Observing Pitch Gestures Modulates Sleep-Dependent Consolidation in L2 Lexical Tone Learning

Derrick Nguyen (dnguyen@health.missouri.edu)
Department of Speech, Language and Hearing Sciences, University of Missouri

F. Sayako Earle (fsearle@udel.edu)
Department of Communication Sciences & Disorders, University of Delaware

Laura M. Morett (lmorett@health.missouri.edu)
Department of Speech, Language and Hearing Sciences, University of Missouri

Abstract

This study examined the effect of sleep on L2 Mandarin lexical tone acquisition by L1 English speakers and how observing pitch gestures modulates this effect. Sixty-four L1 American English speakers without prior tonal language experience learned monosyllabic L2 Mandarin words with or without observing pitch gestures and were tested at pre-test, immediate post-test, and a 24-hour delayed post-test. Higher sleep quality predicted improved accuracy and decreased response latency in the delayed post-test in the no gesture condition. Observing pitch gestures mitigated these sleep-dependent gains, suggesting that pitch gestures may enhance initial encoding so that overnight consolidation is less critical. These results extend research examining the impact of sleep on L2 tone acquisition to atonal L1 speakers and reveal interactions between multimodal learning and sleep-based memory consolidation.

Keywords: pitch gestures; lexical tone; sleep; memory consolidation; second language acquisition; multimodal learning; speech perception

Introduction

Learning a second language (L2) can be challenging when it involves unfamiliar speech sounds (Flege, 1995). This difficulty is especially salient for tonal L2s such as Mandarin, in which lexical tones convey meaning via pitch (Yip, 2002). Distinguishing between L2 lexical tones is difficult for learners without prior tonal language experience, such as L1 English speakers (Kiriloff, 1969; Wang et al., 2006). In the present study, we examine whether observing pitch gestures during learning and subsequent nocturnal sleep enhances adult L1 English speakers' acquisition of L2 Mandarin lexical tones.

Auditory training enhances Mandarin L2 lexical tone perception by atonal L1 speakers (Wang et al., 2003; Morett & Chang, 2015; Zhen et al., 2019; Morett et al., 2022; Wei et al., 2022), with effects persisting beyond six months (Wang et al., 1999). While musical experience has been associated with enhanced pitch perception and may influence sensitivity to lexical tone contrasts (Wong et al., 2007), measurements of learning are assessed by two types of tasks. Lexical tone learning can be assessed with classification tasks, which measure the strength of category representations, and discrimination tasks, which measure acoustic sensitivity. However, most studies only assess classification, with discrimination either omitted or administered in separate blocks or phases, limiting comparison between the two outcomes (Hallé et al., 2004; Hannah et al., 2017; Baills et al., 2019; Zhen et al., 2019; Morett et al., 2022). Moreover, improvement across talkers may reflect item familiarity rather than category generalization when the same words are reused (Nygaard & Pisoni, 1998; Logan et al., 1991).

Beyond auditory training, multimodal approaches that supplement L2 lexical tones with visual cues support their acquisition. One visual cue that supports L2 lexical tone acquisition is pitch gestures. Pitch gestures consist of hand movements conveying pitch height and contour via the vertical conceptual metaphor of pitch, in which higher vertical hand positions correspond to higher pitch and lower positions to lower pitch (Morett et al., 2022; Casasanto et al., 2003; Liu et al. 2011). However, atonal L1 speakers tend to focus on pitch height rather than contour (Wang et al., 1999; Wayland & Guion, 2004). This can be problematic as the contrast between Tone 2 (rising) and Tone 3 (low dipping) is defined primarily by pitch contour rather than height, so attending only to pitch height hinders differentiation (Wang et al., 1999; Yip, 2002). This problem is addressed by observing pitch gestures, which promotes encoding of both pitch contour and height in L2 lexical tones via multimodal input. By highlighting the dimensions of pitch contours, pitch gestures may facilitate the abstraction of tone categories by engaging sensorimotor representations, consistent with accounts of embodied cognition (Barsalou, 2008), while leveraging cross-domain mappings between pitch and spatial representations. (Lakoff & Johnson, 1980; Casasanto, 2003). Indeed, observing pitch gestures during training of the lexical tones of L2 words results in subsequent perceptual gains (Baills et al., 2019; Hannah et al., 2017; Zhen et al., 2019; Morett et al., 2022). By engaging the visual and auditory systems, mental

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

representations of L2 lexical tones are strengthened (Paivio, 1991), supporting the efficacy of observing pitch gestures in lexical tone acquisition (Shams & Seitz, 2008).

