Sleep affects higher-level categorization of speech sounds, but not frequency encoding

Aurélien de la Chapelle, Marie-Anick Savard, Reyan Restani, Pouya Ghaemmaghami, Noam Thillou, Khashayar Zardoui, Bharath Chandrasekaran, Emily B.J. Coffey

SLEEP AFFECTS HIGHER-LEVEL CATEGORIZATION OF SPEECH SOUNDS, BUT NOT FREQUENCY ENCODING

A PREPRINT

Aurélien de la Chapelle
Lyon Neuroscience Research Centre
Lyon, France

Marie-Anick Savard
Department of Psychology
Concordia University
Montreal, QC, Canada

Reyan Restani
Université Paris Nanterre
Paris, France

Pouya Ghaemmaghami
Department of Psychology
Concordia University
Montreal, QC, Canada

Noam Thillou
Department of Psychology
Concordia University
Montreal, QC, Canada

Khashayar Zardoui
Department of Psychology
Concordia University
Montreal, QC, Canada

Bharath Chandrasekaran
Department of Communication Science and Disorders
University of Pittsburgh
Pittsburgh, USA

Emily B. J. Coffey
Department of Psychology
Concordia University
Montreal, QC, Canada
emily.coffey@concordia.ca

May 23, 2023

ABSTRACT

Sleep can increase consolidation of new knowledge and skills. It is less clear whether sleep plays a role in other aspects of experience-dependent neuroplasticity, which underlie important human capabilities such as spoken language processing. Theories of sensory learning differ in their predictions; some imply rapid learning at early sensory levels, while other propose a slow, progressive timecourse such that higher-level categorical representations guide immediate, novice learning, while lower-level sensory changes do not emerge until later stages. In this study, we investigated the role of sleep across both behavioural and physiological indices of auditory neuroplasticity. Forty healthy young human adults (23 female) who did not speak a tonal language participated in the study. They learned to categorize non-native Mandarin lexical tones using a sound-to-category training paradigm, and were then randomly assigned to a Nap or Wake condition. Polysomnographic data were recorded to quantify sleep during a 3 hour afternoon nap opportunity, or equivalent period of quiet wakeful activity. Measures of behavioural performance accuracy revealed a significant improvement in learning the sound-to-category training paradigm between Nap and Wake groups. Conversely, a neural index of fine sound encoding fidelity of speech sounds known as the frequency-following response (FFR) suggested no change due to sleep, and a null model was supported, using Bayesian statistics. Together, these results support theories that propose a slow, progressive and hierarchical timecourse for sensory learning. Sleep's effect may play the biggest role in the higher-level learning, although contributions to more protracted processes of plasticity that exceed the study duration cannot be ruled out.

***K*eywords** categorical learning · frequency-following response · sleep · consolidation · speech

1 Introduction

Sleep is known to play an active role in consolidating several types of learning, leading to performance gains (or preservation) as compared with equivalent periods of wakeful rest [Rasch and Born, 2013]. There is considerable

evidence from human training studies of a benefit of sleep on declarative memory tasks, which involve remembering facts or events and rely on medial temporal lobe structures including the hippocampus. The benefits of sleep on some aspects of non-declarative memories, which generally do not rely on hippocampus, are also well established. For example, some forms of conditioning, implicit learning, and motor control are consolidated in sleep [Morin et al., 2008, Paller and Voss, 2004, Rasch and Born, 2013]. In contrast, the benefit that sleep might confer upon low-level perceptual or sensory learning (i.e., long-lasting performance improvements on perceptual skills) is less clear. These fundamental processes affect how effectively we observe and interact with our environment. Training can tune and re-weight sensory input [Jia et al., 2020], which increases automaticity [Earle and Myers, 2014] and reduces the need for high-level processing. Mechanisms of sensory plasticity have clinical interest, as they can be harnessed to promote better cognitive skills, for example to counteract age-related deficits in sound encoding [Anderson et al., 2013]. A clearer understanding of sleep's involvement in sensory plasticity will complement our knowledge of higher-level consolidation processes, and may have implications for training-based interventions in auditory impairment.

Working in the visual modality, Klinzing et al. recently showed that sleep does not aid binocular disparity-based learning [Klinzing et al., 2020, 2021], which is a low-level perceptual task involving depth assessment using small differences between images received by each eye. The authors argued against an effect of sleep on low-level perceptual neuroplasticity more generally, on the basis that the majority of evidence in its favour had relied on a single learning paradigm (the texture discrimination task; TDT; [Gais et al., 2000, Deliens et al., 2014, Karni et al., 1994, Stickgold et al., 2000, Tamaki et al., 2020]). In the TDT, each trial consists of a visually-presented 'texture' in which a target set of symbols, such as three diagonal lines, is embedded within a grid of lines with different orientations. The stimulus is briefly presented, followed by a mask, after which the subjects report the target's orientation. However, because difficulty is manipulated by changing the stimulus-to-mask onset asynchrony [Gais et al., 2000], better task performance could plausibly be driven by improvements in higher-level cognitive processes, such as learning to better allocate attentional resources, rather than plasticity of the perceptual component itself [Klinzing et al., 2020, 2021].

In the auditory modality, no single paradigm has enjoyed such focused attention. A handful of studies have looked at the effect of sleep on relatively low-level auditory tasks, finding mixed evidence. For example, Gaab et al. asked participants to make match/mismatch judgements of the first and last tones of a short sequence containing intervening distractors [Gaab et al., 2004]. Delayed performance gains occurred only following periods of sleep, regardless of whether sleep occurred immediately, or 12 hours after training. Conversely, Atienza et al. found no performance difference (after 48-72 hrs) between a sleep-deprived and a control group following a task in which participants listened to patterns of eight tones and were asked to identify infrequently occurring deviants of the sixth tone [Atienza et al., 2004, 2005]. Gottselig et al. used a paradigm in which participants learned associations between auditory tones and visual symbols. The authors showed that sleep does not benefit learning sequences of tone-symbol pairs above and beyond an equivalent period of quiet rest (although a trend was suggestive) [Gottselig et al., 2004]. Basic linguistic stimuli, instead of pure tones, have also been studied. Alain et al. observed no differences related to sleep beyond those accounted for by the passage of time after participants learned to identify familiar spoken vowels [Alain et al., 2015]. Roth et al. trained participants to identify consonant–vowel stimuli in increasing levels of background noise [Roth et al., 2005] and, although performance improved following delays of 12 h, there was no additional benefit of sleep during the interval. Fenn et al. studied the ability of participants to learn to recognize words presented in speech streams that were produced by an early text-to-speech synthesizer [Fenn et al., 2003]. The authors reported that sleep was helpful to recover previously learned auditory memories that had degraded throughout the day. Other studies have focused on how sleep might affect specific aspects of learning such as generalization across talkers [Earle and Myers, 2015a] and passive infererence from one's native language on perceptual learning of non-native speech contrasts [Earle and Myers, 2015b] (see also Earle and Myers [2014] for a review), but as regards the effect of sleep on tuning of basic perceptual processes, in brief, the picture is unclear. There is mixed evidence for behavioural gains of sleep on basic auditory learning and its observation may depend on experimental design including the stimuli used, and the delay and amount of sleep occurring between learning and testing.

Of the relatively few studies that have measured both behavioural improvements and neurophysiological indices of sound processing using electroencephalography (EEG) or magnetoencephalography (MEG), there is some evidence of sleep-dependency, particularly for later evoked responses, which are attributed to higher levels of processing. Atienza et al. found sleep-related changes in the late cortical response (P3a) of subjects who slept normally after a first exposure to a deviant detection paradigm as compared to those who were deprived of sleep. The authors suggested that sleep improved automatization of learning by improving the response timing consistency of cortical neural assemblies involved in automatic sound-change detection [Atienza et al., 2004, 2005]. Interestingly, the differences were apparent in the neurophysiological responses but not in the behavioural responses, as there was no sleep-related difference in detection sensitivity (hit/false alarm) nor response time. Alain et al. observed increased magnetic brain responses about 200 ms following sound presentation while participants identified spoken vowels after training and a night of sleep, as compared with an equivalent wakingo period. However, this evoked component was not directly related

to performance gains [Alain et al., 2015]. Earle et al. trained monolingual English-speakers to identify non-native speech sounds in the evening, and found a correlation between sleep duration and next-day behavioural improvement, which was also correlated with changes in ERP response magnitude (i.e., mismatch negativity amplitude between 150-200 ms) Earle et al. [2017]. To summarize, while there is some evidence for a sleep effect on both behavioural and neurophysiological indices of auditory plasticity, the emerging view suggests they do not always align or show the same timecourses. To further complicate matters, studies of sleep's effect on auditory neuroplasticity vary in their sleep design (e.g. nap [Gottselig et al., 2004], overnight [Alain et al., 2015, Fenn et al., 2003, Gaab et al., 2004], sleep deprivation [Atienza et al., 2004, 2005]), in addition to the auditory tasks used [Earle and Myers, 2014].

To clarify how sleep affects auditory perceptual learning both behaviourally and physiologically, we used categorical perception boundary learning as a model to study both behavioural and neurophysiological plasticity, as a function of sleep. To our knowledge, this is the first such study that uses a controlled sleep design (involving participants learning non-native speech sound categories), and in which a range of behavioural performance and electroencephalographic measures are used to assess patterns of representation and change caused by training.

Perception of some human speech phonemes including tones is categorical in nature, an arrangement that makes language perception robust to variability in sound production and listening conditions. Categorical boundaries can be language-specific [Xu et al., 2006a], and develop naturally upon language acquisition in childhood, although they can also be learned in adulthood with training [Zhang et al., 2009, Wong and Perrachione, 2007, Reetzke et al., 2018]. In tonal languages, pitch dynamics carry semantic information. For example, a phoneme spoken with a steady pitch has a different meaning than an otherwise similar phoneme spoken with a rising pitch. Mandarin speakers therefore develop a categorical boundary between level and rising tones, whereas non-tonal language speakers do not.

Because of its behavioural relevance, tonal language speakers must encode pitch information with high fidelity. Neural correlates of these representations can be non-invasively recorded using an evoked auditory response measured with EEG called the frequency-following response (FFR; for reviews, see [Coffey et al., 2019, Skoe and Kraus, 2010, Krizman and Kraus, 2019]). The FFRs of native tonal language speakers show enhanced pitch encoding and more accurate pitch tracking than do those of non-speakers [Krishnan et al., 2005, Swaminathan et al., 2008], demonstrating that neural encoding of pitch in the auditory system is shaped by long-term experience with language (in a parallel fashion, musicians also show encoding benefits [Musacchia et al., 2007, Wong et al., 2007]). Shorter periods of training in non-tonal speakers that span days or weeks can enhance FFRs (e.g. Song et al. [2008], Reetzke et al. [2018], Carcagno and Plack [2011]), which are also somewhat amenable to attentional processes (e.g. [Hartmann and Weisz, 2019, Coffey et al., 2016]).

Categorical perception boundary learning in tonal languages therefore provides an optimal window to study behavioural learning and neuroplasticity simultaneously, because the quality of encoding of pitch dynamics can be non-invasively recorded. Training paradigms in which participants listen to tones and receive feedback for classifying them can be used to study both the development of (perceptual) categorical boundaries between tones, and changes in low-level sensory encoding of the newly acquired speech sound patterns [Reetzke et al., 2018, Xie et al., 2017].

Reetzke et al. recently studied the evolution of Mandarin lexical tone categorical boundaries and FFRs over several weeks of daily training using a naturally-produced speech tokens [Reetzke et al., 2018]. Performance on categorization accuracy showed a typical learning curve, with rapid increases during the first training sessions followed by more gradual improvements for the remaining ones. A separate measure known as the perceptual identification task, in which participants classify seven synthetic stimuli that vary parametrically between two of the categories (level, rising), was used to measure the developing categorical boundary more precisely (note that no feedback was given, such that learning is driven by the main training task). Within a few days' training, novices exhibited clear signs of a perceptual boundary between level and rising tones (evidenced in the slope of the responses to the parametrically varying stimuli and a peak in reaction times at the boundary). Reetzke et al. report sensory tuning as measured by FFR strength, but only at longer training periods (7 days). The authors highlight the differences in timecourses of plasticity, and propose that sensory plasticity accompanies rather than drives non-native speech learning. These results are in accordance with theories that posit faster plasticity in higher-level representations and slower timecourses for earlier, lower-level changes [Ahissar et al., 2009, Kraus, 2021].

