Bilingualism: Language and Cognition

cambridge.org/bil

L2 difficulties in the perception of Mandarin tones: Phonological universals or domain-general aptitude?

Chao Zhou and João Veríssimo

Center of Linguistics, School of Arts and Humanities, University of Lisbon, Lisboa, Portugal

Research Article

Cite this article: Zhou, C. and Veríssimo, J. (2025). L2 difficulties in the perception of Mandarin tones: Phonological universals or domain-general aptitude? *Bilingualism: Language and Cognition* 1–15. https://doi.org/10.1017/S1366728925100114

Received: 25 February 2024
Revised: 23 April 2025
Accepted: 13 May 2025

Keywords:
L2 tonal acquisition; phonological universals; OCP; tonal markedness scale; pitch acuity; domain-general auditory processing

Corresponding author:
João Veríssimo;
Email: jlverissimo@edu.ulisboa.pt

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

What makes lexical tones challenging for second language (L2) learners? Several recent studies suggest that two phonological universals, the Obligatory Contour Principle and the Tonal Markedness Scale, may constrain the L2 acquisition of Mandarin lexical tones, regardless of learners' first language. We assessed the role of these universals in L2 tonal acquisition by learners from a non-tonal background (L1 Portuguese). We implemented a perceptual testing protocol, which contained a number of methodological and analytical improvements relative to previous studies, including the use of Bayesian mixed-effects models to assess evidence for null hypotheses. The results provided evidence for the null effects of both phonological universals. Instead, a clear determinant of tonal identification accuracy was the participants' pitch acuity, suggesting that domain-general auditory processing underlies the learning of L2 phonological categories. All materials, data and code are publicly available in the OSF repository at https://osf.io/ezadw.

Highlights

- We tested the role of the OCP and TMS universals in L2 Mandarin tonal perception.
- Bayesian analyses supported the null effects of the two phonological universals.
- Learners' lexical knowledge of Chinese characters did not predict L2 tonal development.
- Domain-general pitch acuity was associated with more accurate L2 tonal perception.

1. Introduction

Among studies of L2 Mandarin acquisition, lexical tones are arguably the most studied structures. Tones are essential for distinguishing lexical meanings in Mandarin, which makes them one of the primary linguistic components for learners from the very beginning. While it is widely acknowledged that Mandarin Chinese lexical tones pose a serious challenge for speakers from a non-tonal background, previous research has shown that, given enough time and training, many learners are able to identify and produce tone categories fairly well (Lee, Tao, & Bond, 2009; Pelzl, Lau, Guo, & DeKeyser, 2019; H. Zhang, 2007, 2016). However, mastering individual tone categories (i.e., on isolated monosyllables) does not ensure good performance with tones within words, which are predominantly disyllabic in Mandarin (Duanmu, 2007; Wen, Qiu, Leong, & Van Heuven, 2023). Several recent studies suggest that learners' increased difficulties with tone sequences are influenced by two phonological universals, the Obligatory Contour Principle (OCP, Leben, 1973) and the Tonal Markedness Scale (TMS, Hyman & VanBik, 2004; Ohala, 1978). Expanding upon prior research, the present study assesses the role of the OCP and the TMS in L2 tonal acquisition with a perceptual testing protocol, which allowed us to better disentangle the potential effects of phonological universals from those of confounding variables, such as articulatory difficulty and lexical effects.

1.1. Phonetics, phonology and L2 acquisition of Mandarin tones

The primary acoustic correlate of tones is the fundamental frequency (F0). Higher F0 leads to a higher perceived pitch and all spoken languages employ pitch differences linguistically. When spreading over intonational phrases, pitch indicates different sentence types (e.g., in English, rising intonation for yes–no questions, falling intonation for statements). When carried by individual syllables or words, as in Mandarin, pitch displays as tone, signalling lexical contrasts (see Gussenhoven & Chen, 2020, for a recent overview on the pitch from a cross-linguistic perspective).

The four lexical tones in Mandarin are traditionally denoted as Tone 1 (T1) to Tone 4 (T4). By referring to pitch height (high/low) and pitch movement/contour (rising, falling, or dipping),

these tonal categories can be classified as high-level (T1), rising (T2), low-level (or low-dipping) (T3)[1] and falling (T4).

Research on L2 acquisition of Mandarin tones has been primarily concerned with how lexical tones are acquired in isolated syllables (see Pelzl, 2019, for a recent overview). It has been shown that Mandarin lexical tones are very challenging for novice learners, especially for those from a non-tonal background (e.g., So & Best, 2010), but tones can be mastered after adequate training and enough time. For instance, L2 tonal difficulties can be mitigated rapidly with a few phonetic training sessions (Silpachai, 2020; e.g., Wang, Spence, Jongman, & Sereno, 1999) and advanced learners may even achieve near-native perceptual (Lee et al., 2009; Pelzl et al., 2019; Zou, Chen, & Caspers, 2017) and production performance (Song, 2021; H. Zhang, 2007, 2016) in monosyllabic contexts.

Given that around 70% of the words in Mandarin are disyllabic (Duanmu, 2007; Wen et al., 2023), it may come as a surprise that few studies have explored how L2 learners handle tone sequences. Such a research gap might be rooted in the intuition that mastering isolated tones will lead to fairly good performance in identifying tone sequences in disyllabic words, as individual tones can be regarded as the building blocks of tonal sequences. However, several studies demonstrate that compared to individual tones, disyllabic tone sequences are generally more challenging for L2 learners (Chang & Bowles, 2015; Hao, 2012; Pelzl et al., 2019; Silpachai, 2020). The increased difficulty with tone sequences may be attributed to several factors that all play a larger role in sequences than in individual tones: a higher short-term memory load; tone sandhi (e.g., when a T3 precedes another T3, the first one becomes a T2, T3-T3 → T2-T3, Huang & Johnson, 2011); coarticulatory influence (e.g., T1 ends high in isolation but it surfaces as falling before T2, Chang & Bowles, 2015); and the potential need for more extensive lexical access (given that the Mandarin lexicon is largely disyllabic). Moreover, another line of research suggests that learning disyllabic tone sequences is further influenced by some universal processes common to language learning and human language more generally (Song, 2021; H. Zhang, 2007, 2016). In the following subsections, we provide a concise introduction to the two phonological principles that have been objects of inquiry in prior research, the OCP and the TMS, and review the evidence for their active role in L2 tonal acquisition.

1.2. Phonological universals in L2 tonal acquisition

1.2.1. Obligatory contour principle

The restriction on co-occurrence of identical or homorganic sound structures has long been known to phonologists since Greenberg (1950). Similar observations on the Mende tonal system, where no adjacent identical tones are allowed (e.g., *HHL and *LLH), have led Leben (1973) to propose the very first version of OCP: when two identical tones occur on adjacent vowels, the rightmost one is deleted. Subsequent studies have demonstrated that the most general form of OCP (disallowing adjacent identical elements, without referring to which structural property is targeted) leaves unexplained surface counterexamples in many languages (e.g., Goldsmith, 1976; Odden, 1986, 1988). As noted in Boersma (1998), in order to maintain the belief that the OCP is a phonological universal, one has to conceptualise it as a violable constraint (e.g., under the framework of Optimality Theory, Prince & Smolensky, 2004) or relativise it to certain phonological structures in a language-specific manner.

Besides assessing how well the OCP serves as a tool for formal phonological analysis, another line of research that has received much attention concerns its psychological reality. Studies with native listeners of Hebrew (Berent & Shimron, 1997), Arabic (Frisch & Zawaydeh, 2001) and English (Coetzee, 2005, 2008) have consistently reported that nonwords containing OCP-violating sequences are judged to be less well-formed than nonwords that conform to the OCP. In lexical decision tasks, OCP-violating nonwords are rejected faster than control items by native listeners of Hebrew (Berent, Everett, & Shimron, 2001), Arabic (Frisch & Zawaydeh, 2001) and Dutch (Shatzman & Kager, 2007). Dutch listeners may use their language-specific restriction on the co-occurrence of multiple labial consonants (OCP-Labial, e.g., */spVp/) as a cue for speech segmentation (Boll-Avetisyan & Kager, 2014). More recently, Gong (2022) found that a phonotactic pattern conforming to the OCP was easier to learn in artificial language learning experiments, compared to a place harmony (an anti-OCP) process and to an arbitrary pattern.