Under the Complementary Learning Systems framework, consolidation refers to offline processes through which newly encoded memory traces are integrated with existing knowledge, supporting abstraction and generalization of newly acquired episodic information (McClelland et al., 1995; Davis & Gaskell, 2009). Studies on L2 learning after sleep demonstrate that sleep has been linked to improvements in category formation and unfamiliar L2 phoneme generalization (Earle & Myers, 2015b; Earle et al., 2017). Thus, because observing pitch gestures during learning may influence the structure of encoded representations along with the strength of encoding, it may influence how effectively sleep supports consolidation of L2 lexical tones. In alignment with the current study, Qin & Zhang (2019) measured L2 lexical tone perception via classification and discrimination tasks in relation to sleep in tonal L1 speakers. After training, participants completed delayed post-tests after 12 and 24 hours, and earlier sleep times predicted greater classification gains. In a follow up study, classification gains were observed for tonal L1 speakers who napped after learning L2 lexical tones compared to the non-napping group (Qin et al., 2025). Other studies provide evidence that sleep-related benefits may be task-specific and often favor classification over discrimination (Earle & Myers, 2015a, 2015b, Earle et al., 2017). These studies demonstrate effects of sleep timing and duration, which do not capture all sleep-dependent processes related to consolidation, on L2 perceptual learning (Earle & Myers, 2015a, 2015b; Earle et al., 2017; Qin & Zhang, 2019, Qin et al., 2025). Notably, sleep does not benefit immediate recall following multimodal category learning after a period of wakefulness or sleep (Ashton et al., 2018). As such, it is currently unclear how observing pitch gestures interacts with sleep-related consolidation in L2 lexical tone classification and discrimination following a delay, especially in atonal L1 speakers.

The present study addresses whether learning with pitch gestures alters the impact of sleep-related consolidation on L2 lexical tones. Whereas previous work shows that observing pitch gestures during learning improves L2 lexical tone classification (Baills et al., 2019; Eng et al., 2013; Hannah et al., 2017; Morett et al., 2022; Zhen et al., 2019; Farran & Morett, 2024) and that self-reported nocturnal sleep is associated with classification accuracy for segmental and suprasegmental L2 speech sounds (Earle & Myers, 2015; Qin & Zhang, 2019), it remains unclear whether observing pitch gestures supports or modulates sleep-related consolidation of L2 lexical tones. By examining the impact of self-reported sleep quality while experimentally manipulating observation of pitch gestures during learning, the present study tests the hypothesis that observing pitch gestures modulates the relationship between sleep quality and L2 lexical tone acquisition.

Methods

Participants

A total of 85 participants were recruited. 21 participants were excluded prior to analysis for the following reasons: incomplete data/withdrawal (n = 15), prior tonal language experience (n = 3), and internet connectivity issues (n = 3). Therefore, the final sample consisted of 64 L1 English speakers. Participants' ages ranged from 18 to 38 years (M = 24.31 years, SD = 5.72 years), with 94% identifying as women, 4% identifying as men, and 1% identifying as non-binary. The sample was predominantly female (94%), which may limit the generalizability of findings.

Materials

Mandarin words used in this study were monosyllabic, consisted of segmental phonemes present in both Mandarin and English, and differed minimally in lexical tone. 28 quartets of Mandarin words with each of four lexical tones were divided into 4 lists consisting of 7 quartets (e.g., cai1, cai2, cai3, cai4). One of the four lists served as repeated words, which were presented in all study phases (pre-test, immediate post-test, learning, and delayed post-test). The other three lists served as non-repeated words, with one list presented in each test (pre-test, immediate post-test, or delayed post-test). Non-repeated words complemented repeated words by providing insight into generalization of lexical tones to unfamiliar words. Thus, each test contained 14 quartets of Mandarin words (7 repeated, 7 non-repeated), with list assignment to study phase counterbalanced across participants.