Notably, in Reetzke et al., a large jump in categorization performance on the training task itself was observed between the first and second days, suggesting a possible effect of sleep on this task (see Figure 2B in [Reetzke et al., 2018]). Because all participants slept between sessions, it is not possible to determine whether sleep played a causal role in the performance gains between sessions, or if apparent gains can be explained by the process of averaging each session's trials during the most rapidly changing portion of the learning curve. To investigate the role of sleep in perceptual categorical learning, we used a nap design to systematically assess the effect of sleep on both behavioural improvement and measures of neurological encoding of pitch information. Forty healthy adults who did not speak a tonal language participated in the study. They learned to categorize Mandarin lexical tones using the sound-to-category

training paradigm used in previous work (e.g. [Xie et al., 2017]), and were then randomly assigned to a Nap or Wake condition. Polysomnographic data were recorded to quantify sleep depth and duration during a three hour afternoon nap opportunity, or equivalent period of quiet wakeful activity. As well as completing the training task, participants did a perceptual identification task and FFR recording before and after the Nap or Wake period.

To maximize the possibility of observing changes on all measures, we increased the number of training trials in the session from the number used in previous studies [Reetzke et al., 2018], evaluated FFR in both passive and active listening conditions, and used a sensitive machine-learning based analysis as well as more traditional measures (e.g. stimulus-to-response correlation, pitch tracking) to quantify FFR fidelity. We hypothesized that subjects who have a nap will show an increase in task accuracy and measures of speech categorical learning as compared with those who do not sleep, but that during early stages of categorical learning, an effect of sleep will not be found in the neurophysiological measures of sound representation (a null hypothesis that can be evaluated using Bayesian statistics [Wetzels et al., 2009]). Finding a statistically significant effect of sleep on categorization in the training paradigm, we further evaluated whether the duration of NREM sleep (stages 2 and 3) was related to the degree of improvement on the task. The results suggest an effect of a short period of sleep on categorization performance, the higher-level behavioural measure; and no evidence of a sleep effect on the FFR, the low-level physiological measure. Together, these results add to our understanding of the nature of sleep effects in perceptual learning.

2 Methods

2.1 Participants

Forty right-handed, human adults with no tonal language experience participated in this experiment (in-line with previous work on similar topics that found related effects of interest, e.g., [Earle et al., 2017, Reetzke et al., 2018]). Participants were randomly assigned to a Nap group or control group (Wake). One Wake group subject fell asleep for more than an hour and was moved to the Nap group during the experiment (i.e., prior to data observation); control analyses confirmed that the subject's subjective fatigue rating was not unusually high as compared with both groups' mean and that they otherwise appeared to comply with instructions. In total there were $N = 21$ participants in the Nap group (10 female; mean age 22.0; *SD* 2.7), and N = 19 participants in the Wake group ($N = 19$; 13 female; mean age 23.8; $SD = 3.9$).

All subjects reported being in good neurological health with normal hearing, were non-smokers, and were free of any diagnosed psychiatric disorder or sleep disorder. They did not take medication that is known to affect sleep, had not changed time zones or conducted shift work in the 6 weeks preceding the experiment, and had a normal 6-10 hour sleeping pattern in the three days prior to the experiment (confirmed by a self-reported sleep log). The day before and day of the experiment they were asked to refrain from consuming caffeine, alcohol, nicotine, and cannabis. Subjects gave written informed consent, and were compensated for their time with money or course credit. The experimental protocol was approved by Concordia University's Human Research Ethics Committee. In accordance with Journal guidelines, we report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. The study was not pre-registered prior to the research being conducted. Data and code are available online `https://osf.io/p46ga/`.

2.2 Study design

The study design is illustrated in Figure 1A. To increase sleep pressure and facilitate afternoon napping, participants were asked to sleep one hour less than their habitual sleep duration on the night before the experimental session. After completing several questionnaires to document their health, language, and musical experience, subjects were randomly assigned to either the Nap or Wake group and invited to change into comfortable sleeping clothes.

Subjects' frequency following responses (FFRs) were recorded to two rising tones presented via insert earphones (ER-3A, Etymotic Research, Elk Grove Village, IL). Subjects completed a Perceptual Identification (PI) task (during which time FFR was also recorded) and the Training Paradigm (Figure 1B and C), then a second FFR recording was made to establish whether short-term plasticity had occurred. Subjects in the Nap group had a three hour nap opportunity, in which they were asked to lie in bed in a darkened room. While many nap studies involving active manipulation of brain processes in sleep frequently limit nap duration to about 1.5 hours (e.g. Antony et al. [2018]), here we allowed a longer nap opportunity to maximize the possibility that all participants slept well. Participants were instructed to lie still and rest if they were not able to sleep further. Subjects in the Wake group stayed seated with the lights on, and were encouraged to do a quiet activity such as reading, studying, or watching a silent nature

documentary (Yellowstone: Battle for Life, BBC, 2009). Use of telephones, music, and computers was not permitted so as to minimize potential interference with offline processing in the auditory system.

Polysomnography (PSG) was recorded during the break for both groups for sleep analysis in the Nap group, and to confirm that the Wake group subjects did not sleep. After the break, participants were gently awakened if necessary and asked to get up, and after about 10 minutes all subjects were asked if they felt alert and ready to continue before proceeding. There followed a third passive listening FFR recording, and then subjects completed the behavioural tasks a second time to evaluate effects of sleep vs. the passage of time.

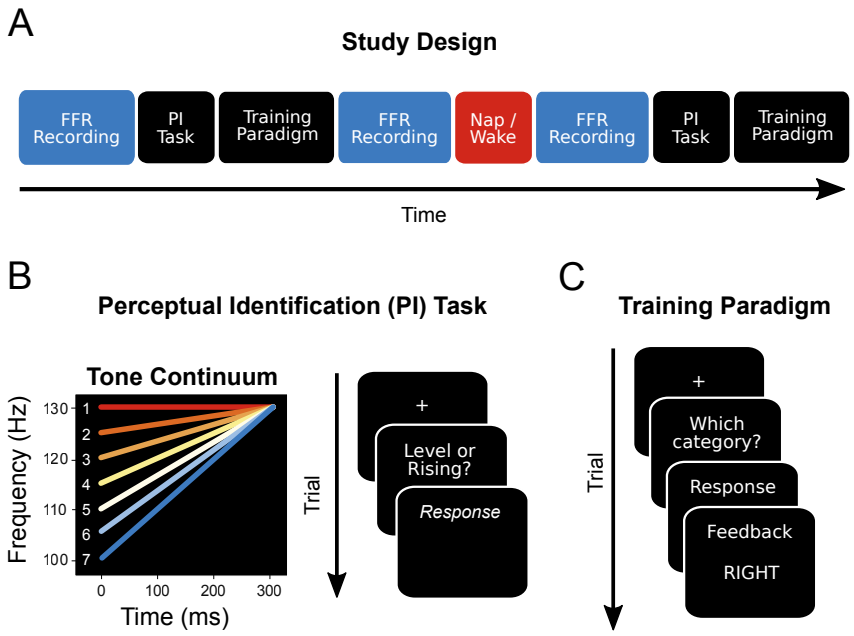


Figure 1: (A) Study design. (B) Perceptual Identification (PI) task, which is used to measure the development of a categorical boundary. The seven-step tone continuum from the level to the rising Mandarin lexical tone (left) and the trial procedure (right; task and illustration adapted from Reetzke et al. [2018]). (C) Training paradigm for the sound-to-category task, which is used to induce the development of categorical boundaries. Each trial began with a 750 ms fixation cross, followed by presentation of a Mandarin lexical tone (440 ms), after which participants had unlimited time to categorize the tone by button press (1, 2, 3, or 4). Feedback was presented for 1000 ms, 500 ms following the participants' response, as in Reetzke et al. [2018].

Quantification of sleep during nap or rest period

Polysomnography (PSG) data were recorded to document the length and depth of sleep during the break. EEG was recorded from the scalp at C3 and C4 (10–20 International System) referenced to the mastoids, electrooculography (EOG) was used to record eye movements, and pairs of electrodes on the chin and neck were used for electromyography (EMG) to record muscle tension (sampling frequency: 1024 Hz). PSG was recorded in the Wake group as well as the Nap group, to confirm that Wake subjects did not fall asleep.

Sleep scoring, in which 30 s windows of data are visually inspected and categorized into wake, non-REM sleep stages 1-3, and REM sleep, was accomplished according to standard AASM practices [Iber et al., 2007] based on the C3 and C4 EEG channels, EOG, and EMG, by a researcher trained in sleep scoring using the 'fMRI Artefact rejection and Sleep Scoring Toolbox' (FAAST; Cyclotron Research Centre, Liège). Total time in wake and each sleep stage was quantified for each subject.

2.3 Questionnaires

To confirm that subjects had normal sleep patterns prior to the experiment, they were asked to fill out a sleep diary three days prior to the experiment which included information about the time they went to bed and woke up. The day of the experiment, the participants filled out the Munich Chronotype Questionnaire (MCTQ; [Roenneberg et al., 2003]);

abnormal chronotypes would have been grounds for exclusion. Participant's fatigue level was measured the day of the experiment using seven questions, five point self-reported Subjective Fatigue Scale questionnaire; abnormally high levels of fatigue (e.g. $>$ 2 SD from the group mean) would have resulted in exclusion.

Both musicianship and multilingualism are correlated with better auditory perceptual skills [Turker and Reiterer, 2021], particularly the neurophysiologial encoding of pitch information [Besson et al., 2007, Musacchia et al., 2007, Bidelman et al., 2011] including speech sounds [Marques et al., 2007, Bidelman et al., 2014]. To better characterize our sample's experience with sound, we used the Montreal Music History Questionnaire (MMHQ)[Coffey et al., 2011]. We quantified total cumulative training and practice hours amongst people reporting musical experience, age of start of formal training, and the number of languages that participants reported being able to speak fluently. The MMHQ also includes questions about demographics, general health, and language experience, which were used to confirm eligibility.

2.4 Sound-to-Category Training Task

A Sound-to-Category Training Task [Xie et al., 2017, Reetzke et al., 2018] was performed twice during the experiment, before and after the break in which the Nap group slept (see Figure 1A). This task serves both to cause and track changes in overall categorization performance. Auditory stimuli were presented via the insert earphones at 70 dBA SPL. The sound-to-category training paradigm included two training features that direct the participant's attention to the relevant acoustic cues: stimuli with high talker variability, and trial by trial feedback (i.e., without prior instruction as to the nature of the categories). These features facilitate categorical learning and retention [Reetzke et al., 2018, Lively et al., 1993, Roelfsema et al., 2010].

The training consisted of presenting auditory stimuli produced by two native Chinese speakers (1 female, 1 male), which comprised five different monosyllabic words (bu, di, lu, ma and mi) and four distinct pitch categories, presented in random order. The four pitch categories were: 1: high-level, meaning the onset frequency was high and remained stable; 2: low-rising, indicating a low starting frequency that dynamically changed to a higher one; 3: low-dipping, meaning a tone that started with a relatively low frequency, dipped to an even lower level and then returned to the starting range; and 4: high-falling, in which the tone's onset frequency was high and then shifted downwards. Each speaker produced 20 tokens (5 syllables x 4 tones), for a total of 40 tokens. These auditory stimuli were the same as the ones used in previous experiments (e.g. [Reetzke et al., 2018, Xie et al., 2017, Yi and Chandrasekaran, 2016, Chandrasekaran et al., 2014]). Participants had to classify the stimuli into one of four categories without a priori knowledge of the categorical dimension, and were given feedback of "Correct" or "Incorrect" after each trial (Figure 1C). Participants completed 6 Blocks of the 40 tokens prior to the nap or rest period, and an additional 6 Blocks (Block numbers 7-12) afterwards. Accuracy was averaged across trials for each block, and the percent change in accuracy between pre-and post-break blocks (6 and 7 respectively) was calculated as the main measure of task performance change. These within-subjects normalized values remove the relatively high variability associated with overall performance differences between subjects.

