The effects of training variability and pitch aptitude on the overnight consolidation of lexical tones

Zhen Qin,[a,*] Rui Jin,[a] and Caicai Zhang[b,*]

[a] Division of Humanities, The Hong Kong University of Science and Technology

[b] Research Centre for Language, Cognition and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

* Correspondence to Zhen Qin: hmzqin@ust.hk and Caicai Zhang: caicai.zhang@polyu.edu.hk

Abstract

Purpose: While variability of training materials has the potential to benefit the learning of lexical tones, the benefit is contingent on an individual's pitch aptitude. Previous studies did not segregate immediate learning and consolidation after an overnight interval, and little is known about how pitch aptitude differences affect consolidation. This study examined whether pitch aptitude predicts overnight consolidation of Cantonese level tones through high-variability (HV) and low-variability (LV) training.

Method: Two groups of Mandarin-speaking participants were first assessed in terms of pitch threshold and tone discrimination, which tapped into different aspects of pitch aptitude. They then received Cantonese level-tone identification training in either an HV or LV condition. The participants were trained in the evening, were tested after training, and returned after 24 hours for overnight consolidation assessment.

Results: The results indicate that pitch aptitude, measured through pitch threshold, may have predicted overnight consolidation and training progress of the HV group, but not those of the LV group. In the HV group, compared with high-aptitude learners, low-aptitude learners benefitted temporarily from training variability but did not consolidate the tonal knowledge as well as their high-aptitude counterparts after 24 hours.

Conclusions: The findings suggest that individual learners had difficulty learning non-native tones by virtue of memory consolidation. Higher pitch aptitude ability (pitch threshold) may provide protection against the decay of learned tones and facilitate tone consolidation. The findings imply that the early emergence of tonal representation is a dynamic process among individuals of non-native speakers who are exposed to training variability.

Introduction

Acoustic variability is ubiquitous in the speech signal, and exposure to a certain amount of variability during training has the potential to benefit second-language (L2) learning of speech sounds (Logan et al., 1991; Melnik & Peperkamp, 2020) and its long-term retention (Bradlow et al., 1999; Lively et al., 1994). For instance, lexical tones have multiple sources of acoustic variation such as that driven by talker voice idiosyncrasies (C. Zhang & Chen, 2016). An identical tone category can have different pitch values among tokens produced by different talkers (i.e., inter-talker variability) as well as by the same talker (i.e., intra-talker variability). In the face of such variability, it is thus crucial for listeners to evaluate the pitch they hear in the signal against the pitch range of a specific talker to identify tones accurately (Peng et al., 2011; C. Zhang & Chen, 2016). For instance, a recent study showed that exposure to inter-talker variability (i.e., high variability) during training would facilitate the formation of abstract tonal categories through overnight consolidation (i.e., changes after an overnight sleep) and thus may result in more robust representations of lexical tones (Qin, Gong, et al., 2021). While the effects of inter-talker variability during training (henceforth, training variability) have been tested in perceptual learning (Wayland & Guion, 2004; P. Wong & Lam, 2021) and overnight consolidation (Qin, Gong, et al., 2021; Qin & Zhang, 2019; for long-term retention, see Wang et al., 1999) of lexical tones, the findings are mixed in terms of its beneficial impact.

A factor that might account for the mixed findings is individual aptitude, that is, some listeners are better than others in learning to discriminate and/or identify non-native tones (Bowles et al., 2016; Ingvalson et al., 2013; McHaney et al., 2021). While individual aptitude has been generally found to predict training outcomes, the specific dimensions of

pitch aptitude that contribute to tone learning appear to differ depending on the types of tones being learned. For contour tones (e.g., rising and falling tones in Mandarin), it has been found that the individual ability to distinguish and categorize changes in pitch contour is indispensable for successful perceptual learning (Chandrasekaran et al., 2010; Gandour, 1983; P. C. M. Wong & Perrachione, 2007). With regard to level tones (e.g., higher and lower tones in Cantonese), on the other hand, Qin et al. (2021) showed that level-tone learning is predicted by two measures of pitch aptitude: 1) threshold for detecting pitch height differences, that is, an explicit categorization of pitch height patterns in a pitch threshold task; and 2) pretraining tone discrimination ability, that is, an implicit sensitivity to pitch height differences measured by a tone discrimination task. Since tone training may not benefit all individual learners similarly, it is important to investigate how pitch aptitude (and which dimension of it) predicts learners' abilities in dealing with training variability in their perceptual learning and consolidation of lexical tones.

A large body of research has shown that whether training variability benefits learners or not is contingent on the learners' individual pitch aptitude (Antoniou & Wong, 2015; Perrachione et al., 2011; Sadakata & McQueen, 2014). Previous studies have shown that learners with higher pitch aptitude (i.e., high-aptitude learners) benefited more from high-variability training than those with lower pitch aptitude (i.e., low-aptitude learners) (Perrachione et al., 2011; Sadakata & McQueen, 2014). The interaction between training variability and pitch aptitude has been relatively well documented in learning Mandarin (contour) tone-word associations (for counter-evidence though, see Dong et al., 2019). For instance, Perrachione et al. (2011) investigated whether English-speaking participants' tone learning depended on an interaction between pitch aptitude and training variability.

High-aptitude and low-aptitude learners were grouped based on a Pitch-Contour Perception Test (PCPT) which was used to assess learners' abilities to identify changes in pitch contour (level, rising, and falling) prior to the perceptual training of Mandarin tones. The learners with different pitch aptitudes were trained either in a high-variability (HV; four talkers with their stimuli mixed) training condition or in a low-variability (LV; one talker) training condition. The results (of Experiment 1) showed an interaction effect of pitch aptitude and training variability. The HV training benefited high-aptitude learners who exhibited greater learning than low-aptitude learners whereas no such pattern of aptitude was found for the LV training. Noticeably, the learning performance of low-aptitude individuals was impaired after the HV training compared to those after the LV training. A further experiment manipulated the degree of trial-to-trial variability through the talker-blocked HV condition (i.e., blocking stimuli produced by each talker) and the talker-mixed HV condition (i.e., mixing stimuli produced by four talkers like in Experiment 1). The results (of Experiment 2) showed that the great amount of trial-to-trial variability in the talker-mixed design was detrimental to low-aptitude learners (but not to high-aptitude learners) and may have obfuscated their perceptual learning of the target cues (i.e., pitch contour) of lexical tones. On the other hand, low-aptitude learners improved significantly in the talker-blocked condition whereas high-aptitude learners did not show additional gain from the reduced degree of trial-to-trial variability. The findings suggest that the trial-to-trial variability of the talker-mixed HV training might have caused an "undesirable" difficulty for low-aptitude learners (Fuhrmeister & Myers, 2020; for a meta-analysis of this effect, see X. Zhang et al., 2021). The HV training without blocking stimuli produced by talkers may prevent an individual from engaging in optimal learning strategies (Sadakata

& McQueen, 2014) and was found to be detrimental for lower-aptitude learners (Perrachione et al., 2011).

In addition to the "undesirable" difficulty caused by the degree of trial-to-trial variability in the talker-mixed design, training variability may interfere with memory consolidation and eventually disrupt long-term retention of learned tones, for instance, among learners with lower aptitude (Fuhrmeister & Myers, 2017, 2020; Tucker & Fishbein, 2008). In general, sleep-dependent memory consolidation, right after training, was found to enhance consolidation of learned speech information by facilitating the transfer of episodic information from an acoustic-sensory-based trace to a more robust (i.e., context-independent) representation of speech sounds (Earle & Myers, 2015; Fenn et al., 2013). To fully understand the (speculated) effects of training variability and pitch aptitude on memory consolidation of non-native tones, it would require a study design that separates the early stage of learning, that is, training-induced (post-training) immediate learning, from (post-sleep) memory consolidation. However, most previous tone learning studies included multiple training sessions over days (Perrachione et al., 2011; Sadakata & McQueen, 2014), and few studies have adopted an overnight design (i.e., over two consecutive days) to segregate post-training performance and post-sleep change (for an example, see Qin & Zhang, 2019).

Recent consolidation studies have shown that scheduling an overnight sleep right after evening training often results in better consolidation of the learned information among learners with higher aptitude (Earle & Qi, 2021; Fuhrmeister & Myers, 2020). Specifically, Fuhrmeister & Myers (2020) showed that individual aptitude (i.e., pretraining discrimination ability) positively predicted learners' identification of non-native sounds (a

Hindi dental and retroflex stop contrast) and the relationship became (numerically) stronger after an overnight sleep. That is, English-speaking learners with higher aptitude demonstrated better consolidation of learned sounds than those with lower aptitude. In other words, an alternative possibility to account for the relationship between training variability and pitch aptitude is that training variability affects learners with different pitch aptitude by virtue of how it interferes with overnight consolidation of learned tones. The questions of whether the advantage of high-aptitude learners in tone consolidation holds when training variability changes (for an interference effect of variablity on consolidation, see Fuhrmeister & Myers, 2017), and how tonal representation emerges for individual learners in early stages of tone learning, remain open. Thus, the current study primarily aims to investigate the relationship of pitch aptitude and training variability in the consolidation of non-native tones with an overnight design.

The Present Study

Different from previous studies which investigated contour-tone learning by novice learners who speak non-tonal languages such as English and Dutch (Dong et al., 2019; Perrachione et al., 2011; Sadakata & McQueen, 2014), the present study focused on Mandarin listeners' identification of Cantonese level-level tonal contrasts (Chang et al., 2017; Qin & Jongman, 2016). The Mandarin listeners were found to show reduced sensitivity to Cantonese level tones, compared with non-tonal language listeners such as English listeners (e.g., Qin & Jongman, 2016). It is likely that the fine-grained differences in pitch height among Cantonese level tones were treated as within-category differences by Mandarin listeners. Based on their well-documented difficulty of level-tone learning, the current study has two research aims.

The primary aim was to examine whether pitch aptitude predicts *overnight consolidation* (i.e., after 24 hours) of non-native tones differently through HV and LV training which is conducted in the evening. Based on the finding that individual aptitude positively predicted consolidation after the overnight sleep (Fuhrmeister & Myers, 2020), we predicted that high-aptitude learners were more likely to consolidate newly-learned tones than low-aptitude learners. The effect of pitch aptitude on consolidation is possibly stronger in the HV training, with low-aptitude learners being more vulnerable to the disruptive effect of training variability than high-aptitude learners (Fuhrmeister & Myers, 2017), than that in the LV training. The secondary aim was to examine whether pitch aptitude predicts *tone training* differently over the course of HV and LV training. Based on the findings of previous tone learning studies (Perrachione et al., 2011; Sadakata & McQueen, 2014), we predict that low-aptitude learners may also benefit from HV training through a talker-blocked design (i.e., with the trial-to-trial variability reduced), for example, by showing comparable tone training outcome to high-aptitude learners.