Further evidence supporting the OCP as a universal phonological principle comes from large-scale quantitative typological reports, which show that nearly all languages are restricted by some kind of similarity avoidance (Graff, 2012; Mayer, Rohrdantz, & Plank, 2010; Pozdniakov & Segerer, 2007).

1.2.2. Tonal markedness scale

Phonetically speaking, some tones are intrinsically more complex than others. Level tones mainly involve F0 height, while contour tones additionally entail an F0 slope. Although rising and falling tones are both contour tones, they appear to show different degrees of phonetic complexity. Ohala (1973) observed in a production experiment that, for a given pitch interval, a rising tone systematically takes longer to produce than a falling tone. Falling tones may thus be seen as phonetically less demanding, since they need less time to reach a certain level of prominence than rising tones (Ohala, 1978; Xu, 2002; Xu & Sun, 2002). These phonetically-grounded differences in tonal complexity were later formalised by Hyman and VanBik (2004) as the TMS (*Rising >> *Falling >> *Level), which was used to explain the tone sandhi pattern in Hakha Lai.

Patterns consistent with the TMS have been attested in first language acquisition and typological studies. In a longitudinal study with four Mandarin-acquiring toddlers, Hua and Dodd (2000) reported that the stabilisation of lexical tones generally followed the TMS (i.e., T2 < T4 < T1). Moreover, in a typological survey by J. Zhang (2002, 2004), the tone distribution of 187 genetically diverse languages corroborated the implicational relationship predicted by the TMS: if a language has contour tones, it also employs level tones; and if it has rising tones, it also employs falling tones.

1.3. The presence of OCP and TMS in L2 tonal acquisition

The effects of OCP and the TMS have been reported in several studies on L2 tonal production. In a reading task of Chinese real words, H. Zhang (2007) observed that L1-English learners, who had studied Mandarin for about 5 months, tended to avoid producing identical tones on adjacent syllables, especially in the case of contour tone pairs (i.e., T2-T2 and T4-T4). Such tonal dissimilation can hardly be attributed to the learners' L1 English, which does not

[1]Although T3 is realised as a low dipping tone (tone value [214]) in a prosodically-prominent position, such as in isolated syllables and prosodic-final positions, it is most often a low-level tone (tone value [21] or [11]). Which allotone corresponds to the underlying form of T3 is an ongoing debate (Duanmu, 2007; Yip, 2002; J. Zhang, 2014) that goes beyond the scope of this study, since T3 was not directly examined.

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

contain any lexical tones, or to the target language Mandarin, where only T3-T3 sequences are avoided. Under the OT-theoretic framework, H. Zhang speculated that the OCP, as part of the learners' Universal Grammar, was responsible for the observed L2 tonal dissimilation. Moreover, after splitting tone sequences into individual tones, H. Zhang observed that the learners' production accuracies corroborated the TMS (i.e., T2 < T4 < T1).

In a subsequent study, H. Zhang (2016) examined L2 tonal production by intermediate learners of Mandarin with three different L1s (English, Japanese and Korean). This study constituted an insightful test of phonological universals, because if L2 phonology was subject to the OCP, the dispreference of identical tone sequences would be attested, regardless of the learners' L1. H. Zhang (2016) found two pieces of evidence suggesting OCP effects. First, for all three learner groups, target identical tone sequences were replaced by non-identical ones more frequently than the reverse direction. Second, identical tone pairs, when taken together, were produced less often than what would be expected given their target proportions. However, the OCP only affected rising and falling tone pairs (T2-T2 and T4-T4), not level pairs (T1-T1). In addition to the OCP effect, H. Zhang (2016) further observed that the production accuracy of both identical tone sequences (T2-T2, T4-T4, T1-T1) and individual tones (T2, T4, T1) followed the TMS (i.e., T2 < T4 < T1).

Following the research direction of H. Zhang (2007, 2016), Song (2021) investigated the OCP and the TMS in the spontaneous YouTube speech of four L1-English learners with near-native Mandarin proficiency. Song observed that the production accuracy of individual tones (i.e., T2 < T4 < T1), but not that of tone sequences (i.e., T2-T2 > T4-T4), corroborated the TMS. The OCP effect, defined as lower accuracy for identical than non-identical sequences, was only obtained for T4-T4.

To summarise, there is some evidence for OCP effects in L2 tonal production, but they might be tone-specific. As for the TMS, it has been consistently verified on the production of individual tones, whereas conflicting evidence exists for its operation on tonal pairs. Given that the aforementioned three studies examined learners at different stages of L2 Mandarin learning, the divergence between their findings raises the possibility that the effects of the phonological universals may change over the course of L2 development (e.g., Major, 2001). In particular, given that the OCP (except on T3-T3) and the TMS are both in conflict with the target Mandarin phonology, one may expect that learners with more advanced Mandarin proficiency are better at overcoming the interference of these two universals. Extending prior research, the current study assesses whether the effects of the phonological universals are modulated by learners' L2 speech proficiency. In the following section, we review two measures that have been shown to be good predictors of L2 Mandarin speech proficiency.

1.4. Predictors of L2 speech proficiency

Decades of research have led to a consensus that L2 learning and processing exhibit significant individual differences (Hulstijn, 2012; Sandlund, Sundqvist, & Nyroos, 2016). Two predictors of L2 speech proficiency, in particular, are a learner's vocabulary size and their auditory processing abilities.

With regard to vocabulary size, it is clear that a solid mastery of lexical knowledge is a prerequisite for effective language use, such that as L2 lexical competence increases, so too does general L2 proficiency (Meara, 1996; Zhou & Li, 2021). In the case of L2 speech development, prior research indicates that learners with larger vocabularies can better distinguish between confusable L2 sound categories (Bundgaard-Nielsen, Best, Kroos, & Tyler, 2012; Bundgaard-Nielsen, Best, & Tyler, 2011; Daidone & Darcy, 2021; Llompart, 2021). This is because learning phonological neighbours may contribute to the consolidation and refinement of existing phonological representations in the lexicon. Furthermore, learners' vocabulary size has been shown to play a critical role in L2 phonotactic learning (Spinelli, Forti, & Jared, 2021), consistent with the view that phonotactic generalisations are made across lexical entries.

In addition to lexical knowledge, the rate of success in L2 speech acquisition has been demonstrated to be closely associated with individual differences in auditory acuity, that is, the ability to detect subtle differences in various aspects of acoustic input at a fine-grained level (Auditory Precision Hypothesis - L2, for an overview, see Saito, 2023). This is presumably because learners with better auditory acuity can increase the precision of their auditory representations, which contributes to the increase of L2 speech proficiency. Auditory acuity can be estimated in a global manner (e.g., Kachlicka, Saito, & Tierney, 2019; Saito, Sun, et al., 2022), namely as the average of duration, pitch, and formant discrimination scores, or in a dimension-specific way. In a recent study, Saito, Sun, et al. (2022) assessed the role of perceptual acuity in L1-Japanese learners' perception of English lateral-rhotic contrast, which is mainly cued by the F3 formant difference. They found that this notoriously difficult English liquid contrast was discriminated more accurately by L1-Japanese learners with more precise processing of the F3 formant. This finding suggests that dimension-specific auditory acuity is strongly linked to the acquisition of phonological contrasts that are robustly distinguished by that particular dimension. By extension, we expect that learners with good pitch acuity show advantages in the acquisition of Mandarin lexical tones.