2.5 Perceptual Identification Task

To assess the development of categorical perception of level to rising Mandarin lexical tone F0 contours independently from the specific learning task stimulus tokens (and thus in a more general manner), we used a perceptual identification (PI) task as has been used in previous work [Reetzke et al., 2018] (see Figure 1B). Auditory stimuli were presented via the insert earphones at 70 dB sound pressure level (SPL). Subjects were asked to make a binary categorization ('level' or 'rising') by button press to speech tokens that varied along a continuum of seven steps between level and rising F0 contours [Reetzke et al., 2018, Xu et al., 2006a]. The stimuli were synthetic (i.e., different from the more variable, natural speech tones used in the Sound-to-Category Training Task), all had the same offset frequency (130.00 Hz), and only differed in onset frequency (130.00 Hz; 125.15 Hz; 120.38 Hz; 115.70 Hz; 111.08 Hz; 106.55 Hz; 102.08 Hz). Response time was unlimited, and button presses were required to move to the next trial (after a 1000 ms delay). No feedback was provided, meaning that categorical learning is driven by the Sound-to-Category Training task. Each of the 7 stimuli was presented 20 times in randomized order via insert earphones for a total of 140 trials. Tone tokens differed only slightly acoustically from neighbouring stimuli on the continuum, but are perceived categorically by native tonal language speakers. A stronger categorical boundary is indicated by a steeper identification labeling curve (i.e. perceptual slope), and greater peak in identification reaction time (peak RT) at the category boundary, where between-category distinctions become ambiguous [Hallé et al., 2004, Reetzke et al., 2018, Xu et al., 2006a].

Identification reaction times (RTs) were calculated as mean response latency across trials, for each session. The peak RT represents the difference of between-category and within-category perceptual sensitivity [Xu et al., 2006a]. In line with previous work [Reetzke et al., 2018], trials with RTs outside of 250-3500 ms were excluded from further analysis on the basis that responses shorter than 250 ms are unlikely to have resulted from task-related cognition, and very long

response times are more likely to indicate distraction or inattention rather than to represent the duration of cognitive processing related to the task. We estimated RT peaks as the difference between mean RT at the categorical boundary (i.e., tone token 4), and the average RT for two categories near the ends of the tone continuum (tone tokens 2 and 6). For each participant, we computed within-subject differences in RT peak across sessions as input to the between-group statistical analysis.

To calculate the slope of the identification curve, we fitted a logistic regression model on the tone identification function for each session, consistent with prior work [Reetzke et al., 2018, Xu et al., 2006a, Huang et al., 2015], using the following formula:

$$y = \frac{1}{1 + e^{-b(x-c)}}$$

Where y refers to the proportion of participant responses that indicate "rising" pitch (ranging from 0 to 100%), x refers to the onset frequency of the stimulus' F0 contour, c refers to the category boundary where the proportion to report the tone as a "rising" pitch was 50%, and b refers to slope of the fitted logistic function and indicates the sharpness of the categorical boundary. In the case that a subjects' responses varied irregularly across the 7 categories, resulting in a poorly fitting curve and slope <0.25 for either or both of the sessions, their results were excluded from the group analysis. The difference in slope between the first and second measurement was computed for statistical comparison between the Wake and Nap group. The model estimation procedures were conducted in MATLAB (The MathWorks, Natick, MA) via the lsqnonlin function using the Levenberg-Marquardt algorithm [Marquardt, 1963] to search for the optimal parameters (b and c) that provides the best fit between the logistic model and the actual value y.

Electroencephalography (EEG) was recorded during the task to enable 7-way classification of FFRs to the seven tones, as an exploratory measure of sound encoding during active task performance. It is considered supplementary to the main (passive listening) FFR recording, which was explicitly designed to elicit high-quality FFRs via a much larger number of trials. Recording parameters are as described for the main FFR recordings, as described below.

2.6 Electrophysiological data collection and preprocessing for FFR analysis

The three FFR recordings were timed so as to be able to assess the effect of sleep as compared with wakeful rest on the FFR (FFR Recording 2 vs. 3), and in case of a positive result, to be able to rule out exposure to the tasks as the cause of any apparent change (FFR Recording 1 vs. 2; see Figure 1). Subjects were seated comfortably in a quiet room and watched a silent nature documentary (Yellowstone: Battle for Life, BBC, 2009). Electrophysiological responses to synthetic tonal stimuli were collected using monopolar active Ag/AgCl electrodes placed at Cz (10–20 International System [Homan et al., 1987]) and both mastoids (BioSemi, www.biosemi.com). Two ground electrodes were positioned above the right eyebrow. Active electrodes reduce nuisance voltages caused by interference currents by performing impedance transformation on the electrode; therefore, we confirmed that direct-current offset was close to zero during electrode placement instead of measuring impedance. Electrode signals were amplified with a BioSemi ActiveTwo amplifier, recorded using open filters with a sampling frequency of 2048 Hz, and stored for offline analysis using BioSemi ActiView software. An audio channel duplicating the auditory information presented to the subject was read into a separate analogue channel in the EEG recording to enable placement of stimulus onset markers with sub-millisecond precision during post-processing.

We reasoned that representation of stimuli close to a categorical boundary would be most susceptible to plasticity and could help to reveal subtle rapid changes in the FFR not observed in previous work, if they exist [Reetzke et al., 2018]. For the main FFR recordings, we selected two of the syllables used in the perceptual identification task described above (/yi2/ and /yi6/). Each has similar moderately rising frequency trajectories, but they are found on either side of the native categorical boundary between level and rising (i.e. tone four, or /yi4/). The stimulus' timeseries and autocorrelograms, which provide a visualization of autocorrelation over a 40-ms sliding window, are shown in Figure 2. Because FFRs closely reflect fine sound encoding, the same analysis can be used on the FFR data to derive a visual representation of the extent to which the brain's response captures the information about fundamental frequency (f0) and its temporal dynamics [Reetzke et al., 2018].

The stimuli were played binaurally at 80 dBA SPL +/- 2 dB. Positive and negative polarity stimuli were presented in an alternating fashion, with an inter-stimulus interval jittered between 90 and 110 ms, and they were grouped into blocks of 200 that alternated between the /yi2/ and /yi6/ stimuli. The total recording time was 28 minutes.

EEG data were re-referenced to create a Cz-to-average mastoid timeseries, which was band-pass filtered from 80 to 450 Hz (128 order Hamming windowed sinc FIR filter; EEGLAB, 2019). Data were then cut into 450 ms epochs including a 50 ms pre-stimulus baseline. For each subject, epochs were sorted by peak absolute value of the amplitude and the highest ten percent were excluded from further analysis, to remove the majority of artifacts contaminated by muscle

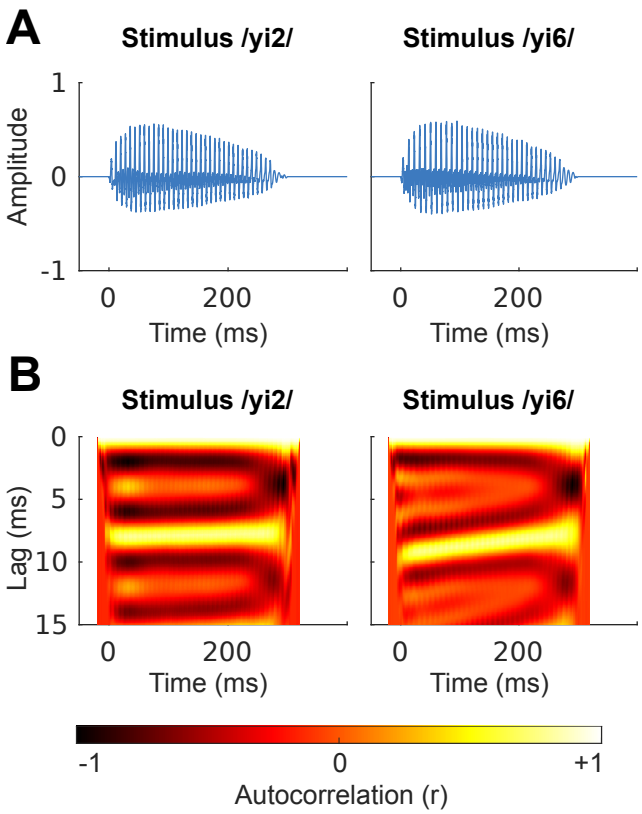


Figure 2: Stimuli used in the frequency-following response (FFR) recordings, which correspond to tones 2 and 6 in the Perceptual Identification (PI) task. Both tones are 300 ms in length and have a pitch rising to 130 Hz, but the tones fall on either side of the Mandarin lexical level-to-rising categorical boundary. (A) Waveforms of the two /yi/ tones appear only subtly different in the time domain. (B) Autocorrelograms of the same stimuli reveal differences in the stimulus' pitch dynamics. Colours represent the strength of correlation, ranging from low (dark red) to high (white) at time lags corresponding to the period of the stimulus' fundamental frequency.

movements while resulting in the same number of retained epochs per subject (i.e. 1800 per stimulus). Averages by subject were created for further analysis.

EEG data was recorded during the PI task with the same montage and parameters, and were preprocessed as described above. Because the PI task had only 20 repetitions of each of the 7 stimuli, no data were discarded. A 7-way SVM-based classifier was used to explore the recordings (described below). Because sufficient signal-to-noise ratios in FFR averages are only achieved with hundreds or even thousands of epochs [Skoe and Kraus, 2010], we did not calculate the more traditional metrics that rely on the averaged FFRs for these data.

2.7 Frequency-following response metrics

We calculated two traditional metrics from the FFR averageas for each session, stimulus-to-response correlation and fundamental frequency (f0) tracking. Stimulus-to-response correlation, which reflects the similarity between the brain's response and the evoking stimulus, was calculated by selecting the peak correlation within a range of lags up to 14 ms, which corresponds to slightly more than one period of the stimulus' lowest f0. Tracking accuracy, which reflects the degree to which FFRs follow f0 changes in dynamic frequency stimuli, was calculated by extracting the f0 contour from the FFR and from the stimulus using a short-term autocorrelation algorithm [Reetzke et al., 2018, Boersma et al., 1993]. We used a 40 ms sliding window (1 ms steps) over the entire 300 ms stimulation period, and searched for maxima within the lag range encompassing the f0 range of both stimuli (100 to 140 Hz). The mean difference between the stimulus and response's f0 contours was then calculated.

To investigate the possibility of decoding audio stimuli category from the captured EEG signals, we employed linear SVM classifiers under a stratified 10-fold cross-validation schema to decode the audio stimuli from the EEG features.

For this, pre-processed 450 ms EEG epochs are considered as features and are fed to the classifier. The within-subject classification was done for each subject and on each recording (i.e. FFR1, FFR2, FFR3) separately. In each fold, parameter "C" of the SVM classifier was tuned in a nested cross validation fashion from a predetermined range of values (0.001, 0.01, 0.1, 1, 10, 100). We computed mean classification accuracy for each subject and for each of the two time points (pre- and post- break). The main measure was the defined as the difference in mean classification accuracy across sessions.

The chance level was computed by performing permutation on the labels of the stimuli (one time per each fold, on each task and on each subject; see Supplementary Material for comparison of the classification accuracy to the null distribution). To visualize the performance of the classifiers on each condition (the two broad audio categories), we computed the confusion matrices (i.e. an N x N matrix that compares the actual audio labels with those predicted by the classifier). These confusion matrices are then summed up across all participants and all tasks and then the values of the final matrix was divided by the total number of the predicted class. Thus, the diagonal value of the final confusion matrix represents the precision of the classifier with respect to each category.

FFR analyses typically derive measures such as stimulus-to-response correlation from averaged responses of $>$2000 epochs, which is necessary due to low signal-to-noise ratio of FFRs within EEG. However, under some conditions, machine learning approaches can be used to classify even single FFR trials above chance levels [Llanos et al., 2017]. An advantage of this approach is that because fewer trials are needed, analysis can be carried out on EEG data collected during active task performance (which would be prohibitively long and fatiguing for 2000 iterations). Classification-based approaches therefore offer an additional means of assessing a variation of our main research hypothesis, namely whether the brain's response to sound is altered by sleep when being actively processed as part of a task, rather than during passive listening. The amplitude of the FFR can be subtly affected by top-down processes such as attention (e.g. [Hartmann and Weisz, 2019]) and according to task demands (e.g. [Coffey et al., 2016]). Sleep may affect only these higher processes [Klinzing et al., 2020, 2021], meaning that a sleep effect could occur in the PI FFR even if it does not occur in the main FFR recording.

A 7-class version of the classifier described above was used to decode the 7 types of level-to-rising PI stimuli from the EEG epochs. We computed mean classification accuracy across the categories for each subject and for each of the two time points (pre- and post- break). As compared with the main FFR recording, classification accuracy is expected to be much lower because of the decreased quantity of data (i.e., 10% of the number of trials per category of the 2-class analysis), the higher number of categories (7 instead of 2), and because the acoustic differences between stimuli are extremely small (i.e. adjacent tokens along the tone continuum differed only by 5 Hz in their onset frequency). To test whether the small number of FFR trials recorded during the PI task contained sufficient information for additional analysis, we first pooled classification averages over all subjects and both time points, and evaluated whether classification accuracy was significantly greater than the theoretical chance level of accuracy for 7-way classification with equally balanced numbers of epochs (14.3%). Positive results on this test suggest that the approach is sensitive to stimulus information despite the small number of epochs available for the PI task FFR analysis, and support its suitability for assessing the research question concerning sleep differences.

