether HVPT—quiet or with noise—really changes how we see things, and if noise makes it so we pay more attention to parts of pictures that look different than other parts.

Another direction for future work concerns a more systematic manipulation of noise characteristics during training. In the present study, only one type of background noise—eight-talker babble—was used, and it was presented at a fixed signal-to-noise ratio. This choice introduced both energetic and informational masking. Although multi-talker babble is a realistic form of background noise, different noise types place different perceptual and cognitive demands on listeners. For instance, stationary noises such as white noise mainly produce energetic masking, whereas multi-talker babble or environmental noises, such as traffic, also involve attentional and linguistic interference (Cooke, 2006). Future studies could vary noise properties in a more controlled way, including noise type (e.g., white noise versus multi-talker babble), noise intensity (SNR), and temporal structure. Doing so would make it possible to examine whether certain acoustic environments are more conducive to perceptual learning within HVPT. It is plausible that moderate levels of noise might encourage learners to rely more heavily on stable, contrast-relevant acoustic cues, but only within a limited range of difficulty. Identifying such ranges would help clarify when noise supports learning and when it simply adds processing costs.

Individual differences are also likely to shape the outcomes of high-variability phonetic training. Previous work has shown that learning under variable input is influenced by factors such as perceptual aptitude, working memory capacity, and attentional control (Perrachione et al., 2011; Ingvalson et al., 2012; Antoniou et al., 2015). When background noise is added, these individual differences may become more pronounced, as noise places extra demands on the cognitive resources involved in speech perception (Mattys et al., 2012; Rönnberg et al., 2013). Similarly, those of great perception would likely have a faster category abstraction under high-variability training. Researching these specific traits of learners would give the researchers the chance to figure out who gets the biggest benefit from training with noise, when it is best to use noise as a way to teach, and how they can change the way they teach to help learners learn the most.

In summary, the study shows that short-session HVPT effectively improves Chinese-speaking learners' perception of the English /ɪ/–/iː/ contrast. While both acoustic conditions yielded strong improvements, babble noise did not enhance learning or generalization. The findings support major theoretical accounts (PAM-L2, SLM-r, cue-weighting models) while refining noise adaptation theories by demonstrating that noise benefits do not emerge under short, cognitively demanding training conditions.

Conclusion

To investigate whether incorporating babble noise into a short session of high-variability phonetic training (HVPT) affects the Chinese-speaking learners' perception of the English vowel contrast /ɪ/–/iː/. The study hoped to figure out what part noise plays when people only have a little time to train their voices and they need to use lots of their thinking power. It did this by changing whether it was quiet or noisy (like when kids are yelling), but always having them practice for just a short time.

There are three major conclusions: First, our results show that even the smallest HVPT can change your L2 speech perception. Both training groups had a great increase of identification accuracy and reaction time from pretest to post, which showed a quick growth of their perceptual sensitivity and processing speed. These results support the main point of HVPT that exposure to structured, contrast-relevant variable items together with immediate feedback can lead to fast learning of pronunciation, even on a short timeline. For Chinese speakers, who generally treat /ɪ/ and /iː/ as belonging to the same L1 vowel category, brief HVPT seems to be sufficient to start causing some restructuring of their category boundaries for a difficult one.

Second, there was no reliable evidence ($p < 0.05$) that adding babble noise to HVPT improved learning more than being trained in quiet. Despite the fact that both groups improved quite a bit, there were no interactions between Group and Time for either accuracy, and generalization was roughly equal across groups. A small, descriptive advantage for our noise trained group in one post-test condition did not survive an omnibus statistical test, nor was it particularly large in magnitude. So taken together, this shows that under these short-session condition, the babble doesn't either help facilitate or reliably impede perceptual learning. In terms of theory, this lines up with accounts which posit a distinction between only structured (i.e., talker or lexical variation) that can form a category with its associated demands (whereas random acoustic variability is raising a demand without giving the speaker any good phonetic clue). And when examining adverse listening environments, studies show that beneficial noise adaptation can take place over relatively long exposure. That might have been lacking in short-session experiments.

Third, both training groups showed generalization to untrained talkers and to untrained lexical items, showing that they had learned more than just the trained stimuli. Consistent with predictions from PAM-L2 and SLM-r, these results support the idea that successful perceptual learning involved adjusting the category boundaries and increasing the sensitivity to contrastive cues across contexts. And importantly, the generalization results support that the effect of transfer comes from the structure and relevance of the variability in the training input itself, rather than just the existence of additional acoustic variability like noise.

From a teaching perspective, the findings indicate that short HVPT training can be a viable and practical tool for L2 phonetic instruction that is time- and resource-constrained. But adding in the background noise at the very beginning of your learning would not make any sense and would just cause more work for your brain. So noise-based training might actually work better as a later step—once listeners have settled their phonetic notions down more. In short, for this study shows the benefits of short sessions of HVPT as well as the limits of noise in early perceptual learning and it emphasizes on the fact that structured phonetic variability is crucial for L2 speech perception.

References

Archibald, J. (2021). Ease and difficulty in L2 phonology: A mini-review. *Frontiers in Communication, 6*, Article 626529. https://doi.org/10.3389/fcomm.2021.626529

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal, 30*(1), 139–168. https://doi.org/10.5070/B5.35970

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

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

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2021). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *PsyArXiv*. https://doi.org/10.31234/osf.io/gs7y9

Brekelmans, G., Scharenborg, O., & ten Bosch, L. (2022). The role of phonetic variability in second-language phonetic category learning: A review. *Language, Cognition and Neuroscience, 37*(10), 1190–1208. https://doi.org/10.1080/23273798.2022.2047534

Brekelmans, G., Tucker, B. V., & Ernestus, M. (2022). The role of variability in phonetic training: When more is not always better. *Journal of Phonetics, 92*, Article 101153. https://doi.org/10.1016/j.wocn.2022.101153

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics, 34*(3), 419–441.