1.5. The present study

The goal of the current study is to further assess the role of two phonological universals, the OCP and the TMS, in L2 tonal acquisition via a perceptual testing protocol with a group of learners of Mandarin, whose L1 European Portuguese does not employ pitch at the lexical level. Unlike the tasks employed in previous studies, our experiment was designed to tap into pre-lexical perception, to better disentangle the influence of phonological universals from other confounding factors, namely, the quality of phono-lexical representation and articulatory difficulty.

First, given that the underlying representations of Mandarin lexical tones may be fuzzy even for advanced learners (Pelzl et al., 2019, 2021a), it is unclear to what extent the prior evidence for phonological universals, which was obtained on the basis of real-word production accuracy (Song, 2021; H. Zhang, 2007, 2016), can be attributed to inaccurate lexical encoding. To give an example, learners' production of target /T4-T4/ as [T1-T4], which was interpreted as an OCP effect, may stem from the fact that learners have wrongly represented disyllabic words carrying two falling tones as /T1-T4/ in their L2 lexicon.

Second, deviations in L2 tonal production may result from articulatory imprecision. It has been shown that, even after establishing distinct tone categories, as revealed by good perceptual discrimination or accurate identification, L2 learners may still not be able to produce lexical tones very well (Elliot, 1991; Nagano-Madsen & Wan, 2017), suggesting that L2 difficulties with Mandarin tones might be articulatorily motivated. This is because, in production, learners need to not only construct the distinct abstract tone categories, but also learn how to phonetically implement them

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

through the corresponding articulatory gestures (e.g., controlling the timing of F0 movement), which can be quite challenging for learners from a non-tonal background.

The presence of these two confounding factors casts some doubt on previous real-word production studies, which have, nevertheless, attributed their findings to phonological universals exclusively (Song, 2021; H. Zhang, 2007, 2016). To address this issue, the current study employs a perceptual identification task with pseudowords, thereby precluding the potential influence of imprecise lexical encoding and articulation. Given that the OCP and the TMS are primarily conceptualised as restrictions on surface phonological forms (Song, 2021; e.g., H. Zhang, 2007, 2016)—which are the outputs of pre-lexical perception—we posit that a pseudoword identification task constitutes a suitable tool for examining the role of these two phonological universals in L2 tonal acquisition.

Apart from directly tapping into the representational level where the OCP and the TMS arguably operate, our experimental design displayed several other methodological improvements. In particular, we have made use of a much larger number of participants and stimuli than those in previous studies: H. Zhang (2007, 2016) employed two test items for each tone pair (each produced twice), and Song (2021) examined the productions of four L2 learners, whereas we have tested 59 participants on 96 tonal syllables (see Method). Moreover, the same set of tonal syllables were used in the identical and non-identical conditions, allowing for a more stringent test of the OCP, while the tone pairs we presented (T2-T2, T4-T4 and T1-T1) were well-matched in various lexical statistics to appropriately test for the TMS.

The current study seeks to answer the following research questions:

1. Is L2 tonal perception subject to OCP effects? We expected to find OCP effects in L2 tonal perception, at least for some tone pairs (e.g., T4-T4, Song, 2021). OCP effects would be borne out if the accuracy rate in the identical condition was lower than in the non-identical condition, because the OCP should force identical tone pairs into non-identical ones, but not vice versa.
2. Is L2 tonal perception subject to TMS effects? We hypothesised that learners' identification accuracy of tone pairs (H. Zhang, 2016) and that of individual tones (Song, 2021; H. Zhang, 2016) would both conform to the TMS (i.e., T2 < T4 < T1).
3. Are the effects of the OCP and the TMS modulated by L2 speech proficiency? We predicted that the effects of OCP and TMS would be more pronounced in learners with lower Mandarin proficiency. Given that both the OCP (except for T3-T3) and the TMS are not active in Mandarin phonology, an increase in L2 speech proficiency should lead learners to gradually overcome the influences exerted by these two universals.

2. Method

2.1. Participants

Fifty-nine native speakers of European Portuguese (48 women, mean age = 21.63 years, SD = 2.51) were recruited at the University of Minho, Portugal, where they were enrolled in degree programmes in Chinese language and culture. Twenty-four participants had studied in a formal classroom setting for 1 year, 32 for 2 years, and 3 for 3 years. None had studied Mandarin prior to entering university (all had an age of onset of acquisition >17 years old), nor lived in Mandarin-speaking countries. All participants reported having normal speech and hearing. No participant reported being fluent in, nor regularly using a tonal language other than Mandarin.

Participants' Mandarin speech proficiency was assessed with the LEXTALE_CH vocabulary test (Chan & Chang, 2018) and with a pitch discrimination task from the auditory processing test batteries developed by Saito, Sun, et al. (2022). The participants of this study are quite heterogeneous in terms of their Mandarin speech proficiency, as revealed by their LEXTALE_CH scores (range: −17 to 47, SD = 10.72) and by their pitch acuity scores (range: 3.71–85.23, SD = 16.96). The two test scores were only weakly (and non-significantly) correlated ($r = -0.16$, $t = -1.20$, $p = .235$), which suggests that these two tests essentially measure different constructs of L2 Mandarin proficiency.

2.2. Materials

Ninety-six Mandarin tonal syllables were selected from the lexical database DoWLS-MAN (Neergaard, Xu, German, & Huang, 2022), to create 48 disyllabic pseudoword items carrying two identical lexical tones (Identical condition: 16 × T1-T1, 16 × T2-T2, 16 × T4-T4). The same set of syllables was also used to create another set of 48 disyllabic items with distinct tones (Non-identical condition: 8 × T1-T2, 8 × T2-T1, 8 × T1-T4, 8 × T4-T1, 8 × T2-T4, 8 × T4-T2). None of the experimental stimuli included low-level tones (T3), because in Mandarin /T3-T3/ undergoes tone sandhi to become [T2-T3], and thus it is not possible to create disyllabic stimuli with identical T3 tones.

The syllables that were used to create the pseudowords were matched as close as possible across the three tones (i.e., T2, T4, T1) in mean (phonological) frequency, mean (phonological) neighbourhood density, and mean homophone density, as shown in Table 1. The matching minimises the possibility that any observed difference between experimental conditions would stem from learners' lexical knowledge (e.g., participants could be more familiar with the monosyllabic stimuli in T2-T2 than those in T4-T4).

A female native speaker of Mandarin was recorded reading the stimulus list in a sound-attenuated room at the University of Minho in Portugal, using a Zoom H4n pro recorder, and a Shure SM58 microphone, at an audio sampling rate of 44.1 kHz. All recorded sound files were adjusted to the average intensity of 70 dB in Praat 6.1.05 (Boersma & Weenink, 2022). The speaker was a Mandarin instructor and was uninformed as to the nature of the research. Four repetitions of the stimulus list were recorded and, for each pseudoword, we selected the token that was produced more naturally and without creaky phonation, without neutralizing durational differences (see Table 1 for the mean durations of each tone in each syllabic position).

Table 1. Summary of syllable characteristics (means and SDs) in each of the three identical tone pairs

Tone	Frequency	Neighbour dens.	Homophone dens.	Duration (1st)	Duration (2nd)
T2	4.52 (0.40)	14.38 (4.43)	5.78 (4.01)	400 (54)	482 (57)
T4	4.58 (0.31)	18.16 (6.29)	5.91 (2.87)	385 (56)	385 (49)
T1	4.54 (0.34)	18.53 (5.67)	5.72 (3.54)	357 (65)	424 (50)

Note: All metrics were obtained from the DoWLS-MAN database (Neergaard et al., 2022) as phonological metrics. Frequency is log10 of number of occurrences in a corpus of 46.8 million characters (Cai & Brysbaert, 2010). Durations for the first and second syllabic positions are expressed in milliseconds.