2.8 Statistical approach

For each behavioural or neurophysiological measure, a within-subjects difference value was computed between the pre- and post- sessions. Directional independent samples t-tests (or non-parametric equivalent in case of violation of assumptions) were used to assess the main research questions, i.e., that different metrics across testing session for each behavioural and neurophysiological measure was greater in the Nap vs. Wake group. The 95% Confidence Intervals reported with the results were computed around the mean difference between groups. In light of previous work [Reetzke et al., 2018] and current theories of sensory plasticity [Kraus, 2021, Ahissar et al., 2009], we expect to see the most evidence of change in metrics that tap into higher-level processing in the short experimental time frame (i.e. $\sim$3.5 hrs between behavioural tests), and the least in the FFR, which in previous work only showed neuroplastic effects after a week of training [Reetzke et al., 2018].

Baysian statistical approaches allow for assessing evidence in favour of both an alternative and null hypothesis, and are gaining acceptance as an alternative to traditional frequentist null hypothesis significance testing in Psychology [Aczel et al., 2018, Wagenmakers et al., 2016, Marsman and Wagenmakers, 2017, Kelter, 2020]. The resulting metric, known as a Bayes factor (BF), is a likelihood ratio of the marginal likelihood of two competing hypotheses, usually a null and an alternative. Bayes factors are expressed as a positive number on a continuous scale. Numbers greater than 1 are interpreted as evidence in favour of an alternative hypothesis (BF10), where bigger numbers indicate stronger evidence; numbers between zero and 1 indicate evidence in favour of the null hypothesis, with smaller numbers indicating stronger evidence [Lee and Wagenmakers, 2014]. Bayes factors for the null hypotheses itself (BF01) can be similarly interpreted, i.e. $>$1 indicates support for the model, whereas values between 0 and 1 support the alternative hypothesis [Aczel et al.,

2018]. We adopt the classification scheme used in JASP to interpret Bayes factors, i.e., separating between "anecdotal", "moderate", "strong", etc. relative evidence for an hypothesis based on the size of the Bayes factor obtained [Kelter, 2020].

To explore trends in support for the existence of a sleep effect across behavioural and neurophysiological measures, we computed two sets of Bayes factors. The first is comparable to the directional hypothesis evaluated using frequentist statistics (i.e., change in Nap group > change in Wake group), as will be referred to as '$BF_{SleepEffect}$'. The second evaluates evidence in favour of a bidirectional null hypothesis (i.e., there is no difference between groups in change on each measure), and will be referred to as $BF_{NoEffect}$. The two calculations are highly related, as they differ only in the directionality of the null hypothesis, but importantly, the null model of directional test (i.e., change in Wake group > change in Nap group) is not as meaningful in the present context as the bidirectional version. This distinction is relevant for the interpretation of trends in the results. The 95% Credible Intervals (i.e., the interval within which the parameter value falls with a 95% probability) reported with the results were computed around the mean difference between groups. Bayes factors were calculated with the default Cauchy prior width of 0.707, as no other information was available to update this prior. As this choice of prior width may impact the results, we also conducted a robustness check of the posterior using two wider prior distributions (1.0, 1.5), and report if the wider prior distributions significantly impacted the results (see Supplementary Materials for an analysis of the effect of prior distribution width, as well as sequential analyses, which are used to evaluate the stability in the data over the last 5-10 subjects). Bayesian statistics were conducted with the software program JASP (Version 0.14.3) [JASP Team, 2021].

FFRs to the /yi2/ and /yi6/ tones were measured three times; before and after the initial behavioural training and testing (FFR1 and FFR2), and then again after the Nap or Wake period (FFR3). To create a parallel analysis to the behavioural measures, which were measured twice, we conducted between-groups analyses on difference scores for each metric, across the two time points that bracket the experimental manipulation of Nap vs. Wake (FFR2 and FFR3). This approach facilitates comparison of the Bayes Factors across the behavioural and physiological measures, while the additional datapoint (FFR1) is available to further interrogate the source and nature of significant results, if present.

3 Results

3.1 Auditory experience

Eleven subjects in each group reported having had some formal instrumental or vocal musical instruction, and in each group the age of start amongst them ranged from 4-16 years of age. The mean cumulative musical practice and training hours in the Wake group was 1931 hrs (*SD* = 2652 hrs, range: 0 - 9609 hrs), and in the Nap group, 1221 hrs (*SD* = 1807 hrs, range: 0 - 6614 hrs). These data did not include 2 subjects in the Nap group and 4 subjects in the Wake group who did not fully complete the musical experience questions in the MMHQ. No subjects had tonal language experience, but the majority were fluent bi- or multilinguals (Wake group: 2 mono, 13 bi, 5 multi, 1 did not answer; Nap group: 1 mono, 10 bi, 4 multi, 4 did not answer). Both groups therefore appeared to be similarly heterogeneous in their musical and linguistic experience groups, consistent with our experience with our local population.

3.2 Sleep analysis

We used the first item on the Subjective Fatigue Scale, which concerns feelings of tiredness ("How tired are you?" on a scale of 1 to 5 where 5 is "very tired"), to confirm that the groups did not differ significantly in their level of tiredness at the beginning of the experiment. Mean Wake group score was 2.90 (*SD* = 0.81), and Nap group mean score was 2.91 (*SD* = 0.94). The two groups did not differ in their initial subjective tiredness ratings (W = 207.5, p = .83; two-tailed; Mann-Whitney test was used due a violated Shapiro-Wilk test of normality).

Sleep scoring analysis confirmed that the 19 Wake group participants did not fall asleep. All of subjects in the Nap group slept, between 103.5 and 171.5 minutes (mean = 148.9; *SD* = 20.1; see Figure 3). The average time that the Nap group spent awake was 34.3 mins (*SD* = 21.2), they spent an average of 24.3 mins in N1 (*SD* = 12.1); 68.8 mins in N2 (*SD* = 18.4); 32.6 mins in N3 (*SD* = 20.3), and 23.04 mins in REM (*SD* = 15.1). EOG and EMG was not available for one Nap group subject, making it difficult to distinguish REM and Wake states; ambiguous epochs were marked as Wake for this participant (a distinction that has no bearing on the current research questions). Cumulative N2 and N3 time was computed as a measure of time in NREM sleep for further analysis (mean = 101.5 mins, *SD* = 24.5), with follow-up tests on N2 and N3 separately in case of positive results.

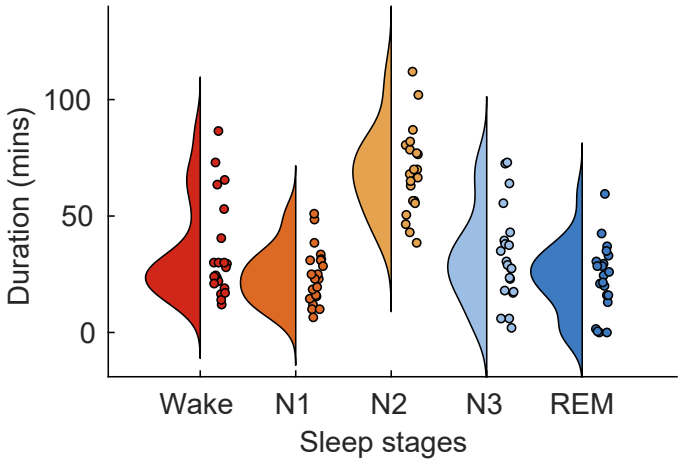


Figure 3: Time spent in each sleep stage for the participants in the Nap group. All subjects slept, and most spent some time in each sleep stage, including rapid eye-movement (REM) sleep. N1-N3 indicate non-REM sleep stages 1-3, where 1 is light and 3 is deep sleep.

3.3 Training paradigm performance gains

The percent change in accuracy between blocks 6 and 7 in the Nap group (N = 21) was 3.24% (*SD* = 6.94), and in the Wake group (N = 19) was -2.83% (*SD* = 7.54). The Nap group had a significantly greater accuracy increase between pre- and post- sessions than did the Wake group ($t(38) = 2.65$, $p = .006$, $d = 0.84$, 95% CI [0.290, $+\infty$]), see Figure 4. Due to an equipment problem, data for some training blocks for several subjects in each group was not stored; mean learning trajectories are therefore not visualized. Critically, all subjects completed all blocks of training, and data from the Blocks 6 and 7 for all subjects was preserved, allowing us to address the research questions. As a control, we confirmed that the two groups did not differ in their pre-break (Block 6) test scores (W = 153.5, $p = .217$; a Mann-Whitney test was used due a violated Shapiro-Wilk test of normality). To control for the possibility that the subject who fell asleep and was moved to the Nap group was driving the observed effects, we recomputed statistics for this the main behavioural research questions; the pattern of results was qualitatively unchanged.

To compare support for the null vs. alternative hypotheses over different performance and physiological measures, we computed Bayes Factors. As with the frequentist statistics, Bayesian analysis indicated a sleep effect on accuracy improvement. There was moderate to strong evidence in favour of the directional hypothesis that the increase in accuracy was greater in the Nap than the Wake groups ($BF_{SleepEffect} = 8.75$, error = 9.319e-5, 95% Credible Interval [0.141, 1.358]). The non-directional null hypothesis also was not supported, $BF_{NoEffect} = 0.23$, error = 2.484e-4, 95% Credible Interval [0.094, 1.356]; see Figure 8 for a comparison with other measures.

In the Nap group ($N = 21$), percent change in accuracy was not significantly correlated with cumulative minutes of N2 and N3 sleep (Spearman's $r(19) = .087$, $p = .35$; one-tailed). We explored the possibility of correlations with the number of minutes in each of the sleep stage separately as well as time spent awake, and found no significant relationships to percent change in accuracy on the training task. The results concerning improvement on the training task therefore indicate support for a sleep effect, although it could not be attributed to duration of time spent in NREM sleep.

3.4 Development of a categorical perceptual boundary

Behavioural results from the PI task, which are used to assess changes in the level-to-rising categorical boundary, are presented in Figure 5. Due to a technical problem, one subject from the Wake group did not have complete PI task data and was excluded from this analysis. Reported analyses include $N = 21$ subjects in the Nap group and $N = 18$ subjects in the Wake group.

To test the main hypothesis that sleep affects the development of a categorical boundary as an outcome of the training paradigm, we first evaluated whether there was a sleep-dependent increase in RT peak, and a sleep-dependent increase in slope (from a fitted a logistic regression curve to each subjects' identification percent values).

The mean RT peak difference in the Nap group was 0.046 (*SD* = 0.11), and in the wake group, -0.01 (*SD* = 0.13). The Nap group had a greater increase in RT peak difference than the Wake group, which just met significance threshold (W = 248, $p = .050$, rank-biserial correlation: $r_{rb} = .312$, 95% CI [0.014, $+\infty$]; a Mann-Whitney test was used due a

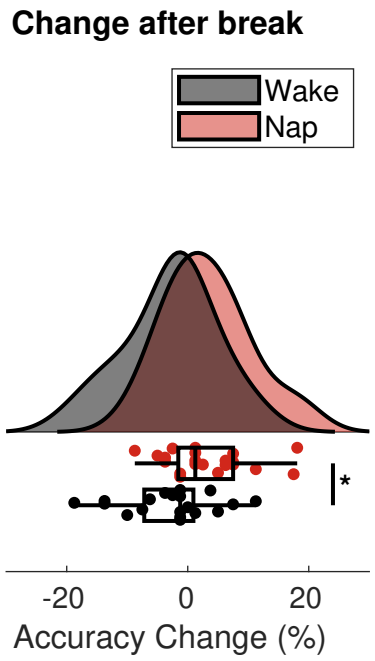


Figure 4: Change in accuracy on the sound-to-category training paradigm over the three-hour break, for Nap and Wake groups (percent change in accuracy difference between Block 7 and Block 6; positive values indicate a post-break performance gain). Shaded bars (left) and whiskers (right) indicate SEM, asterisk indicates significance at p<0.05. Groups did not differ in their pre-training accuracy.

violated Shapiro-Wilk test of normality). Bayes factors were equivocal about a sleep effect on RT peak; the directional hypothesis showed anecdotal evidence in favour, $BF_{SleepEffect}$ = 1.31, error = 5.8e-5, 95% Credible Interval [0.034, 0.999]; but the non-directional null hypothesis was also anecdotally supported, $BF_{NoEffect}$ = 1.37, error = 9.2e-4, 95% Credible Interval [-0.190, 0.985], (see Figure 8).