Two groups of Mandarin-speaking participants were first assessed in two tests of pitch aptitude, that is, pitch height threshold and pretraining tone discrimination ability, following our previous study of level-tone learning (Qin, Zhang, et al., 2021). The two tests were included to tap into different aspects of pitch aptitude: explicit categorization of pitch height patterns assessed by a pitch threshold task (C. Zhang et al., 2021); and implicit sensitivity to pitch height differences measured by a tone discrimination task (Earle & Myers, 2014). After that, they received Cantonese-tone identification training in either a HV condition (two talkers: a female talker and a male talker) or a LV condition (a single talker: the same female talker as in the HV condition) manipulated based on the

presence or absence of inter-talker variability. The performance over the course of training (initial vs. final training block) was used to test the secondary prediction of tone training (Perrachione et al., 2011). Crucially, the participants were all trained in the evening hours, were tested after training, and returned after 24 hours[1] for reassessment (see details in the Procedure section below). The overnight change of performance (first vs. second posttest; before vs. after sleep) was used to test the primary prediction of tone consolidation (Fuhrmeister & Myers, 2020).

Method

Participants

A total of 82 participants (52 females, 30 males; age range: 19-33 years, M = 25.07, SD = 2.30) were recruited via campus advertisements at a university in Hong Kong. To be eligible for the study, all participants met the following inclusion criteria: they (1) must be native Mandarin speakers who did not speak any Southern Chinese dialect (e.g., Shanghainese, Hakka, or Southern Min); (2) had minimal exposure to Cantonese, that is, those who have resided in Hong Kong for less than 15 months and received no more than one month of classroom training in Cantonese; and (3) had no more than three years of professional music lessons, including vocal and musical instrument training (Li & DeKeyser, 2017). None of the participants reported a history of hearing impairment or neurological disorders. In addition, they were pre-screened to ensure that all participants had a regular sleep pattern of at least six hours per night over the past month and no sleep disorders (Espie et al., 2014).

[1] The 24-hour interval was chosen to control the effect of circadian rhythms (i.e., time of day, de Bot, 2015).

The study protocol was approved by the Human Subjects Ethics Sub-committee of the university. All participants provided written informed consent and received monetary compensation for their participation. Two participants failed to complete all experiment sessions and were thus excluded. The remaining 80 participants were randomly assigned into the HV (n = 40) or LV (n = 40) training group. Demographic details of each group can be found in Table 1 (see text below for descriptions of the pretests and sleep questionnaire).

Stimuli

The stimuli were 30 words contrasting three Cantonese level tones, T1 (/55/ a high-level tone), T3 (/33/, a mid-level tone), and T6 (/22/, a low-level tone), carried by 10 base syllables (/jan/, /ji/, /jau/, /jiu/, /fan/, /fu/, /ngaa/, /si/, /se/ and /wai/)[2]. All words are meaningful in Cantonese. Stimuli were recorded by three native speakers of Hong Kong Cantonese (one male talker and two female talkers) in a soundproof room on a PC workstation with an Azden ECZ990 microphone (Azden, Mt. Arlington, NJ). The recordings were made at a sampling rate of 44100 Hz with 16 bits per sample. Each monosyllabic word was embedded in a carrier phrase "呢個係_ lei1 go3 hai6 [target word]" (this is [target word]) and repeated three times. The three tokens were then segmented out of the carrier phrase and normalized to 500 ms in duration (a value similar to the duration of natural stimuli) and 70 dB in intensity using Praat (Boersma & Weenink, 2018). Two out of the three tokens were selected for the experiment based on their intelligibility and

[2] Both syllable and talker variations are sources of variability in lexical tone perception. Given a well-documented talker normalization process in lexical tone perception (Peng, 2006; C. Zhang & Chen, 2016), this study opted for manipulating inter-talker variability (i.e., the number of talkers), but used the same set of (different) syllables across training conditions.

pronunciation accuracy. Both tokens were used in the HV training to increase the token variability of the stimuli; in contrast, only one token was used in the LV training.

To introduce inter-talker variability in the HV training, stimuli produced by both the female (F1) and the male (M1) talkers were used. As illustrated in Figure 1, the three Cantonese level tones produced by F1 are overall higher in semitone and have larger semitone ranges than M1. A linear mixed-effects model showed that the semitone differences between T1-3 and T3-6 in the F1 talker were both significantly larger than those in the M1 talker (T1-3: $\beta = 1.90$, SE = 0.10, $t = 19.80$, $p < .001$; T3-6: $\beta = 0.29$, SE = 0.10, $t = 3.01$, $p = .003$). This inter-talker difference appears to have an influence on the participants' performance during training (see Results below). Using the stimuli allowed us to create acoustic variations within the same tone category and across speakers in the HV training. In contrast, only the stimuli produced by the female talker (F1) were used in the LV training (the stimuli produced by an untrained female talker, F2, was only used in post-training tests to assess context-independent representation of tones after consolidation (Earle & Myers, 2015). All stimuli were presented over headphones (Sony MDR-7506) at a comfortable sound volume, adjustable by participants.

Procedure

Figure 2 shows an overview of the pretest-training-posttest procedure. First, a **pretest session** was conducted prior to the experiment day to ensure that the HV and LV training groups were matched in their short-term memory, musical processing, and attention control (see details in the Pretests). At the end of the pretest session, participants were assessed by the first **pitch aptitude measure**, a pitch threshold test (see details in the Pitch Aptitude Tests). Participants also completed the second aptitude measure, an AX

discrimination test, on the day of training. The task partially shared the tonal stimuli presented in the training task. Therefore, to avoid pre-exposure (and potential overnight consolidation) of the stimuli, it was not administered at the pretest session.

Within the same week of pretests, participants were then asked to complete two experimental sessions at approximately the same time between 7 to 9 PM over two consecutive days. On Day 1, **the HV or LV training session** was administered to the two groups of participants, respectively (see details in the Training Session). Following the training, the participants went through **a tone identification test** in the first posttest (Posttest 1; see details in the Consolidation Tests). After the experiment session, participants went home to sleep, potentially allowing an (immediate) overnight consolidation of tone learning. To test whether and how the learned tone categories were consolidated after a 24-hour interval, participants came back for **a tone identification re-test** (Posttest 2). The two identification tests were identical and were completed at the same time of the day to minimize the potential effect of circadian rhythm (Qin & Zhang, 2019). To ensure that the overnight consolidation of tone learning was not affected by any group difference in sleep, participants were asked to complete a sleep questionnaire on Day 2 which recorded their last night's total sleep duration, self-perceived sleep quality (on a rating scale from 1 to 10), and time spent on activities after the day 1 experiment and before sleep (Espie et al., 2014). Both independent t-tests and Bayes factor tests (Hoijtink et al., 2019; Morey et al., 2021) suggested that the two groups had comparable responses in the sleep questionnaire (see Table 1).

Pretests

The purpose of the pretest session is to primarily ensure that the two groups are comparable in musical aptitude (Qin, Zhang, et al., 2021) and learning-related cognitive abilities (Bowles et al., 2016; Ou & Law, 2017). The pretest session included: (1) pitch (short-term) memory span test, adopted from Williamson and Stewart (Williamson & Stewart, 2010); and (2) three subtests of MBEA which measure pitch-based musical abilities (scale, contour, and interval). MEBA has been used as a screening tool for congenital amusia (Peretz et al., 2003), and individuals with higher MBEA scores performed better in Cantonese tone categorization (Qin, Zhang, et al., 2021). (3) Five subtests of the Test of Everyday Attention (TEA; Robertson et al., 1996), which provide a comprehensive assessment on selective attention (Telephone Search), attentional switching (Visual Elevator, Elevator Counting with Distraction, Elevator Counting with Reversal) and divided attention (Telephone Search While Counting). These subtests probed into participants' visual and auditory attention, which were found to be related to speech perception (Ou & Law, 2017). Both independent-samples t-tests and Bayes factor tests showed that there was little evidence supporting a group difference in any of the three pretest measures. Results are summarized in Table 1.

Pitch Aptitude Tests

Pitch Threshold Test. The test was adopted from Qin et al. (2021) as an aptitude measure to comprehensively assess a participant's ability in explicit categorization of pitch height patterns as well as pitch processing over speech and non-speech tones (C. Zhang et al., 2021). It captures individual sensitivity to detect just-noticeable differences in pitch height, which is a crucial dimension for contrasting Cantonese level tones (Qin et

al., 2021). The speech tones were carried by the Cantonese syllable /ji/ and were produced by a male native Cantonese speaker. The non-speech tones were created in Praat using complex tones, which had more than one single frequency component, and carried the same F0 (100 Hz) as the speech tones. A 15-ms amplitude ramp was applied at the onset and offset of the complex tones to adjust for rise or decay time. The duration of all stimuli was normalized to 250 ms and the same duration was used for the interstimulus interval. The stimuli consisted of a standard stimulus (100 Hz) and 82 target stimuli ranging from 100.07 to 178.17 Hz in steps of 0.01, 0.1, and 1 semitone. Therefore, the pitch height difference between the standard and target stimuli ranged from 10 to 0.01 semitone. In each trial, participants heard the stimulus pair (standard and target), and they needed to judge whether the pitch pattern of the pair was high-low or low-high by pressing the left or right arrow button on the keyboard. The task followed a "two-down, one-up" staircase method in the adaptive tracking procedure (Leek, 2001). Trials began with the largest pitch difference (10 semitones) and were reduced by 1 semitone upon two consecutive correct trials. When the difference reached 1 and 0.1 semitone, respectively, the reduction was adjusted to 0.1 and 0.01 semitone. An incorrect trial led to a reversal to the previous semitone difference. The task ended after 14 reversals. The pitch threshold was calculated as the mean pitch difference between the stimulus pair in the last 6 reversals. The speech and non-speech tones were conducted in separate blocks and their order was counterbalanced across participants. The overall pitch threshold of each participant was the average of their performance in speech and non-speech tones. A lower pitch threshold indicates that participants had successfully detected smaller pitch differences and thus had higher aptitude in pitch (height) processing.

Pretraining Discrimination Test. The AX (same-different) discrimination task was used to measure participants' pretraining implicit sensitivity to pitch height differences, following the design of Qin and Zhang (2021) (also see Fuhrmeister & Myers, 2020). Participants heard two tone stimuli on each trial and were asked to indicate whether they belonged to the same or different tone categories by pressing the left or right arrow button on the keyboard. The tone stimuli were always carried by the same syllable in each trial. The task had 240 trials in total, with an equal number of AA (same tone category) and AB (different tone categories) pairs presented in a counterbalanced manner. AA pairs always included two different tokens of the same tone so that the participants could not categorize the sounds based on low-level acoustic details of the stimuli. The inter-stimulus interval was set at 1000 ms, and no feedback was given.

Training Session

The training, a forced-choice identification (ID) task of the three Cantonese tone categories (T1, T3, or T6), was conducted to assess training progress of tone categorization over the course of training (Bowles et al., 2016; Perrachione et al., 2011). At the beginning of the task, participants read a brief explanation about the pitch height differences of the three tone categories, i.e., T1 being the highest, followed by T3-Mid and T6-Low. They were then given practice trials to familiarize themselves with the task procedure. Additionally, on each trial, a ruler was presented beside each tone category to remind participants of their corresponding pitch height. After hearing each stimulus, participants needed to identify which category it belonged to by pressing one of the three number keys (1, 3, and 6) in a self-paced fashion. Feedback ("Correct" in green or "Incorrect. The correct answer is …" in red) was given immediately after each response.