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

2.3. Procedure

A perceptual identification task was created and hosted using Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Participants performed the task in a quiet room at the University of Minho. The test trials were presented in a single block in a self-paced task. Stimuli were presented over headphones at a comfortable listening level and were randomised in different orders for each participant. Prior to the task, participants were told that they were going to hear a set of disyllabic Mandarin-like pseudowords, one at a time. Their task was to identify the two lexical tones in each stimulus, by typing their answers into a single text box using the keys 1, 2, 3 and 4, representing the four Mandarin tones. Participants were explicitly told that the stimuli were not real words in Mandarin and that the neutral tone would not occur. Note that participants were not informed that T3 would never occur in the task, because excluding T3 from possible responses could increase the accuracy rate of T2, which is often confused with T3 by L2 learners (e.g., Pelzl et al., 2019).

At the beginning of the session, participants responded to a background questionnaire and signed a consent form. After the perceptual identification task, participants performed the LEXTALE_CH vocabulary test and the pitch discrimination task (see Participants). The whole session took approximately 30 minutes to complete.

2.4. Data analysis

One participant performed the task twice (as revealed by the background questionnaire), so her second set of responses was removed from the dataset. No other participants, items, or observations were excluded.

One-character responses (e.g., "2"; 1.09%) were considered incorrect both when analysing tone pairs and individual tones. Responses with only one valid digit within a two-character sequence (e.g., "2?"; 0.16%) were considered incorrect in the analysis of tone pairs but were assessed for correctness in the analysis of individual tones. Responses with no valid digits (e.g., "??"; 0.018% of the total) were considered incorrect in all analyses.

We present three different analyses below, each tailored to a research question of interest. The first analysis assessed the OCP by comparing the identification accuracies of identical versus non-identical tone pairs. The second analysis assessed the TMS by comparing, within identical tone pairs, the identification accuracies of rising (T2-T2), falling (T4-T4) and level (T1-T1) tones. Additionally, the TMS was assessed at the level of individual tones by comparing their identification accuracy in each of the two syllabic positions. Finally, the third analysis investigated the effect of our two L2 speech proficiency measures (i.e., vocabulary size and pitch acuity) on participants' accuracies and assessed whether the OCP and TMS effects were modulated by L2 proficiency.

All analyses made use of mixed-effects binomial (logistic) regression, which is recommended for the analysis of binary data, such as correct versus incorrect responses (Jaeger, 2008; Quené & van den Bergh, 2008; Veríssimo & Clahsen, 2014). Each statistical model contained the appropriate fixed-effect predictors dictated by the research question of interest. Random effects were additionally included to capture variation across participants and items. In order to prevent overconfident results, all models employed a 'maximal' random-effects structure, that is, they included all random slopes that were allowed by the experimental design (Barr, Levy, Scheepers, & Tily, 2013; Oberauer, 2022).

The statistical models were fit in a Bayesian framework. Bayesian analyses combine prior information with evidence from the data to produce a *posterior distribution* for each parameter, which is a probability distribution over a parameter's possible values (for introductions to Bayesian statistics, see Vasishth, Nicenboim, Beckman, Li, & Kong, 2018; Veríssimo, 2025). In the analyses below, we present the posterior distributions for every effect of interest, accompanied by their means and 95% 'credible intervals', which is the range within which a parameter lies with 95% probability.

In addition, Bayesian models allow for conducting hypothesis tests for effects of interest. Bayesian hypothesis testing is based on comparing an 'alternative model' that includes the effect of interest, to a 'null model', which does not. The evidence that the data provides in favour of one versus the other model is the 'Bayes factor' and allows inferring about the existence of an effect (e.g., Schad, Nicenboim, Bürkner, Betancourt, & Vasishth, 2022; Schmalz, Biurrun Manresa, & Zhang, 2023). Importantly, and in contrast to frequentist analyses, Bayes factors *can* provide support for the null hypothesis, that is, for the equality between conditions or groups (Dienes & Mclatchie, 2018; Rouder, Speckman, Sun, Morey, & Iverson, 2009). For each effect reported in the current paper, we calculated the natural logarithm of the Bayes factor in favour of the alternative hypothesis ($\ln BF_{10}$, Kass & Raftery, 1995) using the Savage-Dickey method (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Values of $\ln BF_{10}$ greater than 1 support the hypothesis that an effect is different from zero (H1), while negative values smaller than -1 support the hypothesis that the effect is absent (H0); values of $\ln BF_{10}$ between -1 and 1 are essentially inconclusive (Kass & Raftery, 1995; Veríssimo, 2025).

Bayesian models require specifying prior distributions on model parameters. We employed weakly informative priors on all parameters. In accordance with various recommendations (Gelman, Jakulin, Pittau, & Su, 2008; Ghosh, Li, & Mitra, 2018; McElreath, 2020; Vasishth et al., 2018), priors on fixed effects were normally distributed with mean 0 and SD 2.5, priors on random effects were exponentially distributed with rate 1, and priors on random correlations were LKJ-distributed with shape 2. The appropriateness of these priors was confirmed through prior predictive checks. Additionally, we have conducted prior sensitivity analyses (e.g., Schad, Betancourt, & Vasishth, 2021; Sinharay & Stern, 2002) for the most important effects in this paper (see Appendix S1).

Analyses were performed with the *brms* package in R (Bürkner, 2017; R Core Team, 2020). The procedures for fitting Bayesian models and assessing their convergence followed recent recommendations (Schad et al., 2021; Vasishth et al., 2018; Veríssimo, 2025).

3. Results

3.1. Comparison between identical vs. non-identical pairs

To assess the OCP, we compared accuracy proportions in identical versus non-identical tone pairs (identical: T1-T1, T2-T2, T4-T4; non-identical: T1-T2, T2-T1, T1-T4, T4-T1, T2-T4, T4-T2). The mixed-effects binomial model included condition as a fixed effect (coded with sum contrasts, -0.5 = 'non-identical', 0.5 = 'identical'). Random effects included random intercepts for participant and syllable pair, as well as by-participant random slopes for condition.

Figure 1A shows the model-based predicted proportions of correct responses in identical and non-identical conditions (empirical proportions averaged across subjects were 47.5% in identical and 45.4% in non-identical conditions). Figure 1B shows the posterior

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

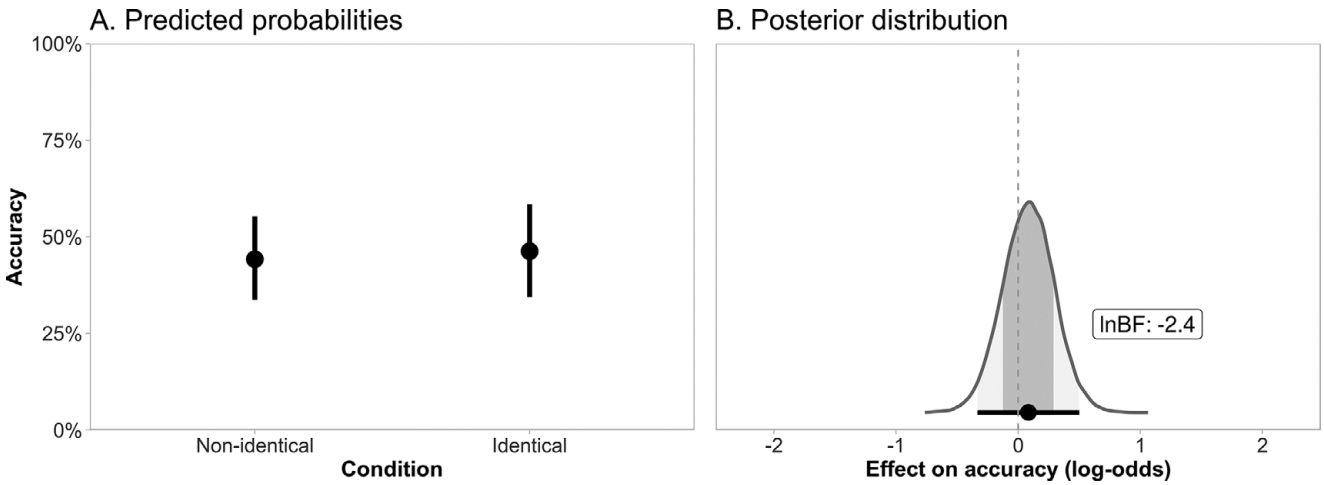

Figure 1. (A) Means (circles) and 95% credible intervals (vertical bars) of predicted proportions of correct responses in identical and non-identical conditions. (B) Posterior distribution of the difference between identical and non-identical conditions in the log-odds scale. Shaded areas show 68% and 95% credible intervals. The black circle and horizontal line represent the mean and 95% credible interval. The numeric label is the natural logarithm of the Bayes factor in favour of the alternative hypothesis (values greater than 1 support the existence of an effect and values smaller than −1 support its absence).