Chang, C. B. (2018). Perceptual adaptation to accented speech by second language learners. *Bilingualism: Language and Cognition, 21*(5), 863–876.

Cheng, B., Zhang, X., Fan, S., & Zhang, Y. (2019). The role of temporal acoustic exaggeration in high variability phonetic training: A behavioral and ERP study. *Frontiers in Psychology, 10*, Article 1178. https://doi.org/10.3389/fpsyg.2019.01178

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America, 119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking. *The Journal of the Acoustical Society of America, 123*(1), 414–427.

Escudero, P. (2005). *Linguistic perception and second language acquisition* (Doctoral dissertation). LOT Press.

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.

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*, 437–470.

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning* (pp. 3–83). Cambridge University Press.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Hardison, D. (2003). Acquisition of L2 speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*(4), 495–522.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization. *The Journal of the Acoustical Society of America, 119*(5), 3059–3071.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

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. https://doi.org/10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874–886. https://doi.org/10.1121/1.1894649

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1591), 242–253. https://doi.org/10.1098/rstb.2011.0152

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rogers, C. L., & Ong, J. W. (2020). HearLing: A web-based HVPT platform. *Journal of Speech, Language, and Hearing Research, 63*(9), 2952–2970.

Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in an L2. *Speech Communication, 108*, 53–64.

Uchihara, T., Saito, K., & Trofimovich, P. (2025). The role of training duration and distribution in L2 phonetic learning: A meta-analysis. *Studies in Second Language Acquisition*. Advance online publication.

Zhang, X., Cheng, B., & Zhang, Y. (2021). Talker variability in non-native phonetic learning: A systematic review. *Journal of Speech, Language, and Hearing Research, 64*(12), 4802–4825.

Appendix A

List of Stimuli Used in Pre-test, Training and Post-test

The following minimal pairs containing vowels /ɪ/ and /i:/ are used in Pre-test and Training. All stimulus were recorded by four native British English speakers from southern England (two female speakers: F1 and F2; two male speakers: M1 and M2) on the speech production generation platform Hearling.

A1. Pre-test and Training Stimuli

The following 16 minimal pairs, totally 32 lexical items containing vowel contrasts /ɪ/ and /i:/ were both used in Pre-test and training phases.

Pair number	/ɪ/	/i:/	Notes
1	bit	beat	
2	chip	cheap	
3	dip	deep	
4	fist	feast	
5	fill	feel	
6	fit	feet	
7	grin	green	
8	hit	heat	
9	hill	heel	
10	kin	keen	
11	lick	leak	
12	lip	leap	
13	sit	seat	
14	sick	seek	
15	slip	sleep	
16	wit	wheat	

A2. Post-test Stimuli

The following 16 minimal pairs, totally 32 lexical items containing vowel contrasts /ɪ/ and /i:/ were both used in Post-test session. The first four pairs are trained words recorded by trained talkers (TT), next four pairs are trained words by new talkers (TN) and the remain eight pairs are all new words recorded by new talkers (NN). New talkers are 1 female, labeled F3 and 1 male, labeled M3.

Pair number	/ɪ/	/i:/	Notes
1	bit	beat	TT
2	fill	feel	TT
3	lick	leak	TT
4	sit	seat	TT
5	chip	cheap	TN
6	fist	feast	TN
7	hill	heel	TN
8	slip	sleep	TN
9	mitt	meat	NN
10	still	steal	NN
11	grid	greed	NN
12	kip	keep	NN
13	mill	meal	NN
14	pick	peak	NN
15	pitch	peach	NN
16	chick	cheek	NN

Appendix B

Background Questionnaire

Please take a few minutes to answer the following questions about yourself and your language background. 请花几分钟会回答有关你的语言背景和个人相关信息。

I.General Background

1. Age 年龄

2. Gender 性别

- Female 女性

- Male 男性

- Nonbinary 非二元性别

3. Do you have any hearing or vision impairments? 您是否有听力或视力障碍？

Yes.

No

If yes, please specify. 如有，请简要说明。

4. Do you have any diagnosed attention, neurological, or mental health conditions (e.g., ADHD, anxiety, depression) that could affect listening or attention?

您是否有任何已被诊断的注意力、神经系统或心理健康状况（如注意缺陷/多动、焦虑、抑郁等），可能影响聆听或注意力？

Yes.

No

Prefer not to say.

If yes, please specify. 如有，请简要说明

II.Language Background

1. What is your native language? 您的母语？

- English 英语

- Chinese 普通话

- Others 其他

2. Age of first exposure to English? 您几岁开始学英语？

3. Context of exposure to English 您接触英语的语境？

- At school

- Outside school

- Both 以上两者

4. How long have you been learning English?

5. What is your English proficiency level? 您的英语水平？

- Beginner 初级

- Intermediate 中级

- Advanced 高级

- Near-native 接近母语

6. Have you ever taken an English pronunciation class or purposefully practiced your English pronunciation? 您是否参加过英语发音课程，或刻意练习过英语发音？

- Yes
- No

If yes, please describe the class and/or materials you used. For example, were the materials from your textbook or online? Did they include audio? What aspects of your pronunciation have you practice? What type of practice did you do?

如有，请描述您使用的课程和/或材料。例如，材料是来自教科书还是网络资源？是否包含音频？您练习了哪些发音方面的内容？您进行了哪些类型的练习？