Participants were instructed to learn via feedback and try their best to correctly categorize each stimulus throughout the training. The training consisted of 600 trials in 10 blocks with 60 trials (3 tones x 2 tokens/repetitions x 10 words) within each block.

The current study adopted an HV training condition with inter-talker variability (Qin, Gong, et al., 2021). To achieve the optimal training outcome at a "desirable" level of difficulty (Fuhrmeister & Myers, 2020), a minimum number of talkers was used (i.e., 2 talkers, one female and one male) in the HV training. Moreover, talkers were presented in a blocked design so that each block contained the stimuli produced by a single talker. Participants in the HV training group learned to categorize the stimuli produced by both a female (F1) and a male (M1) talker that alternated between blocks. In contrast, in the LV training group, only the female talker (F1) was used.

Identification Test (Consolidation)

The identification task was conducted to examine an overnight change in categorizing the three Cantonese level tones. Immediately after the training session (Posttest 1) and after a 24-hour interval (Posttest 2), participants of both groups were asked to categorize the tone of each stimulus (T1, T3, or T6) by pressing the three number keys (1, 3, or 6) within a 6s time window. Unlike the training task, no feedback was given after each response. Both groups were tested on the trained female talker (F1). To assess context-independent representation of tones after consolidation (Earle & Myers, 2015), stimuli produced by an untrained female talker (F2) were used in the post-training identification tests. The identification task had 120 trials (3 tones × 2 talkers × 2 tokens × 10 words) randomly presented in one block.

The pitch memory span test was conducted via MATLAB (The Mathworks, Inc., Natick, MA, USA), MBEA and pitch threshold test via E-Prime 2.0 (Psychology Software Tools, Inc.), and all the other tasks via the Paradigm software (Perception Research Systems, Inc. http://www.paradigmexperiments.com/). The pretest session and the aptitude tests lasted for about 1.5-2 hours, while experiments on Day 1 lasted for around 1 hour (i.e., a 40-min training and a 20-min ID task) and on Day 2 for around 20 minutes. Participants therefore spent around 3 hours in total to complete all three experimental sessions.

Statistical Analysis

Data files, along with analysis scripts, are made publicly available at OSF (https://osf.io/2wkda/). To examine the effect of pitch aptitude on tone training and overnight consolidation, mixed-effects logistic regression models were performed on participants' response accuracy (binary, 1 for correct and 0 for incorrect). The models were fitted in R (R Development Core Team, 2008) using the lme4 package (Bates et al., 2015). All models followed the backwards stepping procedure to determine a maximal random effects structure that was best justified by the data (Matuschek et al., 2017). To standardize the data, pitch threshold in semitone was log-transformed and *z*-normalized prior to the analyses. Participants' performances in AX discrimination were converted to *d-prime* scores, that is, d' = *z* (Hit) – *z* (False Alarm) (Macmillan & Creelman, 2004), to account for response bias. To capture participants' training progress in tone categorization throughout the training blocks, the first two blocks were coded as initial blocks and the last two as outcome blocks. For the training session, group (2 levels: HV vs. LV; deviation coding: −0.5, .05) and block (2 levels: initial vs. outcome; deviation coding: −0.5, .05) were entered as fixed effects. To disentangle the overnight performance changes of the two

groups in the identification tests, group (2 levels: HV vs. LV; deviation coding: −0.5, .05) and test (2 levels: ID posttest1 vs. ID posttest2; deviation coding: −0.5, .05) were entered as fixed effects, and test was nested within group so that the effect of test is examined within each group. Since the focus of the current study is not the differences among individual level tones (1, 2 and 3), and our previous study (Qin & Zhang, 2019) revealed no effect of tone on overnight consolidation of the level-tonal contrasts, tone is thus not included as a fixed effect in the analysis. Nevertheless, tone was put in the random structure to account for its potential effect on response accuracy. All categorical predictors were deviation coded (−0.5, 0.5) to test for the main effect (see details above). Crucially, since pitch threshold and d-prime scores were tapping into the different aspects of individual pitch aptitude, both were entered in the aptitude models as fixed effects (i.e., continuous predictors). The aptitude models are to assess the effects of individual differences and how they interact with the two groups in the training and consolidation process.

Results

Training Progress

As a baseline assessment of tone training, the proportion of correct responses (0-1) in tone categorization throughout the training task was first analyzed for the HV training group and the LV training group (see Figure 3). One participant in the HV group who had below chance performance (< 0.33) up to the end of the training was excluded from further analysis.

To test for group differences in training progress during training, a mixed-effects logistic regression model was performed on participants' response accuracy (See S1 in supplementary materials for details of response accuracy in the training). Fixed effects

included group (HV vs. LV) and block (initial vs. outcome), and the random-effects structure included by-participant intercepts and slopes for group and by-tone intercepts. Results of the model revealed a main effect of group ($\beta = 0.35$, SE = 0.16, $z = 2.25$, $p = .024$) and block ($\beta = -0.33$, SE = 0.03, $z = 9.32$, $p < .001$), but no interaction between group and block ($\beta = 0.07$, SE = 0.07, $z = 0.99$, $p = .323$), indicating that although the performance of the LV training group was in general better than the HV training group due to less stimulus variability, the training-induced improvement from initial to outcome blocks was not significantly different between the two groups. That is, the two groups improved to a similar extent over the course of the training task[3]. In addition, if constraining the data analysis to the stimuli of the F1 talker, which both groups of participants were exposed to during training, the response accuracy was not statistically different ($\beta = 0.24$, SE = 0.15, $z = 1.55$, $p = .122$) between the HV and LV training groups.

Within the HV training group, performance on the F1 talker's stimuli was significantly higher than that on the M1 talker's stimuli ($\beta = -0.23$, SE = 0.03, $z = -7.56$, $p < .001$). Accordingly, performance fluctuation in the HV training group was observed where the accuracy of odd-numbered blocks (stimuli produced by the F1 talker) was consistently higher than even-numbered blocks (stimuli produced by the M1 talker). As mentioned in the Stimuli section, the semitone differences between T1-3 and T3-6 in the

[3] To corroborate this finding, the second half of the participants (n=39 with 19 and 20 in the HV and LV training condition, respectively) were assessed through a pretest and posttest identification task. As expected, there was a significant improvement from the pretest to posttest 1 ($\beta = 0.46$, SE = 0.44, $z = 10.62$, $p < .001$). The pretest-posttest improvement in tone identification suggested that the training paradigm effectively helped both groups to learn the categorization of the three Cantonese level tones.

F1 talker were significantly larger than those in the M1 talker, which could potentially explain the performance fluctuation.

Individual Differences in Training Progress

Pitch threshold and d-prime scores were two measures used to assess individual pitch aptitude before training. Pitch threshold captures participants' abilities in explicit categorization of pitch height patterns, whereas d-prime scores from the AX discrimination task were considered to assess the implicit sensitivity to pitch height differences (Fuhrmeister & Myers, 2020; Qin, Zhang, et al., 2021). To assess whether the two measures tapped independent dimensions of pitch aptitude, as assumed, correlation analysis was first performed between pitch threshold and d-prime scores. No significant correlation was found between the two pitch aptitude measures, r (78) = −0.08, p = .465. A Bayes factor test also provided evidence for the lack of correlation, BF_{10}= 0.18. In addition, both independent t-tests and Bayes factor tests revealed no difference between the HV and LV training groups in pitch threshold (in semitone; HV: M = 2.00, SD = 2.85; LV: M = 2.32, SD = 3.40), t (78) = −0.46, p = .647, BF_{10} = 0.25, and d-prime scores (HV: M = 2.39, SD = 0.34; LV: M = 2.33, SD = 0.41), t (78) = 0.69, p = .489, BF_{10} = 0.29. The two training groups were thus comparable in their pitch processing and pretraining discrimination abilities before learning Cantonese level tones.

To examine the (unique) effect of pitch threshold and d-prime on training performance, the two measures were both entered as continuous predictors in a mixed-effects logistic regression model, besides fixed effects of group (HV vs. LV) and block (initial vs. outcome). By-participant and by-tone intercepts were included as random effects. This analysis resulted in a main effect of group (β = −1.65, SE = 0.76, z = −2.17, p = .030),

a main effect of threshold ($\beta = -0.29$, SE = 0.06, $z = -4.61$, $p < .001$) and a main effect of d-prime scores ($\beta = 0.51$, SE = 0.16, $z = 3.12$, $p = .002$). The results indicate that both the pitch aptitude measures predicted the overall training performance. Importantly, the model resulted in a three-way interaction between pitch threshold, group, and block ($\beta = -0.15$, SE = 0.07, $z = -2.16$, $p = .031$). Figure 4 demonstrated this three-way interaction by showing that there was a negative relationship between pitch threshold and training improvement (i.e., percentage of increase, response accuracy of outcome – initial blocks) in the HV training group but not the LV training group. Post-hoc analyses revealed a marginally significant two-way interaction between pitch threshold and block in the HV group ($\beta = 0.09$, SE = 0.05, $z = 1.80$, $p = .072$), whereas the interaction was not significant in the LV group ($\beta = -0.06$, SE = 0.05, $z = -1.22$, $p = .223$). In other words, low-aptitude (i.e., high pitch threshold) individuals in the HV training group tended to show a larger increase in their response accuracy than those with high aptitude[4]. In contrast, pitch threshold did not seem to predict performance changes in the LV training group with low- and high-aptitude learners showing comparable improvements in tone categorization between the initial and outcome training blocks.

The model also yielded a significant interaction between d-prime scores and group ($\beta = 0.86$, SE = 0.32, $z = 2.69$, $p = .007$). D-prime scores predicted the overall training performance of the LV training group ($\beta = 0.94$, SE = 0.23, $z = 4.08$, $p < .001$), whereas no such relationship was found in the HV training group ($\beta = 0.08$, SE = 0.23, $z = 0.35$, $p = .725$). Importantly, unlike pitch threshold, the three-way interaction between d-prime,

[4] To better understand the relationship of pitch threshold and training variability, as seen in Figure S1, we visualized the interaction between pitch threshold and group on training performance across 10 blocks by splitting participants into high- and low-threshold subgroups based on the median pitch threshold. See S2 in supplementary materials for details.

group, and block was not significant (β = 0.18, SE = 0.19, z = 0.97, p = .333), suggesting that pre-training tone discrimination as an aptitude measure might not predict the training-induced improvement differently between the two groups.