distribution for the effect of condition (i.e., for the accuracy difference between identical and non-identical conditions) in the modelled log-odds scale. The difference between conditions was estimated to be very small in magnitude, with the 95% interval spanning both negative and positive values (b = 0.08 [−0.33, 0.50]). Moreover, the Bayes factor analysis (also shown in Figure 1B) showed evidence *against* a difference between conditions, that is, it supported the null hypothesis that tonal perception is equally accurate in identical and non-identical pairs. A prior sensitivity analysis showed that this conclusion held for a range of reasonable priors (see Appendix S1). In sum, the results showed no indication that identical tone pairs were less accurately identified than non-identical pairs, a pattern that would be expected if the OCP had been applied.

3.2. Comparison between tones

To assess the TMS, we compared accuracy proportions between rising (T2-T2), falling (T4-T4) and level (T1-T1) identical tone pairs. The mixed-effects binomial model included tone as a fixed effect, coded with treatment contrasts (with T4-T4 as the reference level). The model's random-effects structure included random intercepts for participant and syllable pair, as well as by-participant random slopes for tone.

Figure 2A shows the model-based predicted proportions of correct responses for the three tone pairs (empirical proportions were 33.4% for T2-T2, 52.5% for T4-T4 and 56.6% for T1-T1). While the accuracy proportions were numerically consistent with the pattern predicted by the TMS (i.e., T2-T2 < T4-T4 < T1-T1), rising pairs were particularly difficult to identify relative to falling and level pairs.

Figure 2B shows the posterior distributions of differences between rising and falling pairs (T2-T2 vs. T4-T4) and between level and falling pairs (T1-T1 vs. T4-T4), accompanied by their corresponding Bayes factors. The results provided strong support for the hypothesis that rising tones are perceived much less accurately than falling tones, as revealed by a large (negative) estimate

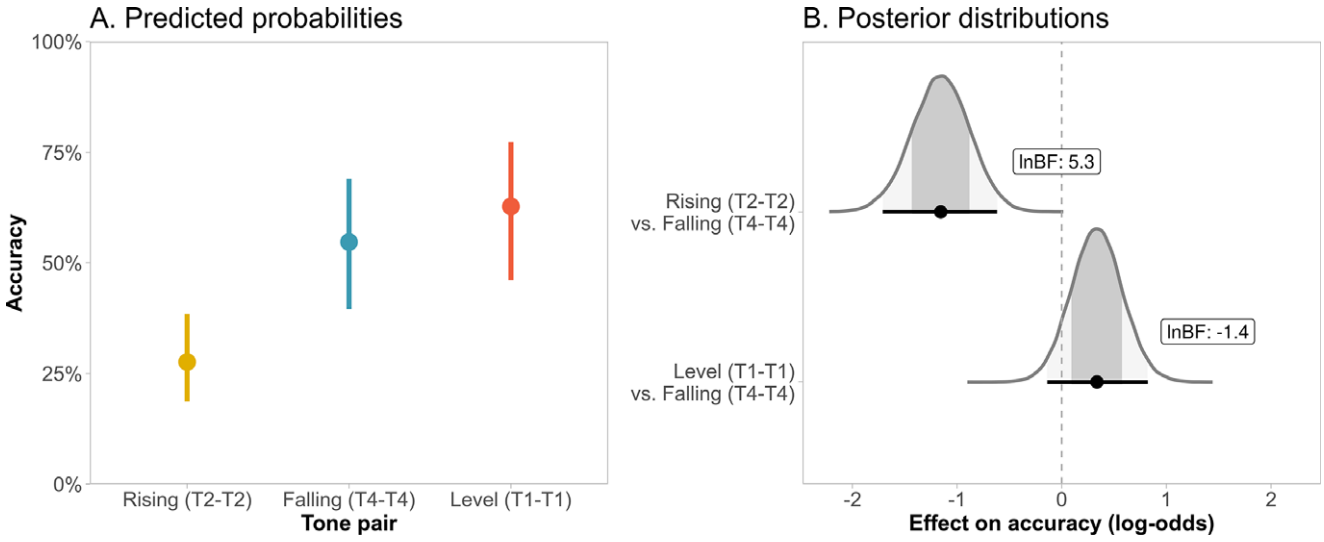

Figure 2. (A) Means (circles) and 95% credible intervals (vertical bars) of predicted proportions of correct responses for the different tone pairs (rising, falling, level). (B) Posterior distributions and natural logarithm of Bayes factors for the differences between tone pairs (rising vs. falling, level vs. falling) (see Figure 1 caption for further details).

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

(b = −1.16 [−1.71, −0.62]) and a large Bayes factor ('very strong' evidence, in the scale of Kass & Raftery, 1995). The sensitivity analysis showed that the same conclusion can be drawn for a range of reasonable priors (see Appendix S1). In contrast, the difference between level and falling tones was estimated to be much smaller (b = 0.34 [−0.14, 0.82]) and the negative Bayes factor for this comparison actually provided evidence for equal accuracy in level and falling tones. With wider priors than employed in our default analyses, the evidence for equality between level and falling tones became stronger, whereas with narrower priors, there was no support for their equality nor for their difference. In sum, our results show that rising pairs (T2-T2) are indeed harder to identify correctly, as predicted by the TMS, but falling tones (T4-T4) cannot be distinguished from level tones (T1-T1) in their identification accuracy.

3.2.1. Comparison between individual tones in each position

We have also investigated whether the individual syllables were subjected to the TMS by comparing the identification accuracies of the different tones in the first and second syllabic positions separately. Figure 3 displays the proportions of responses of each type, averaged across participants, for each of the presented tones (rising, falling, level) and in each syllabic position. Two noteworthy aspects can be gathered from the inspection of the empirical proportions. First, the TMS was clearly violated in the second syllabic position, since falling tones (T4) were more accurately identified than level tones (T1). Second, rising tones (T2) were very frequently misidentified as low-level (T3) in both syllabic positions, despite the fact that the experiment did not include any T3 stimuli (in contrast, T3 responses were almost never produced when T4 and T1 were presented).

To statistically assess the TMS on individual tones, we fitted a mixed-effects binomial model that included tone (rising, falling, or level), position (first or second) and their interaction as fixed effects. The tone was coded with treatment contrasts (with T4 as the reference) and position was coded with nested contrasts, so that separate estimates could be obtained for the differences between tones in each syllabic position (see Schad, Vasishth, Hohenstein, & Kliegl, 2020). Additionally, the model included random effects for participant, syllable pair and individual syllable, as well as all random slopes allowed by the design.

Figure 4A displays the predicted proportions of correct responses for rising (T2), falling (T4) and level (T1) tones in the first and second syllabic positions. The results for the first syllabic position were numerically in line with the TMS (i.e., T2 < T4 < T1; empirical proportions T2: 55.2%, T4: 57.8%, T1: 70.6%). However, in the second position, falling tones were more accurate than level tones (i.e., rising < level < falling; empirical proportions T2: 48.7%, T4: 84.0%, T1: 70.1%). Figure 4B shows the posterior distributions for the differences between rising and falling tones (T2 vs. T4) and between level and falling tones (T1 vs. T4) in the first and second positions, with their corresponding (logged) Bayes factors.