As described in Methods, subjects were removed from the slope analysis if their responses did not allow for a reliable slope difference measure. Three Nap group subjects and three Wake group subjects were excluded from the slope analysis based on poor slope function fit in at least one of the PI sessions, caused for example by responses that did not show a monotonically increasing percentage of flat vs. rising tone identification across the 7-point continuum. The mean change in slope across sessions for the remaining subjects was 0.475 (SD = 0.61) for the Nap group (N = 18), and 0.081 (SD = 0.93) for the Wake group (N = 15). The Nap group had a significantly larger increase in slope than the Wake group (W = 185, p = .037, rank-biserial correlation: r_{rb} = .370, 95% CI [0.052, $+\infty$]; a Mann-Whitney test was used due a violated Shapiro-Wilk test of normality). Bayes factors were also equivocal about a sleep effect on slope; the directional hypothesis showed anecdotal evidence in favour, $BF_{SleepEffect}$ = 1.34, error = 2.2e-4, 95% Credible Interval [0.035, 1.074]; but the non-directional null hypothesis was also anecdotally supported, $BF_{NoEffect}$ = 1.34, error = 0.003, 95% Credible Interval [-0.214, 1.057], (see Figure 8).

Within the Nap group, the change in RT peak was positively correlated with the duration of NREM sleep (Pearson's $r(21) = 0.40$, p = .038, 95% CI [0.032, 1.000]); however, the change in slope was not (Pearson's $r(18) = -0.35$, p = .92, 95% CI [-0.659, 1.000]). For the RT peak correlation, we further explored whether duration of time in N2 or N3, and additionally REM because of the possibility of complementary contributions to learning from these sleep stages [Tamaki et al., 2020]. The clearest relationship was with N2 sleep duration (Pearson's $r(21) = 0.43$, p = .026, 95% CI [0.070, 1.000]). There was no apparent relationship to duration of N3 sleep (Pearson's $r(21) = 0.09$, p = .351, 95% CI [-0.290, 1.000]), nor with duration of REM sleep (Pearson's $r(21) = -0.15$, p = .747, 95% CI [-0.495, 1.000]). The results concerning categorical boundary development indicate modest though equivocal support, which could be partially attributed to duration of time spend specifically in stage 2 NREM sleep.

For completeness, we computed correlations between the three behavioural measures across the whole group (excluding unpaired data points), after the initial training session. Training accuracy was positively correlated with PI slope (Spearman's $r(31) = .41$, $p = .018$; one-tailed), but not with with RT peak (Spearman's $r(31) = .17$, $p = .35$); nor were the two PI metrics correlated (Spearman's $r(38) = .02$, $p = .91$), suggesting that they are tapping into different aspects of performance.

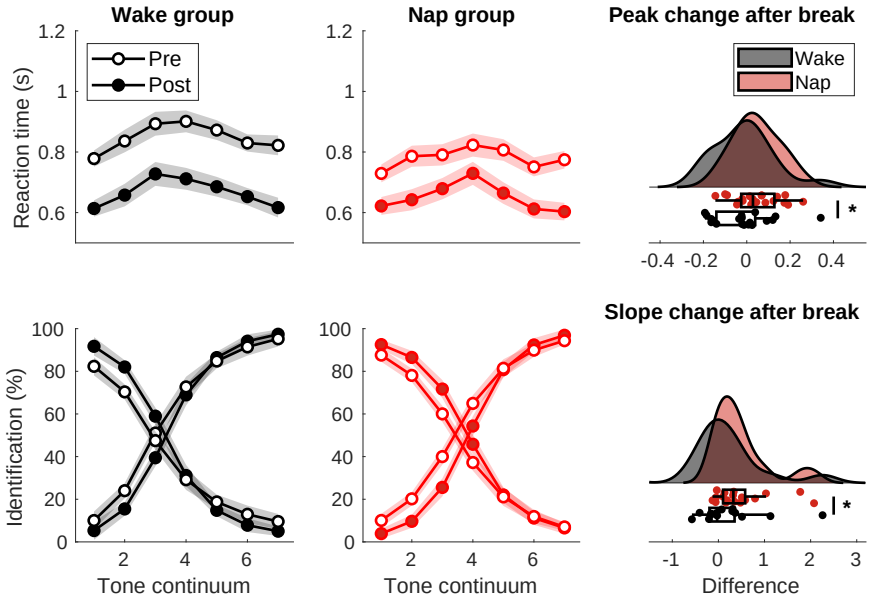


Figure 5: Behavioural results from the perceptual identification task for the Nap and Wake groups. Top row: reaction times to classifying 7 stimuli forming a tone continuum into 'level' or 'rising' categories; relatively slower reaction times close to the middle of the tone continuum (tone token 4) indicate the development of a perceptual boundary. Right: the Nap group showed a significantly greater increase in reaction time peak than the Wake group. Bottom row: identification percent for each stimulus on the tone continuum. A steeper slope close to the crossing point indicates a perceptual boundary. Open circles indicate mean pre-training, closed circles indicate mean post-training, and shaded bars indicate SEM. Right: the Nap group showed a significantly greater increase in slope than the Wake group.

3.4.1 Confirmation of dynamic pitch representation in the FFR

To address the research question concerning whether the representations of the speech tones /yi2/ and /yi6/ are affected by sleep, the three FFR recordings were used. Only subjects with complete, high quality recordings from all three FFR recordings were included in the main FFR analysis. Due to an intermittent technical problem which made it impossible to mark sound onset with the sub-millisecond resolution needed for FFR analysis, five subjects from each group had to be excluded. Two additional subjects from the Nap group and one from the Wake group had to be excluded due to poor electrode contact during at least one of the recordings. The final sample included 14 subjects in the Nap group and 12 in the Wake group.

To verify that FFR captures the subtle differences in pitch dynamics between the two /yi/ tones that were used for the main FFR analysis, we created a grand average timeseries across all subjects, and across the three FFR recordings, for each tone. We then performed an autocorrelation analysis, using identical parameters as used to visualize the stimulus (compare Figure 2 with Figure 6). At the group level, a clear representation of the stimulus' fundamental frequency is observed at lags of 6-7ms in the /yi2/ stimulus, which corresponds to a frequency of 130 Hz. In the /yi6/ FFR, a change in lag over time indicating pitch rising from 105 Hz to 130 Hz is observed. The FFRs appear to capture the acoustic differences between the tones, which is necessary to address the research questions.

3.4.2 Evaluation of a sleep effect on the FFR using traditional methods

We evaluated differences in the change in stimulation-to-response correlation and pitch tracking accuracy (FFR3 - FFR2) between the Nap or Wake period, separately for /yi2/ and /yi6/ stimuli.

The mean change in stimulus-to-response correlation for the /yi2/ stimulus for the Nap group ($N = 16$) was -4.451 ($SD = 9.59$), and for the Wake group ($N = 13$), the mean was 8.95 ($SD = 27.40$). The Nap group's change in stimulus-to-

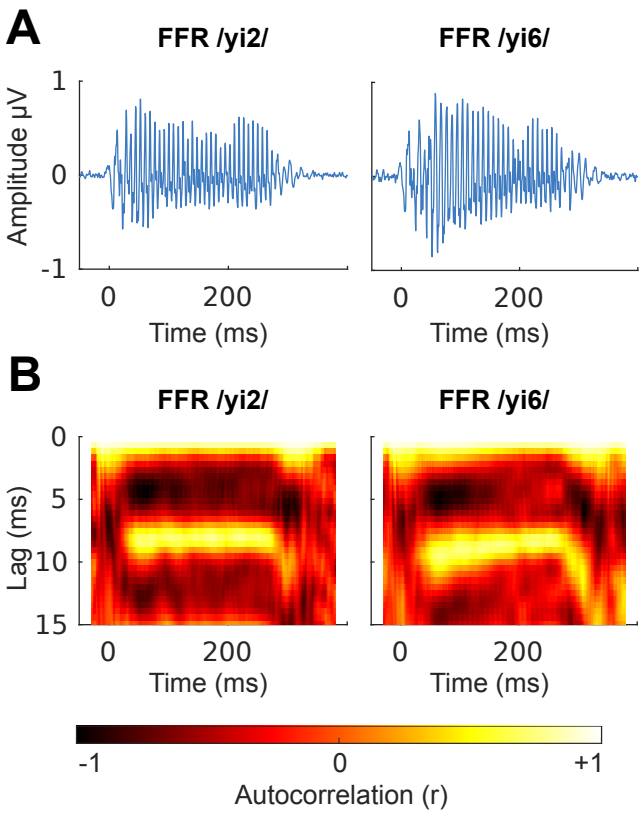


Figure 6: Frequency following responses (FFRs), which correspond to tones 2 and 6 in the Perceptual Identification (PI) task, averaged across all subjects and time points. (A) Waveforms of the two /yi/ tones. (B) Autocorrelograms of the averages reveal that differences in the stimulus' pitch dynamics are captured in the neurophysiological recording (compare with stimulus in Figure 2). Colours represent the strength of correlation, ranging from low (dark red) to high (white) at time lags corresponding to the period of the stimulus' fundamental frequency.

response correlation over sessions was not significantly greater than the Wake group's change (W = 48, p = .99, r_{rb} = -.538, 95% CI [-0.745, $+\infty$]; a Mann-Whitney test was used due to a violated Shapiro-Wilk test of normality). Bayes factors did not support an effect of sleep on stimulus-to-response correlation; $BF_{SleepEffect}$ = 0.148, error = 2.9e-4, 95% Credible Interval [0.004, 0.460]; $BF_{NoEffect}$ = 0.842, error = 0.001, 95% Credible Interval [-1.261, 0.136]. For the /yi6/ stimulus, the Nap group (N = 16) was -2.51 (SD = 8.9), and for the Wake group (N = 13), the mean was -4.71 (SD = 6.3). The Nap group's change in stimulus-to-response correlation over sessions was not significantly greater than the Wake group's change (t(27)=0.747, p = .231, d = 0.279, 95% CI [-0.341, $+\infty$]). Bayes factors did not support an effect of sleep on stimulus-to-response correlation; $BF_{SleepEffect}$ = 0.637, error = 0.008, 95% Credible Interval [0.017, 0.915]; instead, there was anecdotal support in favour of no effect of sleep; $BF_{NoEffect}$ = 2.315, error = 4.593e-4, 95% Credible Interval [-0.418, 0.869].

The mean change in pitch tracking accuracy as measured by autocorrelation for the /yi2/ stimulus for the Nap group was 0.188 (SD = 0.96), and for the Wake group, the mean was 0.822 (SD = 2.82). The Nap group's change in pitch tracking over sessions was not significantly greater than the Wake group's change (W = 106, p = .47, r_{rb} = .019, 95% CI [-0.329, $+\infty$]; a Mann-Whitney test was used due a violated Shapiro-Wilk test of normality). Bayes factors did not support an effect of sleep on pitch tracking; $BF_{SleepEffect}$ = 0.216, error = 9.592e-5, 95% Credible Interval [0.006, 0.576]. Conversely, there was anecdotal support for no effect of sleep; $BF_{NoEffect}$ = 2.186, error = 4.085e-4, 95% Credible Interval [-0.901, 0.392], (see Figure 5). For the /yi6/ stimulus, the Nap group mean was -0.288 (SD = 1.6), and for the Wake group, the mean was 0.44 (SD = 2.1). The Nap group's change in stimulus-to-response correlation over sessions was not significantly greater than the Wake group's change (t(27) = -1.047, p = .85, d = -0.391, 95% CI [-1.008, $+\infty$]). Bayes factors did not support an effect of sleep; $BF_{SleepEffect}$ = 0.197, error = 2.584e-4, 95% Credible Interval [0.006, 0.547]. As for /yi2/, there was anecdotal support in favour of no effect of sleep; $BF_{NoEffect}$ = 1.896, error = 8.953e-5,

95% Credible Interval [-0.972, 0.338]. In sum, across the two traditional FFR measures and repeated for two similar stimuli, we found no evidence of a sleep effect on the passively-recorded FFR.

3.4.3 Evaluation of a sleep effect using classification accuracy

To further explore the possibility that there are subtle changes in the fine sound encoding due to sleep and learning, we decoded the two Mandarin lexical tones from the EEG recordings, for each participant. Table 1 lists the mean performance of the classifier for each FFR recording session and group, using the full 450ms epochs (including baseline and post-stimulus periods). Figure 7a) displays the confusion matrices averaged across all three recordings (i.e., an N x N matrix that compares the actual audio labels with those predicted by the classifier).