Individual Differences in Consolidation

Consolidation was indexed by the overnight changes between ID posttest 1, before sleep, and ID posttest 2, after sleep (See S1 in supplementary materials for details of response accuracy in the ID tests). To test whether and how each training group's identification performance changed after sleep we ran a mixed-effects logistic regression model with test (i.e., time) nested within group as a fixed effect (Schad et al., 2020). By-participant and by-tone intercepts were entered as random effects. Results did not reveal a significant effect of group (β = −0.03, SE = 0.12, z = −0.27, p = .791). A non-significant effect of test was yielded for the HV training group (β = −0.03, SE = 0.05, z = −0.72, p = .471) and the LV training group (β = 0.05, SE = 0.05, z = 1.09, p = .274). The results indicate that the HV and LV training groups did not seem to show overnight changes in tone identification at a group level. However, it is possible that pitch aptitude affects the consolidation process and interacts with other factors such as training variability at an individual level.

To test whether pitch threshold and pretraining discrimination ability can predict the identification and consolidation of tone contrasts (also see S3 in supplementary materials; see S4 in supplementary materials for the talker effect), a mixed-effects logistic regression model was conducted. Again, we examined the effect of pitch threshold and d-prime scores by including both as fixed effects, in addition to test (ID posttest 1 vs. ID posttest 2) which was nested within group (HV vs. LV). By-participant

intercepts and by-tone intercepts were entered as random effects. Results of the model showed a main effect of threshold ($\beta = -0.22$, SE = 0.05, $z = -4.45$, $p < .001$) and a main effect of d-prime scores ($\beta = 0.26$, SE = 0.13, $z = 2.02$, $p = .044$), suggesting that participants with lower pitch threshold (i.e., higher aptitude) and higher d-prime scores had more correct responses in the tone identification tests.

Importantly, the model revealed a significant interaction between pitch threshold and test in the HV training group ($\beta = -0.13$, SE = 0.05, $z = -2.61$, $p = .009$), whereas no such effect was found in the LV training group ($\beta = 0.01$, SE = 0.04, $z = 0.21$, $p = .832$). The results, as illustrated in Figure 5, indicate that the relationship between pitch threshold and identification accuracy (i.e., percentage of change; response accuracy of ID 2 – ID 1) in the HV training group, but not in the LV training group, changed after sleep. After a day, low-aptitude learners did not consolidate what they learned at the training as well as high-aptitude learners, who did not experience such a performance decline at the ID posttest 2, in the HV training group. Unlike the pattern shown in the HV training group, the tone identification performances in the LV training group were not affected by the participants' pitch threshold. Regardless of their aptitude, participants in the LV training group had a similar level of overnight consolidation of newly learned level-tone categories. Taken together, pitch threshold, as an aptitude measure, only predicted 24-hour performance changes in the HV training group. In contrast to pitch threshold, the interaction between d-prime scores and test was not significant in either the HV group ($\beta = 0.08$, SE = 0.13, $z = 0.61$, $p = .544$) or the LV group ($\beta = -0.02$, SE = 0.11, $z = -0.17$, $p = .867$). The finding suggested that while pitch threshold predicted the performance changes between the two identification tests in the HV training group instead of the LV training group, pretraining

discrimination ability might not predict the overnight change differently between the HV and LV training groups.

Discussion

The current study investigated whether two measures of pitch aptitude predicted tone consolidation (i.e., the change after the 24-hour interval) as well as training progress differently through HV (i.e., exposure to inter-talker variability) and LV training (i.e., no exposure to inter-talker variability). Specifically, Mandarin-speaking participants' aptitude was assessed in terms of pitch threshold (in semitone) and pretraining discrimination ability (d-prime scores). Then they received Cantonese-tone identification training in either the HV condition or the LV condition. Crucially, the participants were all trained in the evening hours and returned after 24 hours for assessment of tone consolidation.

Regarding *overnight consolidation*, as expected, the results of identification overnight change revealed a relationship of pitch aptitude and training variability in tone consolidation. Pitch aptitude, measured as pitch threshold, predicted tone consolidation of the HV training group, but not that of the LV training group. Specifically, within the HV training group, learners with lower aptitude (i.e., higher pitch threshold) did not consolidate newly-learned tonal knowledge as well as those with higher aptitude (i.e., lower pitch threshold). The finding is consistent with recent findings of segmental learning, that is, individual aptitude (e.g., an implicit measure of pretraining discrimination ability) predicted learners' perception of non-native segments, with high-aptitude learners outperforming their low-aptitude counterparts, after an overnight sleep (Earle & Qi, 2021; Fuhrmeister & Myers, 2020). The explanation is also consistent with

the literature on sleep and memory consolidation (for a detailed review, see Earle & Myers, 2014).

Since the present study trained all participants in the evening and included an overnight sleep, one would expect the inter-talker variability in HV training to help consolidate tone knowledge for high-aptitude learners (not so much for low-aptitude learners) through the overnight consolidation process (Qin, Gong, et al., 2021). This expectation is confirmed in the results. One possible explanation for the finding is that high-aptitude learners who showed better perception and consolidation of non-native tones than low-aptitude learners might be better physiologically equipped to benefit from sleep-dependent memory processing (Tucker & Fishbein, 2008). For instance, strong positive associations were found for general aptitude tests (e.g., perceptual/analytical skills) and sleep-related psychological correlates (e.g., stage 2 spindle count) in prior studies of sleep and memory consolidation (e.g., Fogel et al., 2007). Furthermore, the possible advantage of high-aptitude learners over low-aptitude learners would be strengthened/amplified when the task was more difficult, for instance, during the HV training with inter-talker variability (also see Fuhrmeister & Myers, 2020 for a stronger relationship of aptitude and ID performance when training variability was increased). Overall, the findings of the current study suggest that individual learners had varying degrees of difficulty learning non-native tones (at least, partially) by virtue of the effect of pitch aptitude on tone consolidation (Fuhrmeister & Myers, 2020; Qin & Zhang, 2019). The findings imply that the early emergence of tonal representation is a dynamic process among individual learners, especially when they face training variability (e.g., the inter-talker variability of tones).

548 An Important difference between the current study and the previous ones is that
549 this study trained participants using a one-day training session (i.e., 40 min) and included
550 an overnight sleep for assessing consolidation (Fuhrmeister & Myers, 2020; Qin &
551 Zhang, 2019), whereas previous studies included multiple training sessions over days
552 (Dong et al., 2019; Perrachione et al., 2011; Sadakata & McQueen, 2014). The immediate
553 learning outcome after each training session was not separated from post-sleep changes in
554 the previous studies (Perrachione et al., 2011; Sadakata & McQueen, 2014). Specifically,
555 it is unclear whether the previous findings regarding the effect of individual aptitude in
556 HV training (e.g., Perrachione et al., 2011) were due to individual learners' poor initial
557 learning, poor consolidation, or a combination of the two. Our finding of tone
558 consolidation suggests that low-aptitude learners might benefit from training variability
559 immediately after training (before sleep) but still had greater difficulty in consolidating
560 tones than high-aptitude learners after the 24-hour interval. In other words, the
561 disadvantage of learning tones previously found for learners with lower aptitude in HV
562 training was possibly yielded by virtue of weaker memory consolidation, together with
563 poor initial learning (Fuhrmeister & Myers, 2020; Zion et al., 2019). The finding has
564 important implications in that sleep-mediated consolidation of non-native tones needs to
565 be assessed separately to deepen our understanding of how robust tonal representations
566 emerge in early stages of L2 tone learning at an individual level. Related to this, few
567 studies, as far as we know, have examined whether sleep-mediated consolidation
568 supports a long-term retention of learned linguistic knowledge from an individual
569 perspective (but see Zion et al., 2019 for individual differences in time-course of
570 consolidation effect across two consecutive nights). Future tone training studies are thus

suggested to separate post-training and post-sleep changes, and to examine the effect of sleep-mediated consolidation on long-term retention (e.g., weeks after training ended).

Another research objective was to examine whether pitch aptitude predicts tone *training* differently over the course of the HV and LV training. At a group level, as expected, the results of training progress showed that the participants in the HV and LV training groups (equally) improved their identification of Cantonese level tones over the course of training regardless of the variability conditions (Qin, Gong, et al., 2021). Consistent with the effect of tone training reported in previous training studies (Bowles et al., 2016; Perrachione et al., 2011), this finding suggests that both training groups improved as the training progressed and stabilized at or above 0.7 (chance-level of response accuracy at 0.33) at the end of the training session (for corroborative evidence of improvement from the pretest to posttest in half of the participants, see footnote 1). Notably, performance fluctuation in the HV training group was observed with the identification accuracy of even-numbered blocks (produced by the M1 talker) being consistently lower than that of odd-numbered blocks (produced by the F1 talker). One straightforward explanation of these results is that the identification of tones produced by the M1 talker may be more perceptually difficult than those produced by the F1 talker due to the reduced acoustic differences between the level tones (see Figure 1). An alternative explanation is related to processing interference that talker changes might have introduced in speech perception (Lim et al., 2021; C. Zhang et al., 2016). The pattern that talker changes between blocks yielded lower accuracy is in line with the results of recent neural studies which showed that presenting stimuli produced by

different talkers disrupted listeners' attention to cues of the target sounds and resulted in processing interference/cost (Lim et al., 2021).

The results further revealed a relationship of pitch aptitude and training variability in tone training progress. Pitch aptitude, measured as pitch threshold, seems to influence tone training (i.e., improvement) of the HV training group, but not that of the LV training group. Specifically, within the HV training group, learners with lower aptitude (i.e., higher pitch threshold) showed a numerical trend of a larger increase in their response accuracy than those with higher aptitude (i.e., lower pitch threshold). The current study reduced the overall and trial-to-trial variability in the HV training group to achieve the "desirable" level of training difficulty. Corroborating with previous studies (Fuhrmeister & Myers, 2020; Experiment 2 of Perrachione et al., 2011), the finding of tone training thus suggests that the talker-blocked HV training (i.e., with the trial-to-trial variability reduced) might enable learners with lower aptitude to deal with inter-talker variability and thus helped them attend to the target cues (i.e., pitch height) which were most relevant to learning the level-level tonal contrasts. It should be noted that the greater numerical trend of improvement found for low-aptitude learners did not necessarily mean that these learners outperformed their high-aptitude counterpart during HV training. The greater improvement could be alternatively driven by initial differences among individual learners with low-aptitude learners having lower accuracy than high-aptitude learners at the beginning of training. In other words, low-aptitude learners who initially underperformed might have gradually caught up with high-aptitude learners through HV training (for similar outcomes in the final training blocks of the two HV subgroups, see S2 in supplementary materials for details).

In addition, these results of pitch threshold indicate an interaction between training variability and pitch aptitude in tone training and consolidation. However, different from the finding of segmental learning and consolidation (Fuhrmeister & Myers, 2020; note that an explicit measure was not included), the results of pretraining discrimination abilities did not reveal a relationship between pitch aptitude and training variability in either tone training or consolidation. It is possible that, for learning to categorize tones produced by different talkers through the HV training, pitch threshold was a more reliable measure than discrimination abilities in assessing pitch aptitude for level-tone learning. The different results of pitch aptitude tests can be explained by the shared nature of pitch threshold test and tone identification (in both training and ID tests), which both required an explicit mapping between auditory stimulus and a tone category (or a pitch pattern). In contrast, tone discrimination is a different task testing a listener's implicit sensitivity to pitch height differences (e.g., Wayland & Li, 2008).