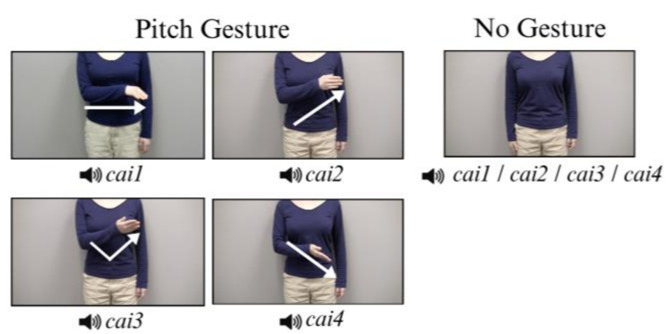

Figure 1: Screenshots of Videos Featuring Actor Producing Pitch Gestures or No Gestures with Examples of Mandarin Words Presented Concurrently.

Video stimuli featured one of two actors, one of whom was female and one of whom was male. Videos for the pitch gesture condition depicted an actor producing a gesture conveying the pitch contour of one of the four Mandarin lexical tones (see Figure 1). Videos in the no gesture condition depicted an actor standing still. In post-

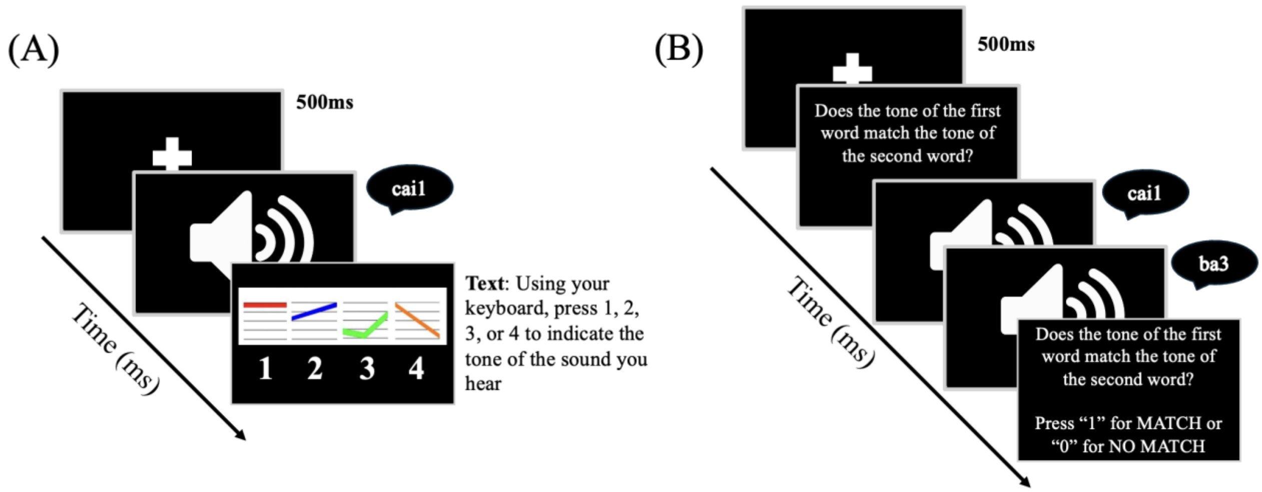

Figure 2: Schematics of (A) Classification and (B) Discrimination Trials

production, these videos were paired with audio clips of Mandarin words containing the lexical tones that they depicted.

Procedure

The study was conducted online, and participants completed all tasks remotely using their own devices. Participants were instructed to use headphones and complete the study in a quiet environment. Participants completed a pre-test, learning phase, immediate post-test, and delayed post-test. In the pre-test, participants completed classification and discrimination tasks to the best of their ability prior to learning Mandarin lexical tones. In the lexical tone classification task, audio clips of Mandarin words were played while graphics displaying the pitch contours of the four Mandarin lexical tones was presented, and participants selected the pitch contour of the lexical tone of each word they heard. In this task, words in each quartet were presented consecutively in random order. In the lexical tone discrimination task, two Mandarin words with the same or different lexical tones produced by the same speaker were presented consecutively, and participants indicated whether the lexical tone of the second word was the same as or different from the lexical tone of the first word (see Figure 2). Each task included both repeated and non-repeated words, with task order counterbalanced across participants.