To confirm that the classifier was using information present during the FFR window itself and not the baseline or post-stimulus period, we repeated the classification using: 1) only the pre-stimulus time segment (0-50 ms), 2) only the stimulus segment (10-370 ms based on Reetzke et al. [2018]), and 3) the post-stimulus time segment (350-400 ms). As expected, both pre-stimulus and post-stimulus time segments showed a chance-level performance (see supplementary materials), and the classification on the whole epoch is highly similar to that using only the stimulus segment. The successful classification is therefore due to neural encoding of sound that is present during and shortly after stimulus presentation. Interestingly, in case of the post-stimulus time segment, the average results are slightly better than the chance-level (1-2%), suggesting that the after-effect of FFR entrainment still contains some information concerning the recently heard sound [Coffey et al., 2021].

As for previous measures, we focused on group differences in the change between the Nap or Wake period (FFR3 - FFR2). The difference in accuracy between Nap and Wake groups was not significant (W = 87, p = .85, r_{rb} = -0.223 [-0.523, $+\infty$]; a Mann-Whitney test was used due a violated Shapiro-Wilk test of normality). Bayes factors also did not support an effect of sleep; $BF_{SleepEffect}$ = 0.154, error = 4.345e-4, 95% Credible Interval [0.004, 0.469]; $BF_{NoEffect}$ = 1.066, error = 0.005, 95% Credible Interval [-1.168, 0.179].

Table 1: Mean (+/-SD) classification performance for the FFR recordings. The training task occurred between sessions 1 and 2, and the Nap or Wake period occurred between sessions 2 and 3.

Task	**Nap**	**Wake**
FFR task 1	0.70 ± 0.07	0.71 ± 0.09
FFR task 2	0.70 ± 0.08	0.69 ± 0.09
FFR task 3	0.68 ± 0.08	0.71 ± 0.10
Average Performance	0.70 ± 0.08	0.70 ± 0.09

3.4.4 Feasibility of using FFRs from the PI task

Three subjects from the nap group and two from the Wake group had to be excluded from the FFR analysis of data collected during the PI task, due to either a technical problem involving precisely detecting sound onsets, or poor electrode contact during at least one of the recordings. The final sample included 18 subjects in the Nap group and 17 in the Wake group. For each recording (i.e. PI1, PI2), we performed a within-subject classification procedure, as in the passive FFR recording, but this time with 7 categories. As previously, chance level (i.e., $\sim$ 14.3% for each category) is computed by performing permutation on the labels of the audio stimuli (see supplementary materials for classifier and permutation distributions).

The mean seven-way classification accuracy amongst the pooled classification averages was 17.4% (SD = 5.2). Table 2 presents the performance of the classifier under each condition for each PI task. Figure 7b displays the confusion matrix for the PI task. Classification accuracy was greater than chance levels (W = 1819.0, p <.001, rank biserial correlation: r = .41). This result confirms that information concerning stimulus identity is captured even within the small number of available PI task FFR trials, which serves as a basis for further exploring our research hypothesis concerning the effect of sleep further in these data.

As for the main FFR analysis, we investigated the effects of classifying different time segments on the classifier's performance. As expected, the successful classification is driven by the time period in which the stimulus is presented, and both pre-stimulus and post-stimulus time segments show a chance-level performance (see Supplementary Materials).

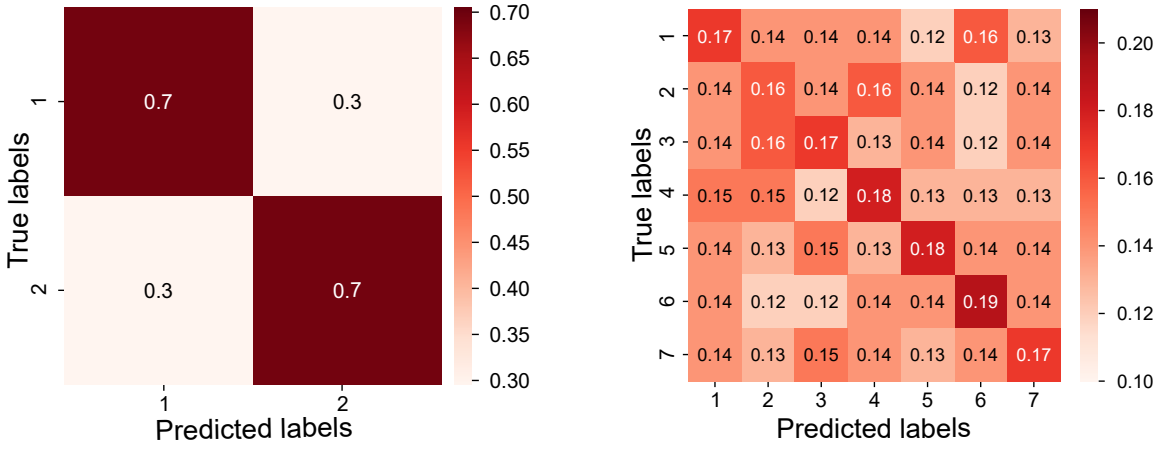
3.4.5 Evaluation of a sleep effect using FFRs from the PI task

As before, we focused on group differences in the change between the Nap or Wake period (PI2 - PI1). The difference in classification accuracy between Nap and Wake groups was not significant (t(33) = -1.420, p = .917, d = -0.480, 95%

Table 2: Mean (+/- *SD*) classification accuracy across time points for the PI task recording

Task	Nap	Wake
PI task 1	0.17 ± 0.11	0.16 ± 0.10
PI task 2	0.17 ± 0.11	0.19 ± 0.10
Average Performance	0.17 ± 0.11	0.18 ± 0.10

CI [-1.041, $+\infty$]). Bayes factors did not support an effect of sleep; $BF_{SleepEffect}$ = 0.155, error = 5.281e-4, 95% Credible Interval [0.004, 0.459]; $BF_{NoEffect}$ = 1.418, error = 0.015, 95% Credible Interval [-1.014, 0.215].



(a) Binary classification (FFR task)

(b) 7-class Classification (PI task)

Figure 7: (A) The normalized confusion matrix for two-way classification of auditory stimuli (i.e. /yi2/ and /yi6) used in the passive FFR recording, averaged across all time points. The brain's responses to both stimuli are classified well above chance level. (B) The normalized confusion matrix of the 7-class classification based on FFR data recorded during the perceptual identification (PI) task). All categories are classified better than the chance level (i.e. 0.142).

3.4.6 Comparison of a sleep effect across measures

To explore trends in support for the existence of a sleep effect across behavioural and neurophysiological measures, we have reported $BF_{SleepEffect}$, which expresses evidence in favour of the directional hypothesis that sleep has a positive effect; and $BF_{NoEffect}$, which indicates support for the non-directional hypothesis that sleep has no effect, in previous sections. Bayes factors for the main research questions can be compared in Figure 8, in which higher-level behavioural measures that are thought to have faster timescales of neuroplasticity are on the left, and lower-level, physiological measures which are thought to change more slowly over training are found on the right. Evidence in support of the $BF_{SleepEffect}$ model is strongest in the training task results, and is also observed at 'anecdotal' levels [Kelter, 2020] in the two measures of category development (reaction time peak and slope). There is no evidence in favour of the $BF_{SleepEffect}$ model in any of the FFR measures. The non-directional $BF_{NoEffect}$ model shows an opposite pattern, with stronger support for the physiological measures.

4 Discussion

To investigate how sleep affects neuroplasticity during training of auditory perceptual skills, we evaluated its influence on learning categories between lexical Mandarin speech tones, in a nap design. Measures of behavioural performance on the sound-to-category training task, which engages higher-level cognitive representations and resources such as attention and use of feedback, were significantly better in a group that had an afternoon nap as compared with a group that stayed awake. The data clearly supported a positive effect of sleep, using both frequentist hypothesis testing and Bayesian statistical approaches.

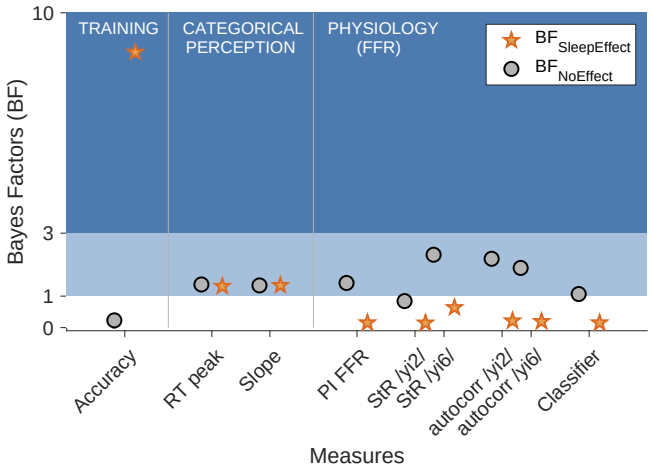


Figure 8: Comparison of sleep effects across behavioural and neurophysiological measures. Bayes factors for each of the main between-group difference comparisons used in this study are arranged according to expected degrees of plasticity within a short time scale; higher-level behavioural measures that are thought to have faster timescales of neuroplasticity are on the left, and lower-level, physiological measures which are thought to change more slowly over training are found on the right. Higher Bayes factors indicate support in favour of the model where values between 1 and 3 indicate 'anecdotal' support (pale blue) and between 3 and 10 mean 'moderate' support (dark blue) according to established scales [Kelter, 2020]. Accuracy: performance on the training task; RT peak: reaction time at the categorical boundary on the PI task; Slope: slope of the identification curve on the PI task; PI FFR: FFR recorded during PI task; StR: stimulus-to-response correlation; autocorr: pitch tracking accuracy by autocorrelation; Classifier: accuracy of FFR classification.

Conversely, neural indices of changes to fine sound encoding fidelity (i.e., the frequency-following response) were not significantly different between groups using frequentist statistics, and the null model was supported using Bayesian statistics (which allow for quantification of evidence in support of null models [Wetzels et al., 2009]). Specifically, there was no evidence of a sleep benefit, neither in the main FFR recordings in which participants passively were exposed to speech stimuli, nor when they were actively engaged in the PI task. We calculated traditional measures of stimulus-to-response correlation and pitch tracking, as well as machine learning-based classification. No effect of sleep was observed at this basic level of representation, despite that we had attempted to increase the sensitivity of the analysis over previous designs that measured change over several sessions (e.g. [Reetzke et al., 2018]).

The measures of categorical boundary development based on the perceptual identification (PI) task can be argued to lie conceptually between the training task and FFR measures in terms of the level of cognitive process they tap into, although with the caveat that the measures are not of the same type (i.e. the training task and PI task are both measures of behaviour, whereas the FFR measures physiology). Binary categorizing a closed set of seven artificially generated sounds that vary subtly and parametrically in their onset frequency is a more constrained task than is the sound-to-category training, in which subjects are confronted with many naturalist exemplars of human speech sounds that vary along multiple dimensions and are asked to categorize them into four bins without first knowing the basis of categorization. The Nap group had a significantly larger increase in the two measures of category development (i.e. reaction time peak and slope) as compared with the Wake group, albeit with smaller effect sizes than with the sound-to-category training measure. Likewise, the Bayes factors for categorical boundary development lay between the training task and the physiological measures (Figure 8).

For the results that were significantly different between groups, we investigated whether there was a correlation within the Nap group between performance and the amount of time in NREM sleep. Neurophysiological events known as sleep spindles and slow oscillations occur during NREM sleep stages 2 and 3 and are involved in memory consolidation, the process by which temporarily stored associations are converted into long-lasting neuroplastic changes [Rasch and Born, 2013, Schönauer, 2018, Niethard et al., 2018]. Some authors have suggested that slow-wave sleep in particular is required for perceptual memory formation, on the basis of texture discrimination learning in mice [Miyamoto et al., 2016]. Of the training, peak RT, and slope measures, only the peak RT difference was correlated with time spent in NREM sleep (i.e., cumulative minutes in N2 and N3), which we interpret as modest additional support in favour of an effect of sleep. For peak RT differences, we additionally explored the relationship to each of the NREM sleep stages (N2, N3), which differ in the proportion of slow oscillations present and can sometimes be distinguished by relative contribution to learning (e.g. [Laventure et al., 2016]), and in REM, which has different neurophysiological features

and may play a complementary role in neuroplastic processes during sleep [Tamaki et al., 2020, Karni et al., 1994]. We found a significant positive correlation only with N2 sleep, in-line with an earlier result showing a relationship between time spent in light NREM sleep (N1 and N2) and an improvement in an auditory task in which participants identified categories of auditory tokens [Earle et al., 2017]. This result may suggest that spindle activity rather than slow oscillations are the driver of sleep-related perceptual gains in categorical perception [Schönauer, 2018, Laventure et al., 2016].