In closing, it is necessary to acknowledge some limitations of this study. As the current study employed Mandarin listeners who have tonal language experience as the learners, future studies should examine English listeners to see to what extent the current findings can generalize to non-tonal language speakers. It will be interesting to investigate whether the relative contribution of pitch threshold and pitch discrimination abilities differ depending on the learners' (tonal or non-tonal) language background. Moreover, due to the constraints of statistical analysis (e.g., accuracy 0-1 as dependent variables), we did not use relative scales of accuracy changes to assess training-induced improvement and consolidation-related changes. Rationalized transformation scale (e.g., arcsine-transformed accuracy Ingvalson et al., 2013, 2017) may also need to be used

639 when the starting scores of performance are close to ceiling or floor level (our
640 participants' performance was above the chance level but not close to ceiling).

641 Conclusion

642 In summary, the results of this study indicate that pitch aptitude, measured through
643 pitch threshold, may predict training and overnight consolidation of non-native tones by
644 the Mandarin-speaking HV trainees, but not by the LV trainees. Importantly, compared
645 with learners with higher pitch aptitude, learners with lower pitch aptitude (and tonal
646 language experience) benefitted temporarily from a "desirable" amount of training
647 variability but did not retain the learning after overnight consolidation. The findings have
648 theoretical implications for an individualized and dynamic emergence of tonal
649 representation in early stages of tone learning (e.g., immediate learning and subsequent
650 consolidation) by tonal language speakers. Future studies should continue to investigate
651 the nature of difficulty in learning and consolidating tones by non-tonal language speakers.

652 Acknowledgements

653 This research was supported by a Start-up Fund and the Anti-epidemic Fund 2.0
654 ("AEF") at the Division of Humanities, the Hong Kong University of Science and
655 Technology, awarded to Zhen Qin. The authors would like to thank Minzhi Gong for her
656 help in data collection.

657 Data Availability Statement (DAS)

658 The data that support the findings of this study is publicly available at OSF
659 (https://osf.io/2wkda/).

References

Antoniou, M., & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571–574. https://doi.org/10.1121/1.4923362

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: linear mixed-effects models using S4 classes. *Journal of Statistical Software*, *67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer [Computer program]. Version 6.0.43*. Retrieved 8 September 2018.

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch Ability As an Aptitude for Tone Learning. *Language Learning*, *66*(4), 774–808. https://doi.org/10.1111/lang.12159

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify english /r/and /1/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456–465. https://doi.org/10.1121/1.3445785

Chang, Y. S., Yao, Y., & Huang, B. H. (2017). Effects of linguistic experience on the perception of high-variability non-native tones. *The Journal of the Acoustical Society of America*, *141*(2), EL120–EL126. https://doi.org/10.1121/1.4976037

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, *2019*(8), e7191. https://doi.org/10.7717/peerj.7191

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. In *Frontiers in Psychology* (Vol. 5, Issue OCT, pp. 1–12). https://doi.org/10.3389/fpsyg.2014.01192

Earle, F. S., & Myers, E. B. (2015). Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds. *The Journal of the Acoustical Society of America*, *137*(1), EL91–EL97. https://doi.org/10.1121/1.4903918

Earle, F. S., & Qi, Z. (2021). Overnight changes to dual-memory processes reflected in speech-perceptual performance. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-021-02418-7

Espie, C. A., Kyle, S. D., Hames, P., Gardani, M., Fleming, L., & Cape, J. (2014). The Sleep Condition Indicator: A clinical screening tool to evaluate insomnia disorder. *BMJ Open*, *4*(3). https://doi.org/10.1136/bmjopen-2013-004183

Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition*, *128*(3), 280–286. https://doi.org/10.1016/j.cognition.2013.04.007

Fogel, S. M., Nader, R., Cote, K. A., & Smith, C. T. (2007). Sleep spindles and learning potential. *Behavioral Neuroscience*, *121*(1), 1–10. https://doi.org/10.1037/0735-7044.121.1.1

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454. https://doi.org/10.1121/1.5009688

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, and Psychophysics*, *82*, 2049–2065. https://doi.org/10.3758/s13414-019-01925-y

Gandour, J. T. (1983). Tone perception in far eastern-languages. *Journal of Phonetics*, *11*(2), 149–175.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A Tutorial on Testing Hypotheses Using the Bayes Factor. *Psychological Methods*. https://doi.org/10.1037/met0000201

Ingvalson, E. M., Barr, A. M., & Wong, P. C. M. (2013). Poorer phonetic perceivers show greater benefit in phonetic-phonological speech learning. *Journal of Speech, Language, and Hearing Research*, *56*(3), 1045–1050. https://doi.org/10.1044/1092-4388(2012/12-0024)

Ingvalson, E. M., Nowicki, C., Zong, A., & Wong, P. C. M. (2017). Non-native speech learning in older adults. *Frontiers in Psychology*, *8*(FEB), 1–10. https://doi.org/10.3389/fpsyg.2017.00148

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics*, *63*(8), 1279–1292. https://doi.org/10.3758/BF03194543

Li, M., & DeKeyser, R. (2017). Perception Practice, Production Practice, and Musical Ability in L2 Mandarin Tone-Word Learning. *Studies in Second Language Acquisition*, *39*(4), 593–620. https://doi.org/10.1017/s0272263116000358

Lim, S. J., Carter, Y. D., Njoroge, J. M., Shinn-Cunningham, B. G., & Perrachione, T. K. (2021). Talker discontinuity disrupts attention to speech: Evidence from EEG and pupillometry. *Brain and Language*, *221*. https://doi.org/10.1016/j.bandl.2021.104996

Lively, S. E., Pisoni, D. B., Yamada, R. A., Yoh'ichi, T., & Yamada, T. (1994). Training japanese listeners to identify english /r/ and /l/. iii. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese Listeners To Identify English /R/ And /1/: A First Report. *Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Macmillan, N. A., & Creelman, C. D. (2004). Detection Theory: A User's Guide: 2nd edition. In *Detection Theory: A User's Guide: 2nd edition*. Psychology, Mahwah, N.J. https://doi.org/10.4324/9781410611147

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McHaney, J. R., Tessmer, R., Roark, C. L., & Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from

computational modeling and pupillometry. *Brain and Language*, *222*. https://doi.org/10.1016/j.bandl.2021.105010

Melnik, G. A., & Peperkamp, S. (2020). High-Variability Phonetic Training enhances second language lexical processing: evidence from online training of French learners of English. *Bilingualism (Cambridge, England)*, 1–10. https://doi.org/10.1017/S1366728920000644

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2021). Bayesfactor: Computation of bayes factors for common designs. R package version 0.9. 12-4.3. In *CRAN Repository*. https://cran.r-project.org/web/packages/BayesFactor/index.html

Ou, J., & Law, S. P. (2017). Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Attention, Perception, and Psychophysics*, *79*(3), 945–963. https://doi.org/10.3758/s13414-017-1283-z

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of mandarin and cantonese. *Journal of Chinese Linguistics*, *34*(1), 134–154.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & Wang, W. S.-Y. (2011). The Effect of Intertalker Variations on Acoustic–Perceptual Mapping in Cantonese and Mandarin Tone Systems. *Journal of Speech, Language, and Hearing Research*, *55*(2), 579–595. https://doi.org/10.1044/1092-4388(2011/11-0025)

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. *Annals of*

770 *the New York Academy of Sciences*, *999*(1), 58–75.

771 Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel
772 phonological contrast depends on interactions between individual differences and
773 training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1),
774 461–472. https://doi.org/10.1121/1.3593366

775 Qin, Z., Gong, M., & Zhang, C. (2021). Neural responses in novice learners' perceptual
776 learning and generalization of lexical tones: The effect of training variability. *Brain*
777 *and Language*, *223*, 105029.
778 https://doi.org/https://doi.org/10.1016/j.bandl.2021.105029

779 Qin, Z., & Jongman, A. (2016). Does Second Language Experience Modulate Perception
780 of Tones in a Third Language? *Language and Speech*, *59*(3), 318–338.
781 https://doi.org/10.1177/0023830915590191

782 Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual
783 learning of non-native tonal contrasts. *PLOS ONE*, *14*(12), e0221498.
784 https://doi.org/10.1371/journal.pone.0221498

785 Qin, Z., Zhang, C., & Wang, W. S. (2021). The effect of Mandarin listeners' musical and
786 pitch aptitude on perceptual learning of Cantonese level-tones. *The Journal of the*
787 *Acoustical Society of America*, *149*(1), 435–446. https://doi.org/10.1121/10.0003330

788 Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of
789 normal human attention: The Test of Everyday Attention. *Journal of the*
790 *International Neuropsychological Society*, *2*(6), 525–534.
791 https://doi.org/10.1017/s1355617700001697

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*(NOV), 1318. https://doi.org/10.3389/fpsyg.2014.01318

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*. https://doi.org/10.1016/j.jml.2019.104038

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep*, *31*(2), 197–203. https://doi.org/10.1093/sleep/31.2.197

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, *54*(4), 681–712. https://doi.org/10.1111/j.1467-9922.2004.00283.x

Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics*, *36*(2), 250–267. https://doi.org/10.1016/j.wocn.2007.06.004

Williamson, V. J., & Stewart, L. (2010). Memory for pitch in congenital amusia: Beyond a fine-grained pitch discrimination problem. *Memory*, *18*(6), 657–669. https://doi.org/10.1080/09658211.2010.501339

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Wong, P., & Lam, K. Y. (2021). Characteristics of effective auditory training: Implications from two training programs that successfully trained nonnative cantonese tone identification in monolingual mandarin and bilingual mandarin–taiwanese tone speakers. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2490–2512. https://doi.org/10.1044/2021_JSLHR-20-00436

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1252–1268. https://doi.org/10.1037/xhp0000216

Zhang, C., Ho, O. Y., Shao, J., Ou, J., & Law, S. P. (2021). Dissociation of tone merger and congenital amusia in Hong Kong Cantonese. *PloS One*, *16*(7), e0253982. https://doi.org/10.1371/journal.pone.0253982

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & Wang, W. S. Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*, *124*, 536–549. https://doi.org/10.1016/j.neuroimage.2015.08.064

Zhang, X., Cheng, B., & Zhang, Y. (2021). The Role of Talker Variability in Nonnative Phonetic Learning: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825. https://doi.org/10.1044/2021_jslhr-21-00181

Zion, D. Ben, Nevat, M., Prior, A., & Bitan, T. (2019). Prior knowledge predicts early consolidation in second language learning. *Frontiers in Psychology*, *10*(OCT), 1–15. https://doi.org/10.3389/fpsyg.2019.02312

Figure Captions

Figure 1. Tonal contours of the three Cantonese level tones produced by the trained female talker (F1; left) and the male talker (M1; right) in semitone (reference 50 Hz). Tonal contours were measured using ten measurement points.

Figure 2. Schematic of the experimental procedure and tasks.