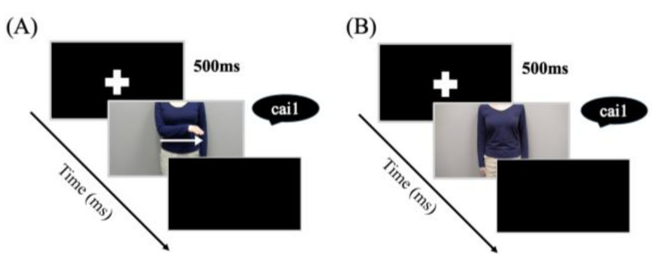

Figure 3: Screenshots of (A) Pitch Gesture and (B) No Gesture Learning Trials

Following the pre-test, participants completed the learning phase, in which they watched videos of an actor producing or not producing pitch gestures while hearing repeated words with corresponding lexical tones (see Figure 3). The learning phase consisted of three blocks, each containing 28 trials (total = 84 trials), which each lexical item presented three times across the learning phase. Following each block, brief interim quizzes with feedback containing both classification and discrimination trials were presented to maximize learning and engagement. The learning phase lasted approximately 15-20 minutes.

Once the learning phase was completed, participants moved on to the immediate post-test, which was identical to the pre-test except that non-repeated words from a different list were presented. Participants completed the follow-up session approximately 24 hours later. Before completing the delayed post-test, participants completed the St. Mary’s Hospital Sleep Questionnaire (Ellis et al., 1981), a self-report questionnaire measuring sleep quality. Because sleep was assessed via self-report, sleep analyses reflect individual differences in sleep quality rather than experimentally manipulated sleep effects. The delayed post-test was identical to the pre-test and immediate post-test, with non-repeated words drawn from the remaining list.

Sleep and Music Composite Score Calculation

For the St. Mary’s Hospital Sleep Questionnaire, we used a scoring method implemented by Karkare et al. (2013). This method involved summing Likert-style items asking about sleep depth, instances of waking up, difficulty falling asleep, and sleep satisfaction to calculate a sleep quality composite score. Scores ranged from 6 to 38, with lower scores indicating worse sleep quality.

We measured music experience by generating a music composite score based on a five-item self-report derived from the demographics survey. Items assessed music reading ability, instrumental experience, voice lessons, self-rated music-making ability, and average time spent per day making music. These metrics contribute to individual differences in musical engagement and training modulating sensitivity to novel phonological contrasts and facilitating L2 learning (Shook et al., 2013). Three items were coded dichotomously (0/1), and the remaining two were recoded to a 0-1 scale based on categorical responses. We summed the five-item scores for each participant, resulting in composite scores ranging from 0 to 5, with higher scores indicating greater self-reported music making experience.

Data Cleaning and Variable Characterization

Discrimination sensitivity was measured using d′ (d-prime), a signal detection measure that accounts for response bias by incorporating both hit rates and false alarm rates, providing a more accurate estimate of perceptual sensitivity than raw accuracy. Latencies for incorrect responses were excluded, and outliers three standard deviations beyond the participant-specific mean were trimmed from response latencies and sleep composite scores. All continuous variables included as fixed factors in models were grand mean centered, and categorical variables included as fixed factors in models were coded using weighted mean centered (Helmert) contrast coding, with the level mentioned first for each factor coded as most negative and the level mentioned last coded as most positive. These variables included age (continuous; 18 – 36), L2 exposure duration (continuous; 0 – 16), music making composite score (continuous; 0 – 4.1), sleep quality composite score (continuous; 7 – 39), test (categorical; pre-test, immediate post-test, delayed post-test), learning condition (categorical; gesture, no gesture), word type (categorical; repeated, non-repeated), and tone match for discrimination trials (categorical; different, same).

Data Analysis

All data were analyzed in R via the lme4 package (Bates et al., 2015), with *p*-values obtained using the lmerTest package (Kuznetsova et al., 2017) and estimated marginal means for interactions reaching significance obtained using the emmeans package (Lenth, 2019), with Tukey-Kramer corrections applied. We implemented a probit mixed effect model including tone match and its interactions with other fixed effects of interest to analyze lexical tone discrimination, a logit mixed effect model to analyze lexical tone classification accuracy, and linear mixed effect models to analyze lexical tone discrimination and classification latency. All models included random intercepts for participants and items.