In the first syllabic position, Bayes factors indicated positive evidence for a more accurate identification of level tones than falling tones (i.e., T1 > T4: b = 0.71 [0.29, 1.14]), but equal accuracy for falling and rising tones (i.e., T2 = T4: b = −0.37 [−0.83, 0.08]). In contrast, in the second syllabic position, we have obtained strong evidence that both rising and level tones are identified less accurately than falling tones (i.e., T2 < T4: b = −2.49 [−3.07, −1.95]; T1 < T4: b = −0.96 [−1.50, −0.42]). We have additionally compared accuracies in the first versus second positions for each of the three tones. Whereas falling tones were identified much more accurately in second than in first position (T4: b = 1.90 [1.46, 2.35], $\ln BF_{10}$ = Inf), rising and level tones were equally accurate in both positions (T2: b = −0.22 [−0.55, 0.11], $\ln BF_{10}$ = −2.4; T1: b = 0.23 [−0.16, 0.64], $\ln BF_{10}$ = −2.4). These positional effects were supported by interactions. Specifically, the differences between accuracies on falling versus the other two tones were larger in the second than in the first position (T2 vs. T4: b = 2.12 [1.54, 2.71], $\ln BF_{10}$ = 39.5; T1 vs. T4: b = 1.67 [1.08, 2.25], $\ln BF_{10}$ = 11.7). In contrast, the difference between level and rising tones did not interact with position (b = 0.45 [−0.04, 0.95], $\ln BF_{10}$ = −1.4).

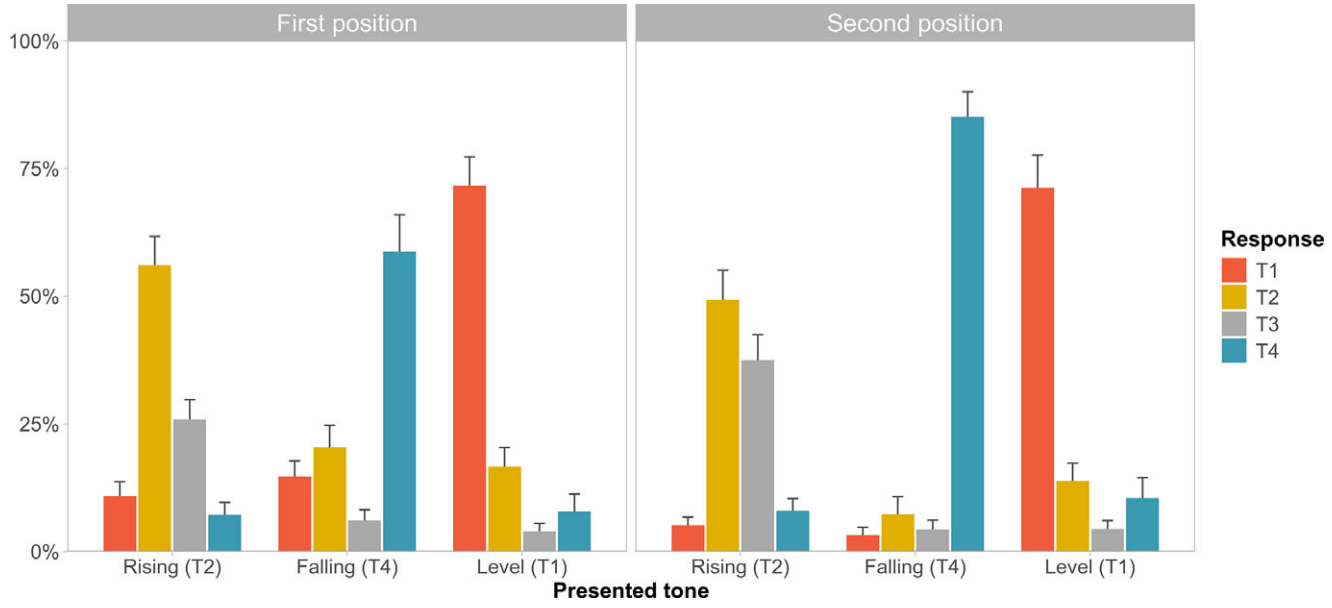

Figure 3. Mean proportions (across participants) of T1, T2, T3 and T4 responses for each of the presented tones (rising, falling, level) and in each syllabic position. Error bars show 95% confidence intervals.

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

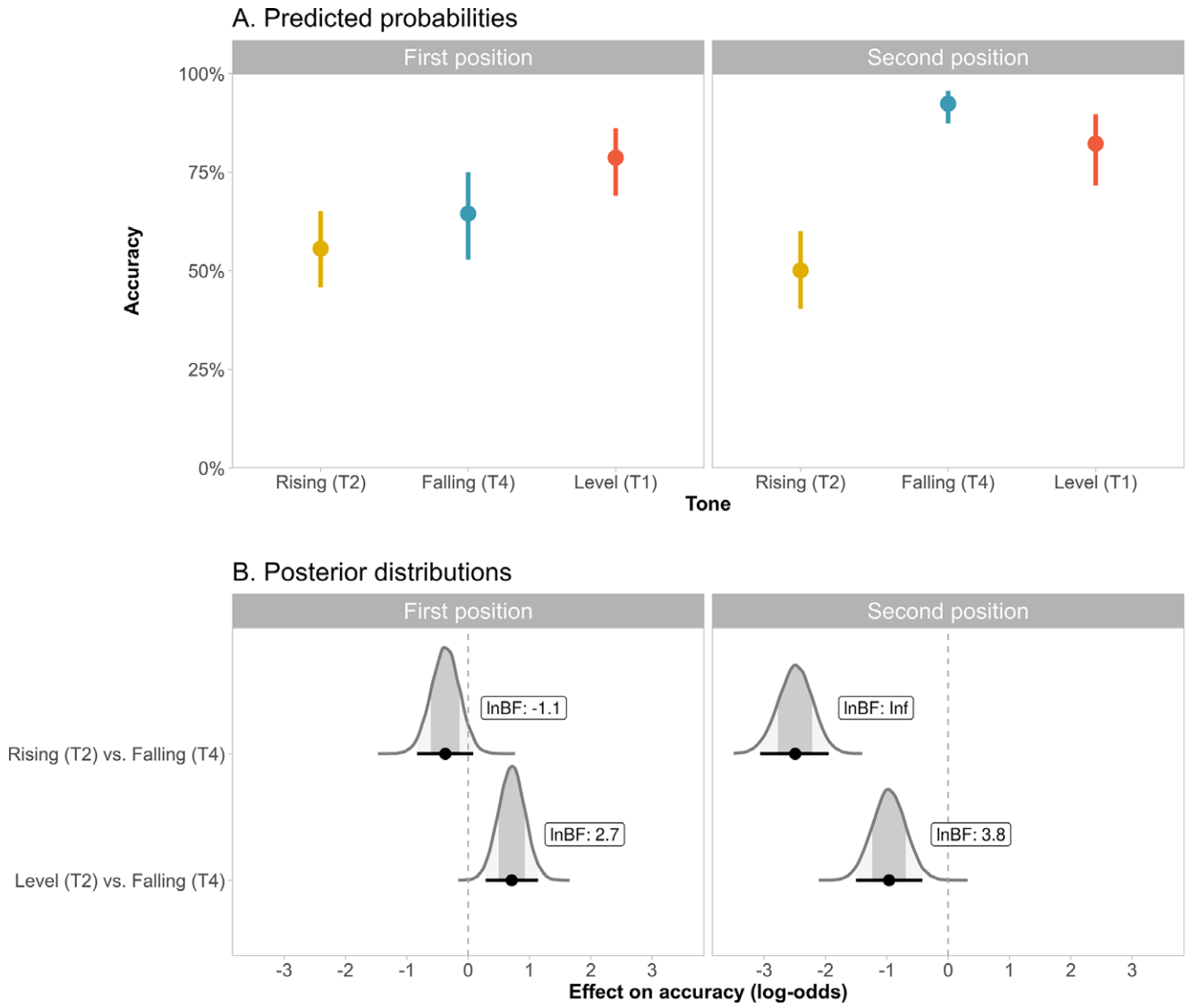

Figure 4. (A) Means (circles) and 95% credible intervals (vertical bars) of predicted proportions of correct responses for the different tone pairs (rising, falling, level), separately for the first and second syllabic positions. (B) Posterior distributions and natural logarithm of Bayes factors for the differences between tone pairs (rising vs. falling, level vs. falling) in each syllabic position (see Figure 1 caption for further details).