Figure 3. Training progress (i.e., proportion of correct responses across the 10 training blocks) of the HV training group (red) and the LV training group (green). Stimuli produced by the female talker (F1; odd-number blocks) and the male talker (M1; even-number blocks) alternated between blocks for the HV training group. Error bars indicate standard error of the mean.

Figure 4. Training improvement (i.e., percentage of increase in response accuracy) of individuals with different pitch threshold in the HV training group (red) and the LV training group (green). Higher pitch threshold indicates lower aptitude, and vice versa. The shaded area indicates 95% confidence intervals.

Figure 5. Overnight consolidation (i.e., percentage difference in response accuracy between ID posttest 2 and posttest 1) of individuals with different pitch thresholds from the HV training group (left; red) and the LV training group (right; green). The dashed line indicates no change in performance. Higher pitch threshold indicates lower aptitude, and vice versa. The shaded area indicates 95% confidence intervals.

Supplementary Materials

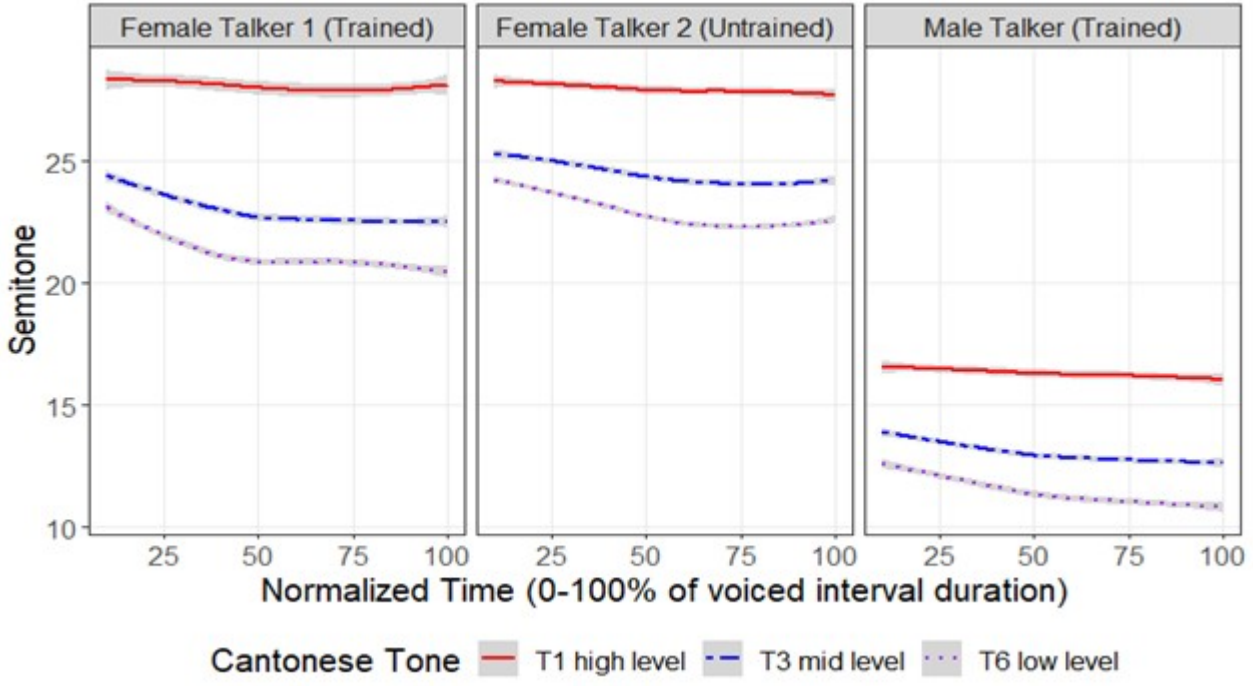
Figure S1. Proportion of correct responses (0-1) across 10 training blocks of the HV training group (left; red) and LV training group (right; green). The high-aptitude (solid lines) and low-aptitude (dashed lines) subgroups were divided based on the median pitch threshold for the purpose of *visualization*.

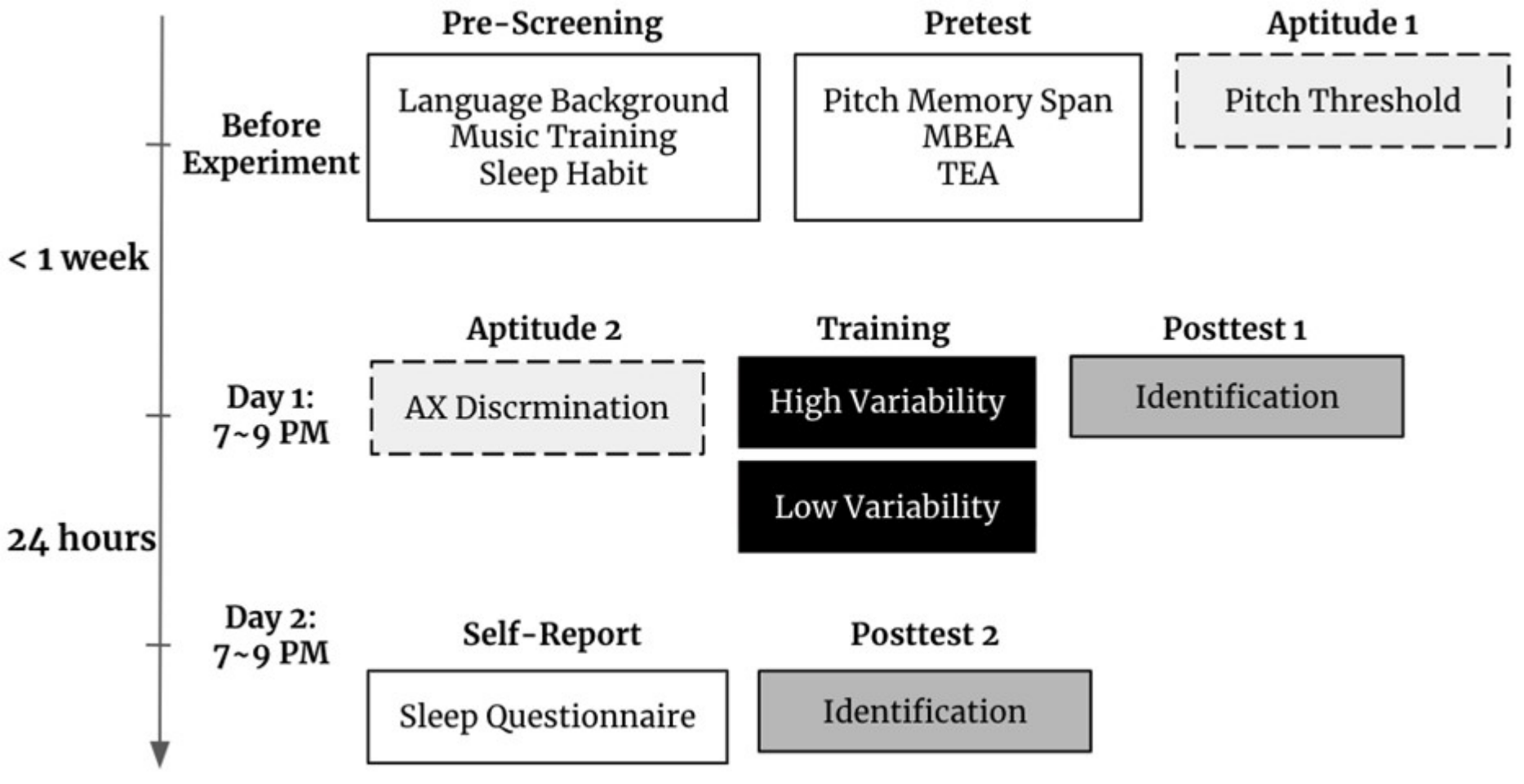
Figure S2. Proportion of correct responses in two identification tests (ID Posttest1 and Posttest 2) of two training groups (HV vs. LV) predicted by individual pitch threshold. The shaded area indicates 95% confidence intervals.

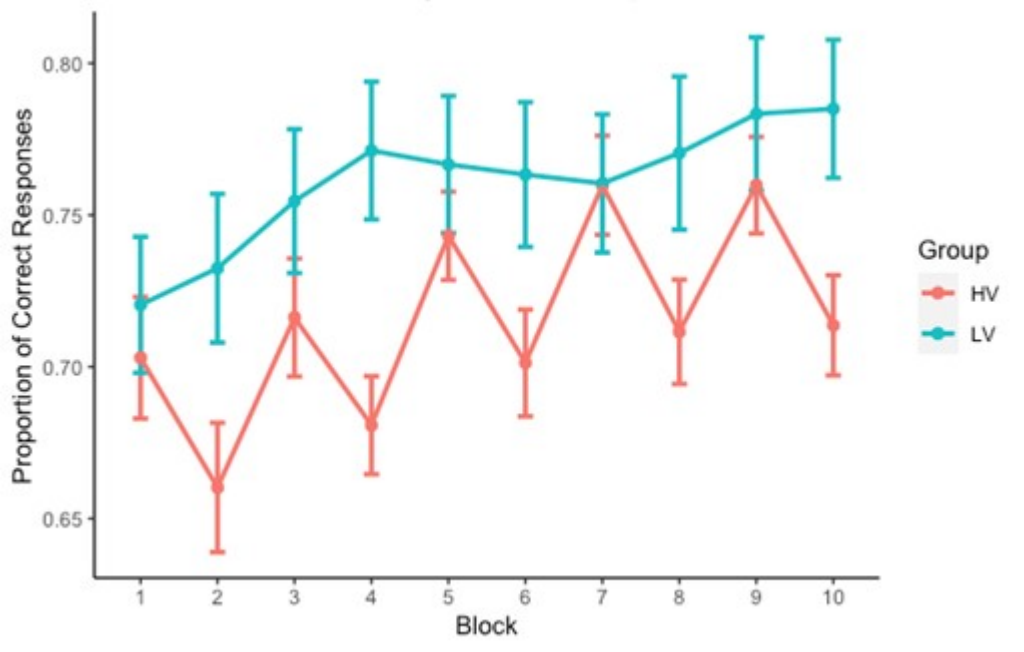
Table 1. Demographics, pretest performance, and sleep questionnaire results of the HV and LV training groups. For the age, pretests and sleep questionnaire, the numbers indicate the mean and SD (in brackets) of the HV and LV training group respectively.