Results

Preliminary Analyses

Preliminary analyses identified age as a significant predictor for discrimination sensitivity (B=-0.03, SE=0.01, z=-3.17, p=.002); consequently, it was included as a covariate in the model for discrimination sensitivity. Music experience and L2 duration were non-significant predictors in all models and were therefore excluded from them.

Lexical Tone Discrimination

We observed an interaction between tone match and sleep composite, indicating that discrimination sensitivity for matching and mismatching lexical tones increased in tandem with sleep quality composite score (B= 0.10, SE=0.03, t=3.58, p<.001). This interaction was qualified by a further interaction with test and age (B=-0.22, SE=0.10, t=-2.19, p=.03). Estimated marginal means indicated a non-significant trend for lexical tone discrimination sensitivity to

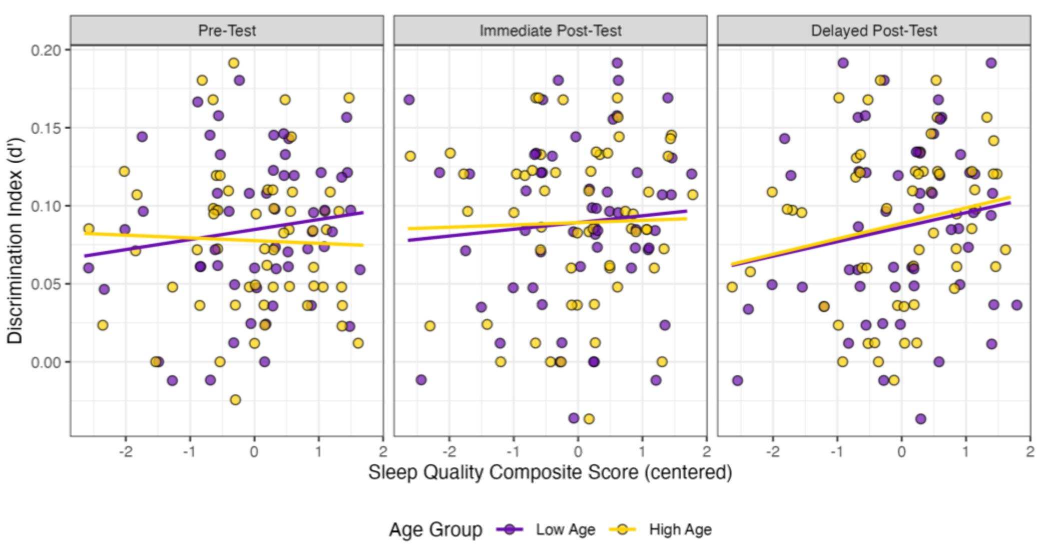

Figure 4: Lexical Tone Discrimination Index (*d'*) by Test, Learning Condition, and Sleep Quality Composite Score.

increase in accordance with sleep quality composite score without modulation by age in the delayed post-test (*B*= 0.44, *SE*=0.20, *t*=2.19, *p*=.07), but not the immediate post-test (*B*= 0.24, *SE*=0.20, *t*=1.20, *p*=.45) or the pre-test (*B*=0.04, *SE*=0.20, *t*=0.18, *p*= .98; see Figure 4).

For latency, we observed a non-significant trend towards a main effect of sleep, with latency decreasing as sleep quality composite score increased (*B*=-126.53, *SE*=66.04, *t*=-1.92, *p*=.06). This effect was qualified by a non-significant trend towards an interaction with test and word type (*B*=134.46, *SE*=72.19, *t*=-1.86, *p*=.06). This finding suggests that, in the delayed post-test, sleep may decrease lexical tone discrimination latency for repeated relative to non-repeated words.

Lexical Tone Classification

For accuracy, we observed an interaction between sleep, test, and learning condition (*B*=0.28, *SE*=0.11, *t*=2.55, *p*=.01). Estimated marginal means indicated that lexical tone classification accuracy increased in accordance with sleep quality composite score to a greater degree in the no gesture than the gesture condition in the delayed post-test compared to the pre-test (*B*=-0.17, *SE*=0.06, *t* =-2.87, *p*=.006) and the immediate post-test (*B*=-0.20, *SE*=0.06, *t*=-3.43, *p*=.002), but not the immediate post-test compared to the pre-test (*B*=0.33, *SE*=0.06, *t*=0.57, *p*=.57; see Figure 5).