3.3. The role of L2 speech proficiency

In our final analyses, we assessed whether the OCP and TMS were modulated by L2 speech proficiency. A first statistical model, focussed on the OCP, included by-participant vocabulary size and pitch acuity scores as predictors (both centred and standardised), as well as their interactions with condition (identical vs. non-identical).

Figure 5A shows the posterior distributions and logged Bayes factors for the overall main effects of the two proficiency measures and their interactions with condition. Because the condition was coded with sum contrasts (i.e., −0.5/0.5), the main effects reflect the average effects of vocabulary size and pitch acuity across the two conditions, while the interactions reflect how much the difference between identical and non-identical pairs (the putative OCP effect) changes for every standard deviation in proficiency.

Vocabulary size did not have an effect on tone identification. Although its posterior mean (and the majority of the posterior distribution) was estimated to be positive (b = 0.29 [−0.08, 0.66]), the Bayes factor analyses provided support for the null hypothesis. Moreover, vocabulary size played no role in modulating the OCP effect, as revealed by a near-zero interaction (b = −0.03 [−0.32, 0.26]; see Figure 5A).

In contrast to vocabulary size, we obtained a clear effect of pitch acuity on tone identification: participants with better pitch acuity were more accurate at identifying tones, across identical and non-identical pairs (b = 0.54 [0.17, 0.91]. The predicted effect of pitch acuity on identification accuracy is displayed in Figure 6. This effect was found to be relatively large, with predicted accuracies ranging from approximately 25% to 60% (for participants who were 2 SDs below and above the mean, respectively).

The second proficiency analysis concerned the TMS and again included participant vocabulary size and pitch acuity scores as predictors, as well as their interactions with the relevant tone pair contrasts (i.e., rising vs. falling and level vs. falling). The posterior distributions and logged Bayes factors for these effects are displayed in Figure 5B. Recall that, unlike the OCP analyses, an assessment of the TMS on the identification of tone pairs requires comparing

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

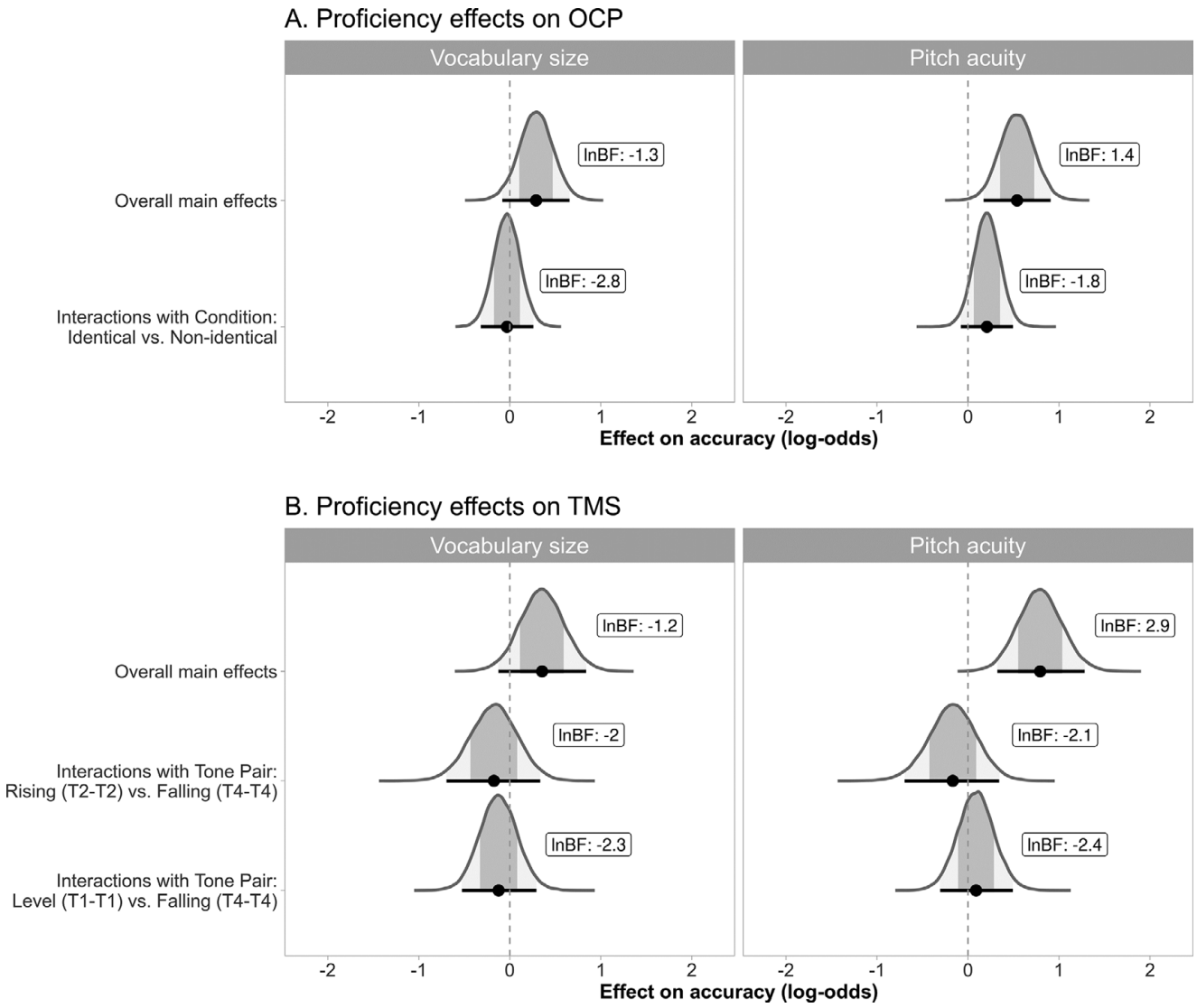

Figure 5. (A) Posterior distributions for the main effects of vocabulary size and pitch acuity and their interaction with the (A) OCP contrast, i.e., tone pairs in identical vs. non-identical conditions, and (B) the TMS contrasts, i.e., rising (T2-T2) vs. falling (T4-T4) and level (T1-T1) vs. falling (T4-T4) identical tone pairs.

tones within the identical condition. Thus, proficiency effects in this model were estimated for the identical tone pairs only.

As in the OCP proficiency analysis, there was no overall main effect of vocabulary size on accuracy (b = 0.35 [−0.13, 0.84]), and the differences between tone pairs were not modulated by vocabulary size (T2-T2 vs. T4-T4: b = −0.17 [−0.69, 0.33]; T1-T1 vs. T4-T4: b = −0.12 [−0.53, 0.29], with Bayes factors supporting the null hypotheses in all cases (see Figure 5B). As for pitch acuity, we again obtained a main effect of pitch acuity on identification accuracy (b = 0.79 [0.32, 1.28]), but no interactions with the tone pair contrasts (T2-T2 vs. T4-T4: b = −0.17 [−0.70, 0.34]; T1-T1 vs. T4-T4: b = 0.09 [−0.31, 0.49]).