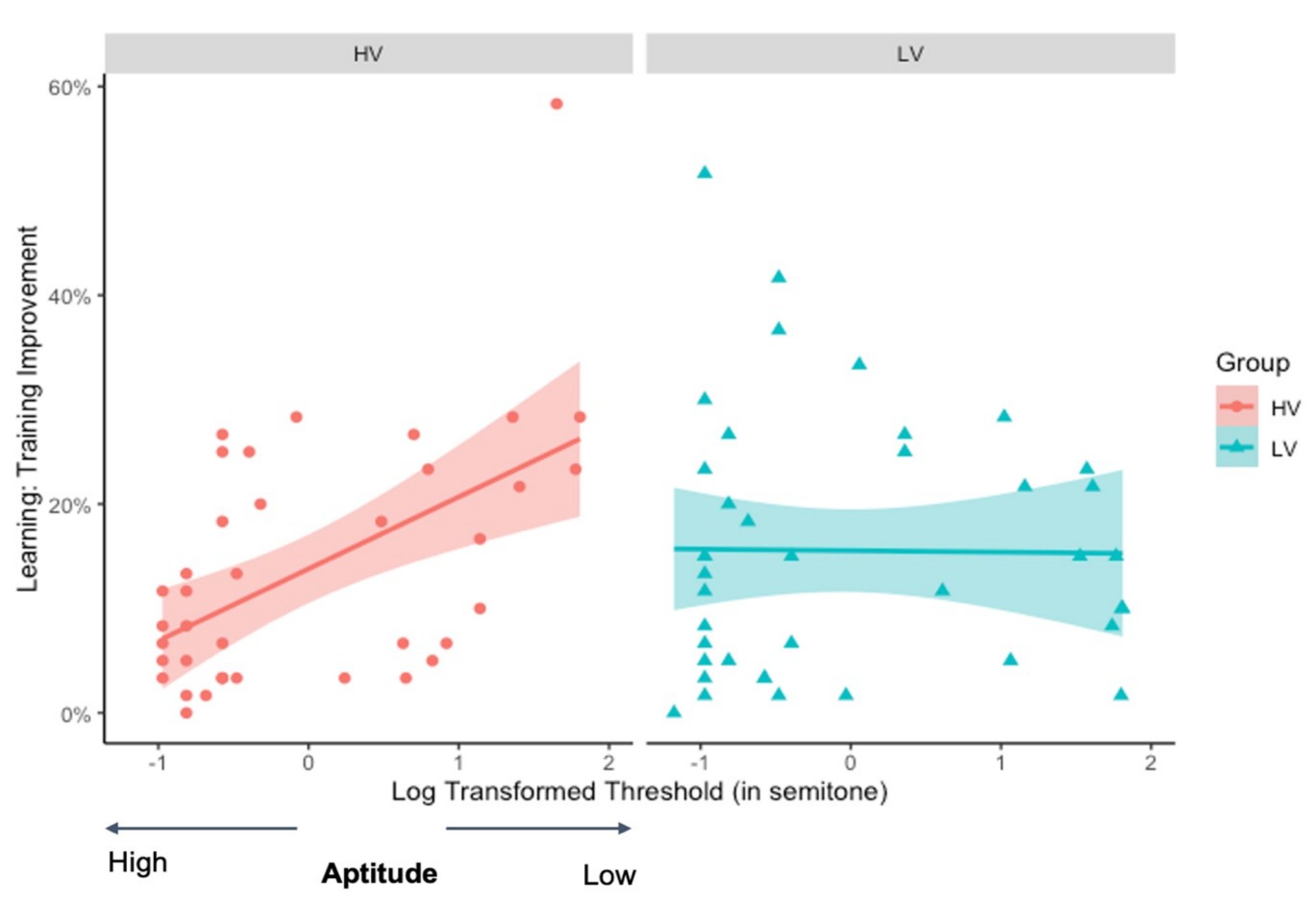
	HV	**LV**	***p* value**	***BF*$_{10}$**
No. of participants	40 (25F, 15M)	40(25F, 15M)		
Age (year)	24.90 (2.43)	25.30 (2.24)	.446	0.30
Pretests				
Pitch memory span (no. of tones)	6.47 (1.39)	5.96 (1.66)	.740	0.62
MEBA pitch (%)	81 (8.35)	82 (8.61)	.489	0.25
TEA (e.g., count of correct trials)	11.44 (1.25)	11.34 (1.45)	.275	0.24
Sleep Questionnaire				
Sleep duration (hr)	7.68 (0.85)	7.54 (1.12)	.694	0.25
Sleep quality (rating 1-10)	6.82 (1.59)	6.63 (2.03)	.636	0.26
Time spent before sleep (hr)	2.65 (0.98)	2.46 (1.07)	.430	0.31

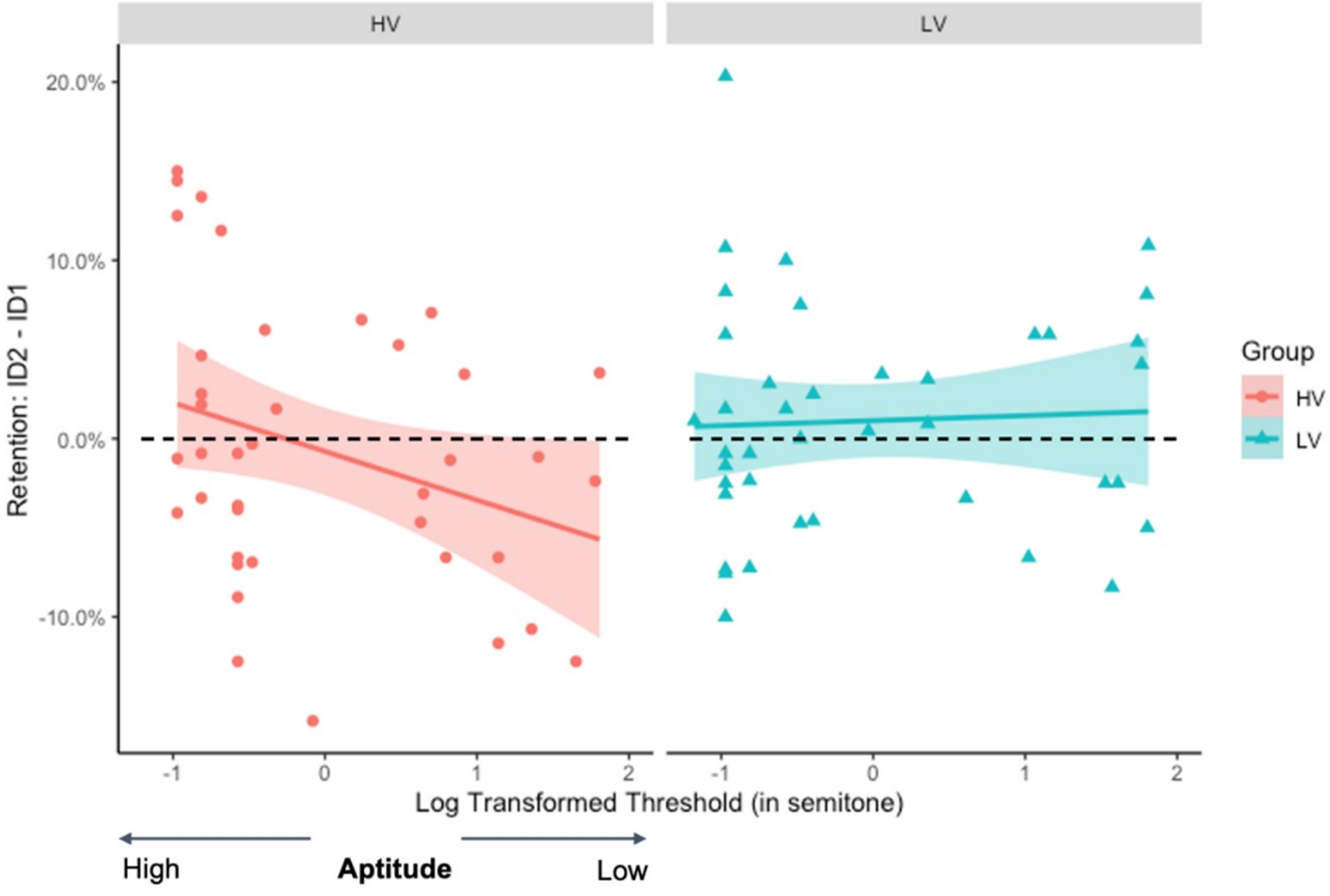
Note. Independent t-tests and Bayes factor tests revealed no significant differences between the two groups in any of these measures; MBEA = Montreal Battery of Evaluation of Amusia; TEA = Test of Everyday Attention.


Female Talker 1 (Trained)
Female Talker 2 (Untrained)
Male Talker (Trained)
Semitone
25
20
15
10
25
50
75
100
Normalized Time (0-100% of voiced interval duration)
Cantonese Tone
T1 high level
T3 mid level
T6 low level


Pre-Screening
Pretest
Aptitude 1
Before Experiment
Language Background
Music Training
Sleep Habit
Pitch Memory Span
MBEA
TEA
Pitch Threshold
< 1 week
Aptitude 2
Training
Posttest 1
Day 1: 7~9 PM
AX Discrmination
High Variability
Identification
Low Variability
24 hours
Day 2: 7~9 PM
Self-Report
Posttest 2
Sleep Questionnaire
Identification


0.80
0.75
0.70
0.65
Proportion of Correct Responses
1
2
3
4
5
6
7
8
9
10
Block
Group
HV
LV


HV
LV
60%
40%
20%
0%
Learning: Training Improvement
-1
0
1
2
Log Transformed Threshold (in semitone)
Group
HV
LV
High
Aptitude
Low


HV
LV
20.0%
10.0%
0.0%
-10.0%
Retention: ID2 - ID1
-1
0
1
2
Log Transformed Threshold (in semitone)
Group
HV
LV
High
Aptitude
Low

Supplementary Materials

S1. Additional Details of Response Accuracy

We conducted a series of one-sample t-tests and Bayes factor tests to verify whether participants' proportion of correct response (i.e., 0-1) was close to ceiling during or after training.

Table S1. The descriptive and statistical results of participants' response accuracy (0-1) of the training session and the two ID posttests.

		Mean (SD)	*t*	*p*	BF_{10}
Training Initial Blocks					
	HV	0.68 (0.12)	−16.39	<.001	>100
	LV	0.73 (0.14)	−12.09	<.001	>100
Training Outcome Blocks					
	HV	0.74 (0.10)	−17.13	<.001	>100
	LV	0.78 (0.15)	−9.15	<.001	>100
ID Posttest 1					
	HV	0.67 (0.11)	−22.98	<.001	>100
	LV	0.65 (0.15)	−17.12	<.001	>100
ID Posttest 2					
	HV	0.66 (0.14)	−17.75	<.001	>100
	LV	0.66 (0.13)	−18.20	<.001	>100

As illustrated in Table S1, the response accuracy of both the groups was significantly different from the ceiling performance (i.e., 1.0 proportion of correct response) at the initial blocks (i.e., first two blocks) and outcome blocks (i.e., final two blocks), indicating that neither group was had ceiling performance *during* training. Likewise, the same tests were conducted for the immediate ID test (i.e., posttest 1) and the 24-hour delayed ID test (i.e., posttest 2). The results suggest that neither group had ceiling performance *after* training.

S2. Additional Details of Individual Differences in Training Progress

To better understand the relationship of pitch threshold and training variability in training progress, as seen in Figure S1, we visualized the interaction between pitch threshold and group on training performance across 10 blocks by splitting participants into high- and low-threshold subgroups based on the median pitch threshold. It should be noted that there was a negative relationship between pitch threshold and aptitude. That is, a higher threshold means lower pitch aptitude.