For latency, we observed an interaction between sleep and test (*B*=-195.08, *SE*=69.27, *t*=-2.82, *p*=.01). Estimated marginal means indicated that lexical tone classification latency decreased in accordance with sleep quality composite score to a greater degree in the immediate post-test compared to the pre-test (*B*=-413.1, *SE*=145, *t*=-2.84, *p*=.007), as well as in the delayed post-test compared to the pre-test (*B*=-474.0, *SE*=146.0, *t*=-3.24, *p*=.004), but not in the delayed post-test compared to the immediate post-test (*B*=-60.9, *SE*=141.0, *t*=-0.43, *p*=.67). This interaction was qualified by a further interaction with learning condition (*B*=-412.70, *SE*=123.97, *t*=-3.33, *p*<.001). Estimated marginal means indicated that lexical tone classification latency decreased in accordance with sleep quality

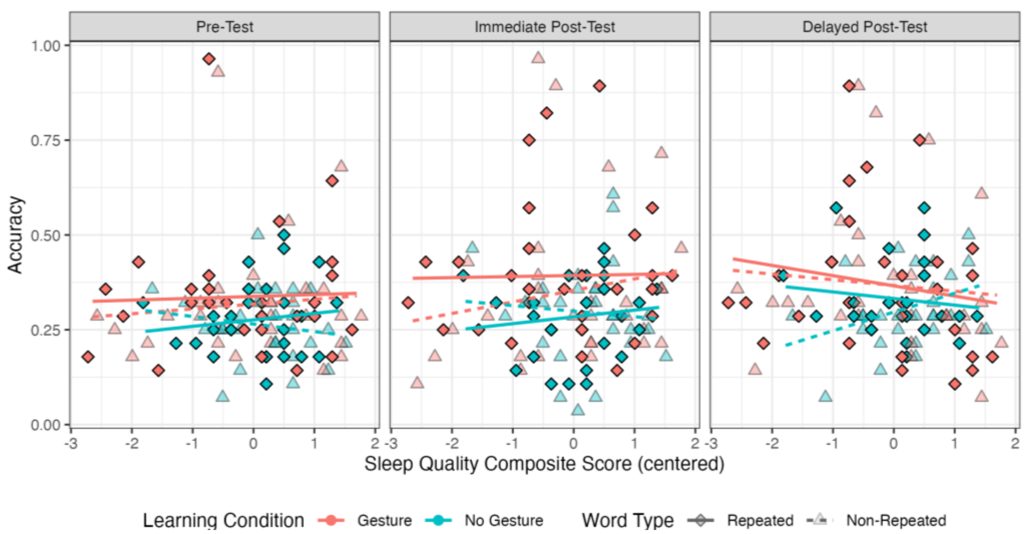

Figure 5: Lexical Tone Classification Accuracy by Test, Learning Condition, Word Type, and Sleep Quality Composite Score.

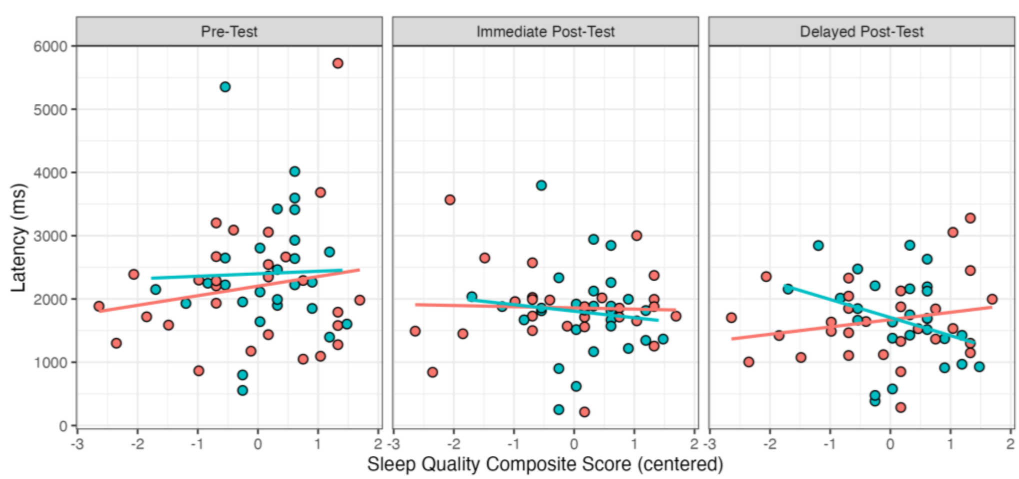

Figure 6: Lexical Tone Latency by Test, Learning Condition, and Sleep Quality Composite Score.