An interesting question is how perceptual learning relates to previous work on procedural vs. declarative memory consolidation in sleep [Earle and Myers, 2014], which show somewhat different patterns with respect to sleep stage and microarchitecture (i.e. spindles, slow oscillations) [Rasch and Born, 2013]. The dual-system model of learning posits the existence of an explicit, hypothesis-driven 'reflective' system as well as an implicit, procedural-based 'reflexive system' [Chandrasekaran et al., 2014]. Minimal, immediate feedback promotes use of the reflexive system, and delayed, detailed feedback promotes use of the reflective system. Speech category learning of the nature examined in this study has been argued to be more procedural, on the basis that it uses the reflexive learning system. Although the current study does not address these issues, future work on sleep and perceptual learning should consider that specific task demands may influence the system used and thus learning and neuroplasticity (see [Earle and Myers, 2014] for further discussion). The current study was designed to address the main research questions concerning the overall effect of sleep vs. wake on behavioural and neural indices of categorical perception development. More work using a larger sample of sleeping participants, longer sleep duration, and an analysis of fine differences in sleep architecture (e.g. number, density, and amplitude of sleep spindles and slow oscillations) will be necessary to clarify which sleep process are most critical to this type of perceptual learning [Cordi and Rasch, 2021].

In general, our results agree with theories suggesting that higher-level categorical representations are the most flexible and easily adapted to novel task demands, whereas lower-level sensory changes emerge only at later stages after considerable practice [Kraus, 2021, Ahissar et al., 2009, Reetzke et al., 2018]. According to the Reverse Hierarchy Theory (RHT), which is a general theory about learning, higher-level categorical representations guide immediate, novice learning whereas lower-level sensory changes do not emerge until later training stages [Ahissar et al., 2009]. The BEAMS framework ('Brain, via Efferent influence, attaining a new default Afferent state that represents the Memory of Sound') more specifically addresses plasticity of auditory representation. [Kraus, 2021]. According to BEAMS, over time, the sounds that are behaviourally relevant to an individual become prioritized to be processed automatically, rapidly, and preferentially [Kraus, 2021]. This organization, in which higher-level and lower-level processes have different speeds of plasticity is in a sense strategic, as it allows for flexible behaviour while maintaining system stability.

Gradual tuning and sharpening of lower-level processes such as the representation of dynamic pitch that we are measuring with the FFR can be achieved with training and experience [Reetzke et al., 2018, Krishnan et al., 2009, 2010]. Better neural encoding of acoustic information including sound frequency can reduce the reliance on cognitive resources for processing relevant environmental information [Peelle, 2018]. In the case of novice Mandarin learners, they might rapidly learn that pitch contours are relevant and must be attended to. With exposure and practice, improvements in neural encoding emerge [Ahissar et al., 2009], making the process of using acoustic information less effortful, and perhaps more robust to degradation [Krishnan et al., 2010, 2009].

As in previous work using the same learning paradigm [Reetzke et al., 2018], our results support that sensory enhancement of incoming stimulus feature is not critical for early stages of perceptual learning, and rather that novice performance is guided by abstract categorical representations [Caras and Sanes, 2017]. A possible caveat to our study design is that neuroplasticity associated with tone category learning may be specific to the parameters of naturally produced tones, or of the specific acoustic features of the stimuli used in training [Xu et al., 2006b]. However, differences in neural responses to between vs. within-category deviations have also been observed in native Mandarin speakers to linearly varying pitch tones that are similar to those used in the present study [Li et al., 2020].

A limitation to this study is its sample size of forty participants, across two experimental groups. Although this sample size is in-keeping with the majority of studies of sleep, behaviour, and neurophysiology, a recent review argued for larger samples across sleep and memory research [Cordi and Rasch, 2021]. Studies of this nature are resource-intensive and ideal sample sizes may be difficult to achieve; however, design choices may improve the robustness of results. The present study attempts to compensate for modest sample size in several ways: there are multiple measures taken across time and within-subject, some of them internal replications (e.g. measures of stimulus-to-response correlation are computed for each of two stimuli, /yi2/ and /yi6/, which show the same patterns). The Bayesian statistical approaches allow us to quantify support in favour of both a null and alternative hypotheses. Sequential analyses are used to evaluate the stability over the last 5-10 subjects, which supports the robustness of the results. Finally, our interpretation of the findings concern the pattern of results across a number of behavioural and physiological metrics (Figure 8). In the future, resource-intensive laboratory studies could be complemented with larger in-home studies using wearable equipment and a lighter experimental protocol. Another way to increase sensitivity of the study design would be to

decrease variability between participants, for example by training participants to a performance criterion in the first session, and discontinuing the experiment for participants who fail to reach it within some fixed limit. This design choice would be less generalizable to the population at large, some of whom do struggle to learn new non-native speech categories [Paulon et al., 2021], but would perhaps offer a clearer view of sleep dependency among participants who do learn the task well. Another potential improvement would be to include an objective measure of alertness following the rest period, such as the psychomotor vigilance task [Lee et al., 2010], to control for differences in levels of fatigue (or of sleep inertia) between groups. Reaction time measures are most often used in such cases, as they are more sensitive to fatigue and sleepiness than are other performance metrics (e.g., Balkin et al. [2004]). In the present study, reaction times were used only in the PI task, and a relative measure of peakiness was computed between categories for each subject, making overall speed differences unlikely to have affected our metric. None of the other measures were dependent upon reaction times, subjective feelings of tiredness did not differ between groups before the experiment, and all participants reported feeling alert before the second set of behavioural tasks; the present results are therefore unlikely to have been driven by post-sleep differences.

This study expands our understanding of the ways in which sleep contributes to memory consolidation beyond the standard declarative vs. non-declarative dichotomy, and shows that sleep can benefit these early stages of perceptual learning. At least in the short timeframe of this study (3 hrs), sleep appeared to play the largest role in the higher-level aspects of learning. An important open question is whether sleep also plays a role in the slow changes that are observed in the sensory encoding of relevant stimulus features over multiple days or weeks. This research question poses challenges to traditional nap or overnight sleep designs when the timecourse of neuroplasticity exceeds an experimentally reasonable duration of wakefulness, but may be accessible in animal models using causal tools. For example, the specific sleep events relevant for the process could be identified and selectively manipulated via optogenetic tools to probe the sleep effect, as has been used in other sleep and memory contexts (e.g. in demonstrating a role of REM sleep in memory consolidation Boyce et al. [2016]).

A common approach to understanding the extent of sleep's affect on shaping brain function and behavioural performance is to define tasks that isolate specific aspects of performance. This generalization is particularly true in the domain of perceptual learning, where tasks that tap only into 'low-level' processes may be sought (e.g. [Klinzing et al., 2021]). In agreement with Klingzing et al., who concluded against a sleep effect on early perceptual learning on the basis of a visual task, we do not observe changes in the frequency-following response, a basic index of auditory encoding plasticity. However, many learning processes are extended over days or weeks, and may involve different cognitive systems over the timecourses of their mastery. This appears to be true of categorical perceptual learning [Reetzke et al., 2018], in which FFR changes are observed only after about a week. Long neuroplastic timecourses are a common feature of many sensorimotor skills (e.g. [Savion-Lemieux and Penhune, 2005]), in which higher and lower levels of representation and processing within the sensory systems work together to achieve behavioural adaptation in a changing environment [Coffey et al., 2019, Kraus, 2021]. Progressive adaptation to the cognitive demands of the task and the test procedure, e.g., changes in attentional demands or task learning, may differ in their relationships to sleep [Alain et al., 2015]. We propose that considering sleep's role at and between levels of representation, and over longer timecourses, may be a necessary and valuable approach to understanding sleep's role in perceptual learning.

5 Acknowledgments

The authors thank Caroline Holden and Camille Bouhour for assisting with data collection; Casey Roark, G. Nike Gnanateja and Fernando Llanos Lucas for sharing code and materials, some of which are credited to Jessica Roeder for the development of the dual-task and Erika Skoe for providing the MATLAB codes to create the autocorrelograms and implement the F0 tracking analysis; Monika Schönauer for advice concerning the nap design; Robert Zatorre and Giovanni Beltrame for discssion; and Aaron Johnson for advice on Bayesian statistics. We thank the International Laboratory for Brain, Music and Sound Research (BRAMS) for loan of the EEG equipment. Funding was provided by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to EC.

6 Author contributions

Aurélien de la Chapelle: Conceptualization, Methodology, Investigation **Marie-Anick Savard:** Formal analysis, Visualization, Writing - Original Draft, Writing - Review and Editing **Reyan Restani:** Methodology, Investigation, Writing - Original Draft **Pouya Ghaemmaghami:** Formal analysis, Visualization **Noam Thillou:** Investigation, Data Curation **Khashayar Zardoui:** Investigation, Data Curation **Bharath Chandrasekaran:** Methodology, Resources, Writing - Review and Editing **Emily B. J. Coffey:** Conceptualization, Methodology, Data Curation, Resources, Formal analysis, Visualization, Writing - Original Draft, Writing - Review and Editing, Supervision, Project administration, Funding acquisition

References

Björn Rasch and Jan Born. About sleep's role in memory. *Physiological reviews*, 2013.

Amélie Morin, Julien Doyon, Valérie Dostie, Marc Barakat, Abdallah Hadj Tahar, Maria Korman, Habib Benali, Avi Karni, Leslie G Ungerleider, and Julie Carrier. Motor sequence learning increases sleep spindles and fast frequencies in post-training sleep. *Sleep*, 31(8):1149–1156, 2008.

Ken A Paller and Joel L Voss. Memory reactivation and consolidation during sleep. *Learning & Memory*, 11(6): 664–670, 2004.

Ke Jia, Elisa Zamboni, Valentin Kemper, Catarina Rua, Nuno Reis Goncalves, Adrian Ka Tsun Ng, Christopher T Rodgers, Guy Williams, Rainer Goebel, and Zoe Kourtzi. Recurrent processing drives perceptual plasticity. *Current Biology*, 30(21):4177–4187, 2020.

F Sayako Earle and Emily B Myers. Building phonetic categories: An argument for the role of sleep. *Frontiers in Psychology*, 5:1192, 2014.

Samira Anderson, Travis White-Schwoch, Alexandra Parbery-Clark, and Nina Kraus. Reversal of age-related neural timing delays with training. *Proceedings of the National Academy of Sciences*, 110(11):4357–4362, 2013.

Jens G Klinzing, Lena Herbrik, Hendrikje Nienborg, and Karsten Rauss. Binocular disparity-based learning is retinotopically specific and independent of sleep. *Philosophical Transactions of the Royal Society B*, 375(1799): 20190463, 2020.

Jens G Klinzing, Hendrikje Nienborg, and Karsten Rauss. Sleep does not aid the generalisation of binocular disparity-based learning to the other visual hemifield. *Journal of Sleep Research*, page e13335, 2021.

Steffen Gais, Werner Plihal, Ullrich Wagner, and Jan Born. Early sleep triggers memory for early visual discrimination skills. *Nature neuroscience*, 3(12):1335–1339, 2000.

Gaétane Deliens, Rémy Schmitz, and Philippe Peigneux. Interocular transfer of perceptual skills after sleep. *Journal of vision*, 14(1):23–23, 2014.

Avi Karni, David Tanne, Barton S Rubenstein, JJ Askenasy, and Dov Sagi. Dependence on rem sleep of overnight improvement of a perceptual skill. *Science*, 265(5172):679–682, 1994.

Robert Stickgold, LaTanya James, and J Allan Hobson. Visual discrimination learning requires sleep after training. *Nature neuroscience*, 3(12):1237–1238, 2000.

Masako Tamaki, Zhiyan Wang, Tyler Barnes-Diana, DeeAnn Guo, Aaron V Berard, Edward Walsh, Takeo Watanabe, and Yuka Sasaki. Complementary contributions of non-rem and rem sleep to visual learning. *Nature neuroscience*, 23(9):1150–1156, 2020.

Nadine Gaab, Miriam Paetzold, Markus Becker, Matthew P Walker, and Gottfried Schlaug. The influence of sleep on auditory learning: a behavioral study. *Neuroreport*, 15(4):731–734, 2004.

Mercedes Atienza, Jose L Cantero, and Robert Stickgold. Posttraining sleep enhances automaticity in perceptual discrimination. *Journal of cognitive neuroscience*, 16(1):53–64, 2004.