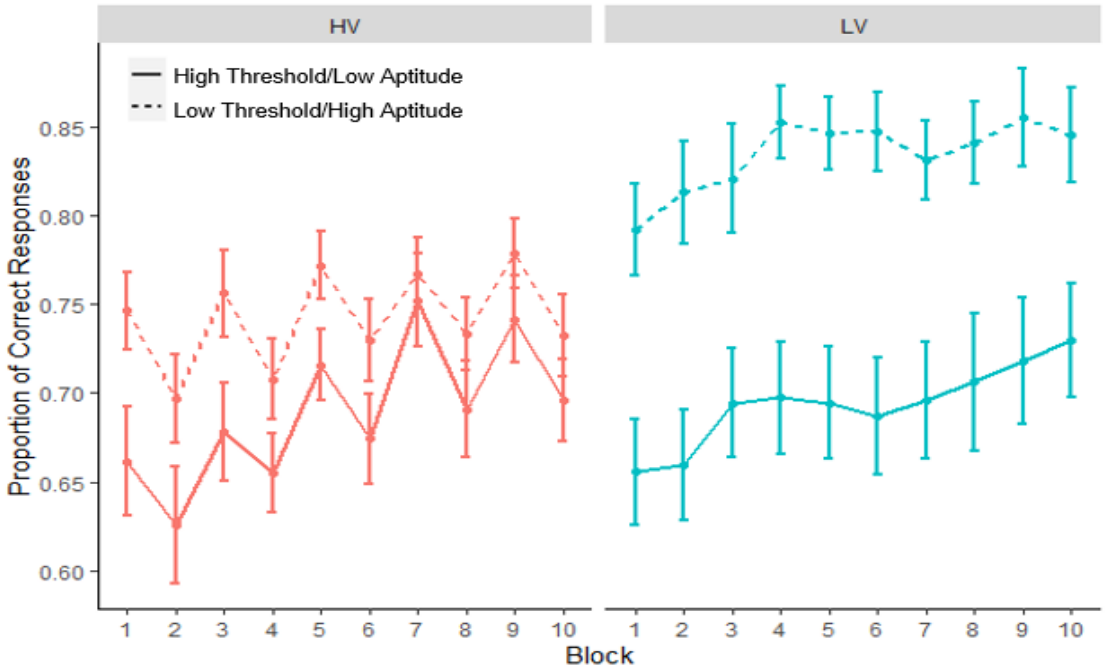


Figure S1. Proportion of correct responses (0-1) across 10 training blocks of the HV training group (left; red) and LV training group (right; green). The high-aptitude (solid lines) and low-aptitude (dashed lines) subgroups were divided based on the median pitch threshold for the purpose of visualization.

The figure on the right showed that individuals in the LV training group held their initial differences throughout the training. High-aptitude learners (i.e., mean accuracy: 0.80)

and low-aptitude learners (i.e., mean accuracy: 0.66) of the LV group had different starting points of accuracy at the initial blocks but showed a similar degree of improvement (e.g., an increase of accuracy by 0.05) at the end of training. In contrast, for the HV training group illustrated in the left figure, low-aptitude learners had lower accuracy than high-aptitude learners in the initial blocks but gradually caught up with high-aptitude learners in the outcome blocks (i.e., mean accuracy around 0.70).

S3. Additional Details of Individual Differences in Consolidation

To better illustrate the relationship of pitch threshold, training variability and test (i.e., time) in the consolidation of non-native tones, as seen in Figure S2, we plotted the raw accuracy of tone identification at an individual level. The plot highlights individual differences of identification accuracy along with pitch threshold in the posttest 1 (i.e., an immediate test) and in the posttest 2 (i.e., a re-test with a 24-hour delay).

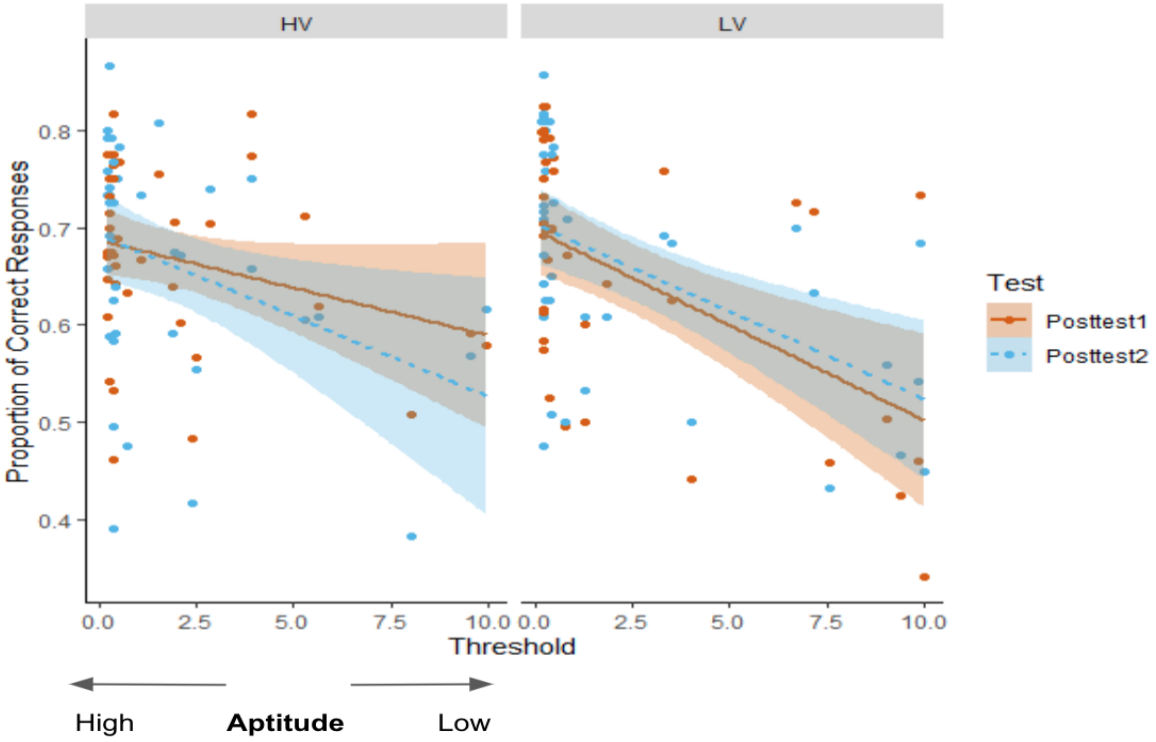


Figure S2. Proportion of correct responses in two identification tests (ID Posttest1 and Posttest 2) of two training groups (HV vs. LV) predicted by individual pitch threshold. The shaded area indicates 95% confidence intervals.

Consistent with the results reported in the main text, for HV group, the relationship between pitch threshold and response accuracy became stronger after the 24-hour interval at posttest 2 (blue dashed line) than that of posttest 1 (brown solid line). In HV group, higher aptitude individuals retained their performance after a day whereas lower aptitude

individuals did not (i.e., larger individual differences of identification accuracy). However, no such an over-time change was found in the LV group. The relationship between pitch threshold and identification performance was held constant during the 24-hour period.

S4. Identification Performance on Untrained-Novel Talker

Referring back to our method section, in the identification tests, we included an untrained female talker (F2) in addition to the trained talker (F1). It is our intention to examine whether the current training paradigm is effective in generalizing trained level tone categorization to stimuli produced by an untrained novel talker, and whether the effect of talker interacts with the consolidation process discussed in the main text.

First, participants were able to generalize trained categorization such that their performance was significantly different from the chance level of response accuracy (i.e., 0.33) on the trained stimuli (mean accuracy: 0.71), t (157) = 35.98, $p < 0.001$, and untrained stimuli (mean accuracy: 0.65), t (157) = 30.63, $p < 0.001$, in the identification tests.

Second, in the main text, the aptitude model on identification revealed a significant interaction between pitch threshold and test at HV but not LV group. To test whether this interaction can be further explained by the effect of talker (2 levels: F1 Trained vs. F2 Untrained; deviation coding: −0.5, .05), we ran a mixed effects model with talker, tested nested within group, and pitch threshold as fixed effects, and by-participant and by-tone intercepts as random effects. Results showed a main effect of talker ($\beta = -0.44$, SE = 0.03, $z = -13.55$, $p < .001$), indicating better tone categorization in stimuli produced by the trained talker (F1) than the untrained talker (F2). There was significant interaction between talker and pitch threshold ($\beta = 0.10$, SE = 0.03, $z = 3.02$, $p = .003$). Post-hoc analyses revealed that although different in magnitude, for both trained and untrained talker, pitch threshold predicted overall identification accuracy (Trained: $\beta = -0.33$, SE = 0.07, $z = -4.97$, $p < .001$; Untrained: $\beta = -0.20$, SE = 0.05, $z = -3.99$, $p < .001$).

The results showed that training variability and pitch aptitude influenced tone consolidation (and training progress), however, neither of them contributed to the talker

generalization of tone categorization (i.e., with identification accuracy of the stimuli produced by the untrained talker being significantly above chance). Training variability has been shown to promote generalization to new talkers in phonetic training studies on segmental contrasts (Bradlow et al., 1999; Lively et al., 1994; Logan et al., 1991). For tone learning, Perrachione et al. (2011) showed that, regardless of pitch aptitude, learners had better generalization abilities (measured by a ratio of posttest performance with untrained talkers and training performance with trained talkers; see note 1 of Dong et al., 2019) following the HV training than the LV training. However, Dong et al. (2019) found no evidence for better generalization following the HV training than the LV training when comparing the accuracy of stimuli produced by the trained and untrained talkers in the ID posttests. Thus, the finding is not conclusive regarding the effect of training variability, which is further modulated by other factors such as training length and talker presentation, on talker generalization in tone learning (for a systematic review, see Zhang et al., 2021). On the other hand, the issue of talker generalization became complicated for the current design including an overnight sleep, which potentially has promoted both the consolidation of trained materials as well as the generalization to novel talkers (Qin & Zhang, 2019). Thus, the overnight consolidation might have obscured the effect of talker generalization, resulting in a lack of interaction between talker and other factors (e.g., training variability). Since learners need to efficiently map variable acoustic input to tonal categories produced by different talkers for successful learning, this research calls for future studies to test the effect of training variability (and pitch aptitude) on tone consolidation and generalization.

References

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify english /r/and /1/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Lively, S. E., Pisoni, D. B., Yamada, R. A., Yoh'ichi, T., & Yamada, T. (1994). Training japanese listeners to identify english /r/ and /l/. iii. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese Listeners To Identify English /R/ And /1/: A First Report. *Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472. https://doi.org/10.1121/1.3593366

Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLOS ONE*, *14*(12), e0221498. https://doi.org/10.1371/journal.pone.0221498

Zhang, X., Cheng, B., & Zhang, Y. (2021). The Role of Talker Variability in Nonnative Phonetic Learning: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825.

https://doi.org/10.1044/2021_jslhr-21-00181