composite score to a greater degree in the no gesture than the gesture condition in the delayed post-test compared to the immediate post-test (B=195.5, SE=67.5, t=2.90, p=.01), but not in the delayed post-test compared to the pre-test (B=108.7, SE=69.5, t=1.56, p=.18) or the immediate post-test compared to the pre-test (B=-86.8, SE=65.9, t=-1.32, p=.19; see Figure 6).

Discussion

The present findings reveal that higher sleep quality is associated with improved delayed learning of L2 lexical tones, particularly when tones are learned without observing pitch gestures. When pitch gestures are observed, the benefit of sleep is reduced. This pattern contrasts with prior findings in tonal L1 speakers and extends evidence for sleep-related benefits to atonal L1 speakers (Qin & Zhang, 2019; Qin et al., 2025). One possible explanation for the reduced sleep-related benefit in the gesture condition is that gestures may create different memory traces. In addition to encoding auditory information, observing pitch gestures may recruit motoric and visuospatial representations, resulting in a more multimodal trace (e.g., Barsalou, 2008; Goldin-Meadow, 2005). This multimodal encoding may alter consolidation processes during sleep. Alternatively, the presence of multiple representational formats may introduce competition or interference during consolidation, particularly if learners attend to gesture features that are not fully aligned with the relevant phonological distinctions. This account may help explain why sleep-related gains were attenuated when gestures were present. Music experience did not significantly predict performance in any model, aligning with mixed prior findings where musical training effects on L2 tone perception are not consistently observed.

With respect to lexical tone discrimination, there was a marginal association with sleep quality in the delayed post-test after accounting for age in the model, but not the pre-test or immediate post-test, consistent with a possible delayed relationship between sleep quality and discrimination performance within the present sample of young adults. This finding diverges from previous evidence that sleep duration does not benefit tonal L1 speakers' discrimination sensitivity for L2 lexical tones (Qin & Zhang, 2019). With respect to lexical tone classification, the beneficial effect of sleep quality on accuracy and latency in the delayed post-test was mitigated by learning L2 lexical tones by observing pitch gestures. This finding contradicted our prediction that sleep would enhance the beneficial effect of observing pitch gestures on tonal L1 speakers' classification of L2 lexical tones in the delayed post-test (Baills et al., 2019; Hannah et al., 2017; Morett et al., 2022; Zhen et al., 2019).

Moreover, this interaction between gesture-based training and sleep diverges from prior L2 lexical tone learning research demonstrating robust immediate benefits, suggesting that while sleep may support generalization of newly learned tonal categories, the beneficial effect of sleep may be influenced by the structure of encoding during embodied learning cues. Consistent with prior work on multimodal category learning (Ashton et al., 2018), learning L2 lexical tones by observing pitch gestures did not produce immediate benefits. However, this previous work showed that benefits of multimodal learning can emerge across multiple sessions. Because the present study only included one learning session and a 24-hour delayed post-test, future research should investigate whether benefits emerge over multiple sessions and delayed post-tests. Given the correlational nature of self-reported sleep, these findings cannot be interpreted causally and may reflect other factors such as fatigue or general cognitive variability. Overall, the results provide evidence that sleep quality exerts a more robust effect on L2 lexical tones learned without than with pitch gestures. Future research should address limitations of the present study, including the use of self-report sleep measures and one talker of each sex, to further clarify these dynamics.

In conclusion, the current study provides initial evidence of a positive association between sleep quality and subsequent L2 lexical tone classification and discrimination by atonal L1 speakers. These findings reveal how sleep is associated with L2 lexical tone learning both independently and in conjunction with pitch gestures, providing insight into the impact of overnight sleep consolidation on classification and discrimination of newly learned L2 lexical tones by atonal L1 speakers.