Mercedes Atienza, Jose Luis Cantero, and R Quian Quiroga. Precise timing accounts for posttraining sleep-dependent enhancements of the auditory mismatch negativity. *Neuroimage*, 26(2):628–634, 2005.

JM Gottselig, G Hofer-Tinguely, AA Borbely, SJ Regel, H-P Landolt, JV Retey, and P Achermann. Sleep and rest facilitate auditory learning. *Neuroscience*, 127(3):557–561, 2004.

Claude Alain, Kuang Da Zhu, Yu He, and Bernhard Ross. Sleep-dependent neuroplastic changes during auditory perceptual learning. *Neurobiology of learning and memory*, 118:133–142, 2015.

Daphne Ari-Even Roth, Liat Kishon-Rabin, Minka Hildesheimer, and Avi Karni. A latent consolidation phase in auditory identification learning: Time in the awake state is sufficient. *Learning & Memory*, 12(2):159–164, 2005.

Kimberly M Fenn, Howard C Nusbaum, and Daniel Margoliash. Consolidation during sleep of perceptual learning of spoken language. *Nature*, 425(6958):614–616, 2003.

F Sayako Earle and Emily B Myers. 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, 2015a.

F Sayako Earle and Emily B Myers. Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6):1680, 2015b.

F Sayako Earle, Nicole Landi, and Emily B Myers. Sleep duration predicts behavioral and neural differences in adult speech sound learning. *Neuroscience Letters*, 636:77–82, 2017.

Yisheng Xu, Jackson T Gandour, and Alexander L Francis. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, 120(2):1063–1074, 2006a.

Yang Zhang, Patricia K Kuhl, Toshiaki Imada, Paul Iverson, John Pruitt, Erica B Stevens, Masaki Kawakatsu, Yoh'ichi Tohkura, and Iku Nemoto. Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. *Neuroimage*, 46(1):226–240, 2009.

Patrick CM Wong and Tyler K Perrachione. Learning pitch patterns in lexical identification by native english-speaking adults. *Applied Psycholinguistics*, 28(4):565–585, 2007.

Rachel Reetzke, Zilong Xie, Fernando Llanos, and Bharath Chandrasekaran. Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, 28(9):1419–1427, 2018.

Emily BJ Coffey, Trent Nicol, Travis White-Schwoch, Bharath Chandrasekaran, Jennifer Krizman, Erika Skoe, Robert J Zatorre, and Nina Kraus. Evolving perspectives on the sources of the frequency-following response. *Nature communications*, 10(1):1–10, 2019.

Erika Skoe and Nina Kraus. Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing*, 31(3):302, 2010.

Jennifer Krizman and Nina Kraus. Analyzing the ffr: A tutorial for decoding the richness of auditory function. *Hearing research*, 382:107779, 2019.

Ananthanarayan Krishnan, Yisheng Xu, Jackson Gandour, and Peter Cariani. Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, 25(1):161–168, 2005.

Jayaganesh Swaminathan, Ananthanarayan Krishnan, and Jackson T Gandour. Pitch encoding in speech and nonspeech contexts in the human auditory brainstem. *Neuroreport*, 19(11):1163, 2008.

Gabriella Musacchia, Mikko Sams, Erika Skoe, and Nina Kraus. Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences*, 104(40): 15894–15898, 2007.

Patrick Wong, Erika Skoe, Nicole M Russo, Tasha Dees, and Nina Kraus. Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, 10(4):420–422, 2007.

Judy H Song, Erika Skoe, Patrick CM Wong, and Nina Kraus. Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of cognitive neuroscience*, 20(10):1892–1902, 2008.

Samuele Carcagno and Christopher J Plack. Subcortical plasticity following perceptual learning in a pitch discrimination task. *Journal of the Association for Research in Otolaryngology*, 12(1):89–100, 2011.

Thomas Hartmann and Nathan Weisz. Auditory cortical generators of the frequency following response are modulated by intermodal attention. *Neuroimage*, 203:116185, 2019.

Emily BJ Coffey, Emilia MG Colagrosso, Alexandre Lehmann, Marc Schönwiesner, and Robert J Zatorre. Individual differences in the frequency-following response: relation to pitch perception. *PloS one*, 11(3):e0152374, 2016.

Zilong Xie, Rachel Reetzke, and Bharath Chandrasekaran. Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of neurophysiology*, 117(3):1409–1424, 2017.

Merav Ahissar, Mor Nahum, Israel Nelken, and Shaul Hochstein. Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1515):285–299, 2009.

Nina Kraus. Memory for sound: The beams hypothesis [perspective]. *Hearing research*, 407(10829):108291, 2021.

Ruud Wetzels, Jeroen GW Raaijmakers, Emöke Jakab, and Eric-Jan Wagenmakers. How to quantify support for and against the null hypothesis: A flexible winbugs implementation of a default bayesian t test. *Psychonomic bulletin & review*, 16(4):752–760, 2009.

James W Antony, Luis Piloto, Margaret Wang, Paula Pacheco, Kenneth A Norman, and Ken A Paller. Sleep spindle refractoriness segregates periods of memory reactivation. *Current Biology*, 28(11):1736–1743, 2018.

Conrad Iber, Sonia Ancoli-Israel, Andrew L Chesson, Stuart F Quan, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, volume 1. American academy of sleep medicine Westchester, IL, 2007.

Till Roenneberg, Anna Wirz-Justice, and Martha Merrow. Life between clocks: daily temporal patterns of human chronotypes. *Journal of biological rhythms*, 18(1):80–90, 2003.

Sabrina Turker and Susanne M Reiterer. Brain, musicality and language aptitude: A complex interplay. *Annual Review of Applied Linguistics*, pages 1–13, 2021.

Mireille Besson, Daniele Schön, Sylvain Moreno, Andréia Santos, and Cyrille Magne. Influence of musical expertise and musical training on pitch processing in music and language. *Restorative neurology and neuroscience*, 25(3-4): 399–410, 2007.

Gavin M Bidelman, Ananthanarayan Krishnan, and Jackson T Gandour. Enhanced brainstem encoding predicts musicians' perceptual advantages with pitch. *European Journal of Neuroscience*, 33(3):530–538, 2011.

Carlos Marques, Sylvain Moreno, São Luís Castro, and Mireille Besson. Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, 19(9):1453–1463, 2007.

Gavin M Bidelman, Michael W Weiss, Sylvain Moreno, and Claude Alain. Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *European Journal of Neuroscience*, 40(4):2662–2673, 2014.

EBJ Coffey, SC Herholz, S Scala, and RJ Zatorre. Montreal music history questionnaire: a tool for the assessment of music-related experience in music cognition research. In *The Neurosciences and Music IV: Learning and Memory, Conference. Edinburgh, UK*, 2011.

Scott E Lively, John S Logan, and David B Pisoni. Training japanese listeners to identify english/r/and/l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, 94(3):1242–1255, 1993.

Pieter R Roelfsema, Arjen van Ooyen, and Takeo Watanabe. Perceptual learning rules based on reinforcers and attention. *Trends in cognitive sciences*, 14(2):64–71, 2010.

Han Gyol Yi and Bharath Chandrasekaran. Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. *The Journal of the Acoustical Society of America*, 140(2):1332–1335, 2016.

Bharath Chandrasekaran, Han-Gyol Yi, and W Todd Maddox. Dual-learning systems during speech category learning. *Psychonomic bulletin & review*, 21(2):488–495, 2014.

Pierre A Hallé, Yueh-Chin Chang, and Catherine T Best. Identification and discrimination of mandarin chinese tones by mandarin chinese vs. french listeners. *Journal of phonetics*, 32(3):395–421, 2004.

Wan-Ting Huang, Chang Liu, Qi Dong, and Yun Nan. Categorical perception of lexical tones in mandarin-speaking congenital amusics. *Frontiers in psychology*, 6:829, 2015.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382, 1987.

Paul Boersma et al. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Citeseer, 1993.

Fernando Llanos, Zilong Xie, and Bharath Chandrasekaran. Decoding linguistically-relevant pitch patterns from frequency-following responses using hidden markov models. *The Journal of the Acoustical Society of America*, 141 (5):3702–3702, 2017.

Balazs Aczel, Bence Palfi, Aba Szollosi, Marton Kovacs, Barnabas Szaszi, Peter Szecsi, Mark Zrubka, Quentin F Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1(3):357–366, 2018.

Eric-Jan Wagenmakers, Richard D Morey, and Michael D Lee. Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3):169–176, 2016.

Maarten Marsman and Eric-Jan Wagenmakers. Bayesian benefits with jasp. *European Journal of Developmental Psychology*, 14(5):545–555, 2017.

Riko Kelter. Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to bayesian inference with jasp. *BMC medical research methodology*, 20:1–12, 2020.

Michael D Lee and Eric-Jan Wagenmakers. *Bayesian cognitive modeling: A practical course*. Cambridge university press, 2014.

JASP Team. JASP (Version 0.14.3)[Computer software], 2021. URL `https://jasp-stats.org/`.

Emily BJ Coffey, Isabelle Arseneau-Bruneau, Xiaochen Zhang, Sylvain Baillet, and Robert J Zatorre. Oscillatory entrainment of the frequency-following response in auditory cortical and subcortical structures. *Journal of Neuroscience*, 41(18):4073–4087, 2021.

Monika Schönauer. Sleep spindles: timed for memory consolidation. *Current Biology*, 28(11):R656–R658, 2018.

Niels Niethard, Hong-Viet V Ngo, Ingrid Ehrlich, and Jan Born. Cortical circuit activity underlying sleep slow oscillations and spindles. *Proceedings of the National Academy of Sciences*, 115(39):E9220–E9229, 2018.

Daisuke Miyamoto, Daichi Hirai, CCA Fung, Ayumu Inutsuka, Maya Odagawa, Takayuki Suzuki, Roman Boehringer, Chinnakkaruppan Adaikkan, Chie Matsubara, Norio Matsuki, et al. Top-down cortical input during nrem sleep consolidates perceptual memory. *Science*, 352(6291):1315–1318, 2016.

Samuel Laventure, Stuart Fogel, Ovidiu Lungu, Geneviève Albouy, Pénélope Sévigny-Dupont, Catherine Vien, Chadi Sayour, Julie Carrier, Habib Benali, and Julien Doyon. Nrem2 and sleep spindles are instrumental to the consolidation of motor sequence memories. *PLoS Biology*, 14(3):e1002429, 2016.

Maren Jasmin Cordi and Björn Rasch. How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology*, 67:1–7, 2021.

Ananthanarayan Krishnan, Jayaganesh Swaminathan, and Jackson T Gandour. Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *Journal of cognitive neuroscience*, 21(6):1092–1105, 2009.

Ananthanarayan Krishnan, Jackson T Gandour, Christopher J Smalt, and Gavin M Bidelman. Language-dependent pitch encoding advantage in the brainstem is not limited to acceleration rates that occur in natural speech. *Brain and language*, 114(3):193–198, 2010.

Jonathan E Peelle. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and hearing*, 39(2):204, 2018.

Melissa L Caras and Dan H Sanes. Top-down modulation of sensory cortex gates perceptual learning. *Proceedings of the National Academy of Sciences*, 114(37):9972–9977, 2017.

Yisheng Xu, Ananthanarayan Krishnan, and Jackson T Gandour. Specificity of experience-dependent pitch representation in the brainstem. *Neuroreport*, 17(15):1601–1605, 2006b.

Xiaolin Li, Xiaochen Zhang, and Qin Gong. Evidence of both brainstem and auditory cortex involvement in categorical perception for chinese lexical tones. *NeuroReport*, 31(4):359–364, 2020.

Giorgio Paulon, Fernando Llanos, Bharath Chandrasekaran, and Abhra Sarkar. Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults. *Journal of the American Statistical Association*, 116(535): 1114–1127, 2021.

In-Soo Lee, Wayne A Bardwell, Sonia Ancoli-Israel, and Joel E Dimsdale. Number of lapses during the psychomotor vigilance task as an objective measure of fatigue. *Journal of clinical sleep medicine*, 6(2):163–168, 2010.

Thomas J Balkin, Paul D Bliese, Gregory Belenky, Helen Sing, David R Thorne, Maria Thomas, Daniel P Redmond, Michael Russo, and Nancy J Wesensten. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *Journal of sleep research*, 13(3):219–227, 2004.

Richard Boyce, Stephen D Glasgow, Sylvain Williams, and Antoine Adamantidis. Causal evidence for the role of rem sleep theta rhythm in contextual memory consolidation. *Science*, 352(6287):812–816, 2016.

Tal Savion-Lemieux and Virginia B Penhune. The effects of practice and delay on motor skill learning and retention. *Experimental brain research*, 161(4):423–431, 2005.