Acknowledgments

Funded by NSF Faculty Early Career Development Program Award #2446895

References

Ashton, J. E., Jefferies, E., & Gaskell, M. G. (2018). A role for consolidation in cross-modal category learning. *Neuropsychologia*, 108, 50-60.

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition, 41*(1), 33–58.

Barsalou, L.W. (2008) Grounded cognition. *Annu Rev Psychol.* 59:617-45.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1).

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). Do we think about music in terms of space? Metaphoric representation of musical pitch. *Proceedings of the Annual Meeting of the Cognitive Science Society, 25*.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behaviouralevidence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3773–3800.

Earle, F. S., Landi, N., & Myers, E. B. (2017). Sleep duration predicts behavioral and neural differences in adult speech sound learning. *Neuroscience Letters, 636*, 77–82.

Earle, F. S., & Myers, E. B. (2015a). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1680–1695.

Earle, F. S., & Myers, E. B. (2015b). 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.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953.

Ellis, B. W., Johns, M. W., Lancaster, R., Raptopoulos, P., Angelopoulos, N., & Priest, R. G. (1981). The St. Mary's Hospital sleep questionnaire: A study of reliability. *Sleep, 4*(1), 93-97.

Eng, K., Hannah, B., Leong, L., & Wang, Y. (2013). *Can co-speech hand gestures facilitate learning of non-native tones?* 060225–060225.

Farran, B. M., & Morett, L. M. (2024). Multimodal cues in L2 lexical tone acquisition: Current research and future directions. *Frontiers in Education, 9*, 1410795.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*(6958), 614–616. https://doi.org/10.1038/nature01951

Fischer, S., Drosopoulos, S., Tsen, J., & Born, J. (2006). Implicit Learning–Explicit Knowing: A Role for Sleep in Memory System Interaction. *Journal of Cognitive Neuroscience, 18*(3), 311–319.

Flege, J. E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics, 16*(4), 425–442.

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics, 32*(3), 395–421.

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal Association between Auditory and Visuospatial Information in Mandarin Tone Perception in Noise by Native and Non-native Perceivers. *Frontiers in Psychology, 8*, 2051.

Karkare, K., Sinha, S., Taly, A. B., & Rao, S. (2013). Prevalence and profile of sleep disturbances in Guillain-Barre Syndrome: A prospective questionnaire-based study during 10 days of hospitalization. *Acta Neurologica Scandinavica, 127*(2), 116–123.

Kiriloff, C. (1969). On the Auditory Perception of Tones in Mandarin. *Phonetica, 20*(2–4), 63–67.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13).

Lenth, R. V. (2019). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.11.1). CRAN.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a Tonal Language by Attending to the Tone: An In Vivo Experiment. *Language Learning, 61*(4), 1119–1141.

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.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457.

Morett, L. M., & Chang, L.-Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*(3), 347–353.

Morett, L. M., Feiler, J. B., & Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning. *Cognition, 222*, 105014.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376.

Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLOS ONE, 14*(12), e0221498.

Qin, Q. Z., Wu, R., & Zhang, C. (2025). Daytime naps consolidate Cantonese tone learning through promoting cross-talker perception: The role of prior knowledge. Brain and language, 265, 105568.

Shook, A., Marian, V., Bartolotti, J., & Schroeder, S. R. (2013). Musical experience influences statistical learning of a novel language. *The American Journal of Psychology, 126*(1), 95–104.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113*(2), 1033–1043.

Wang, Y., Sereno, J. A., & Jongman, A. (2006). L2 acquisition and processing of Mandarin tones. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (1st ed., pp. 250–256). Cambridge University Press.

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.

Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics, 36*(2), 250-267.

Wei, Y., Jia, L., Gao, F., & Wang, J. (2022). Visual–Auditory Integration and High-Variability Speech Can Facilitate Mandarin Chinese Tone Identification. *Journal* *of Speech, Language, and Hearing Research, 65*(11), 4096–4111.

Wong, P.C., Skoe, E., Russo, N.M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat Neurosci*. 10(4):420-2. doi: 10.1038/nn1872.

Yip, M. (2002). *Tone* (1st ed.). Cambridge University Press.

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition, 187*, 178–187.