4. General discussion

In the current tone identification experiment, we have tested the effects of two different phonological universals that have been proposed to constrain the L2 acquisition of Mandarin tones (Song, 2021; H. Zhang, 2007, 2016): (i) the OCP, which manifests as a dispreference for adjacent identical tones, and (ii) the TMS, according to which rising (T2) tones are dispreferred relative to falling (T4) tones and both of these are dispreferred relative to level (T1) tones. The OCP was tested by comparing accuracy rates in disyllabic pseudowords with identical and non-identical tone pairs. Bayesian mixed-effects analyses revealed that L2 learners were equally accurate in both conditions, suggesting that the OCP is not involved in tone identification. The TMS was tested by comparing accuracies in rising (T2-T2), falling (T4-T4) and level (T1-T1) tone pairs. Although participants were less accurate on rising tone pairs, there was little evidence for a difference between falling and level tones. An analysis of accuracy rates in each syllabic position (across identical and non-identical pairs) also showed that the TMS was not fully supported in either position. We have additionally explored whether the potential effects of universals are modulated by L2 phonological proficiency but found no evidence for such interactions. Instead, a clear effect of learners' pitch acuity was observed on tone identification accuracy.

4.1. OCP effects

The absence of OCP effects in L2 tonal perception is at odds with the findings of prior production studies (Song, 2021; H. Zhang,

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

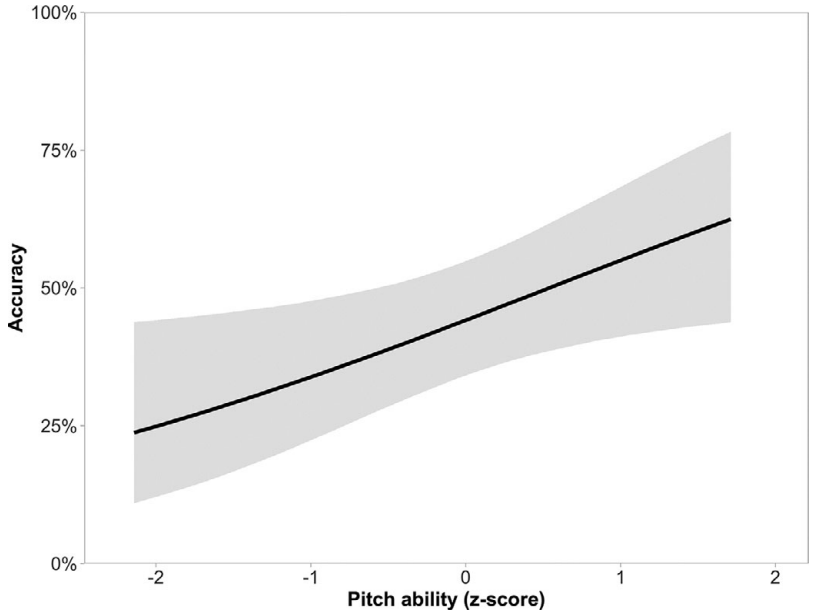

Figure 6. Mean and 95% credible interval of the predicted effect of (centred and standardised) pitch acuity on the proportion of correctly identified tone pairs (averaged across identical and non-identical tone pairs).

2007, 2016). To a certain extent, the null result of OCP in L2 perception poses a challenge to the generalizability of previous findings, especially considering the methodological improvements implemented in the current experiment (larger statistical power and better-matched stimuli between conditions). Given that the previously attested tonal dissimilation cannot be attributed to cross-linguistic influence, H. Zhang reasoned that L2 learners must have access to Universal Grammar, which supplies an innate universal constraint set, including the OCP. Nevertheless, the innateness of the OCP remains contentious. For example, a series of learning experiments conducted by Boll-Avetisyan and colleagues suggests instead that the OCP operates as a language-specific phonotactic constraint and is acquired on the basis of input distribution (Boll-Avetisyan, 2012; Boll-Avetisyan & Kager, 2014).

If the OCP were not responsible, what would explain the prior production results? As mentioned in the Introduction, one possibility is that the apparent OCP effects (e.g., /T4-T4/ produced as [T1-T4]) are actually due to miscoded phono-lexical forms (e.g., a target disyllabic word carrying two falling tones wrongly represented as /T1-T4/ in the L2 lexicon), which are retrieved as input in L2 tonal word production. Supporting evidence can be found in a series of studies conducted by Pelzl and colleagues (Pelzl et al., 2019, 2021a; Pelzl, Lau, Guo, & DeKeyser, 2021b), who found that, even for advanced learners with excellent tone identification abilities and a good command of vocabulary, the tonal representations in their mental lexicon may still be fuzzy (i.e., missing, incorrect, or uncertain). For a detailed discussion of factors that may contribute to such fuzziness, interested readers are referred to Pelzl et al. (2021b). Despite the plausibility of this alternative account, we believe that future work combining comparable perceptual and production experiments (Nagle & Baese-Berk, 2022) with the same group of L2 learners is needed before the OCP can be fully rejected.

Another way to explain the asymmetry between perceptual and production evidence is to acknowledge modality-specific OCP effects, which are attainable in several theoretical frameworks. The first one is to assume that the OCP is articulatory in nature (rearticulating the same gesture successively entails more effort than realising two different gestures, Dell, 1986), thus affecting L2 production exclusively. However, this straightforward account is not uncontroversial, because many studies that conceptualise the OCP as an articulatory constraint (e.g., Benus, Smorodinsky, & Gafos, 2004; Gafos, 2002, 2006) adhere to the theoretical view that gestures are perceptual primitives. That is to say, in order to model the production-specific effects, the articulation-based approach to OCP would need to depart from its fundamental premise. The second possibility is that L2 learners have developed distinct perception and production grammars (Ramus et al., 2010). This view is consistent with a growing number of studies showing that L2 speech perception and production do not always develop in tandem (see Nagle & Baese-Berk, 2022, for a review). However, it remains unclear why L2 Mandarin learners only integrate tonal similarity avoidance into their production grammar. Furthermore, even without assuming distinct grammars, modality-specific patterns may emerge, which leads us to a third possibility. In generative phonology, the mismatch between speech perception and production has been explicitly formalised in a number of studies. For instance, in Smolensky's (1996) optimality-theoretic model, a single grammar was proposed to explain why young English children may pronounce 'cat' as [kæ] themselves but would at the same time object to [kæ] when uttered by an adult. In speech perception (mapping from phonological surface forms to underlying forms), structural constraints such as the OCP cannot exert their effects, because they target surface forms, which are the non-evaluable input to perception (richness of base, Prince & Smolensky, 1993). In contrast, during production (mapping from underlying forms to surface forms), structural constraints have an effect, because the surface forms they evaluate are now the output. More recently, discrepancies between speech perception and production have also been modelled (Boersma & Hamann, 2009b; Cavirani & Hamann, 2022; Zhou & Hamann, 2024), under the Bidirectional Phonology and Phonetics Model (Boersma, 2011; Boersma & Hamann, 2009a). An explicit formal account is beyond the scope of this paper. Future

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

modelling studies, as well as experiments examining both L2 tonal perception and production, will be necessary to proceed with a complete formalisation along these lines.[2]

4.2. TMS effects

Our results did not fully support the predictions of the TMS, neither for tone pairs nor for individual tones. Accuracy for falling tone pairs (T4-T4) was comparable to that of level tones (T1-T1), against the TMS. Moreover, we have found that the results on tone pairs stemmed from particular patterns in each syllabic position. In the word-initial (first syllabic) position, learners were indeed more accurate at identifying level tones (T1), but falling tones (T4) and rising tones (T2) showed similar low accuracy rates. This pattern does not strictly conform to the TMS (*Rising >> *Falling >> *Level), but is better characterised as an instantiation of tone complexity (Contour >> Level). In the word-final (second syllabic) position, rising tones were again very difficult to identify, but the identification of falling tones was much more accurate and it even surpassed that of level tones—again, a pattern that does not align with the TMS.

The comparable accuracy rates for rising and falling tones in word-initial position contradict prior production studies, all of which have found TMS effects on individual tones. We speculate that this discrepancy may be attributable to the articulatory nature of the TMS. Unlike OCP, which is widely considered phonological in the literature, the TMS was formalised by Hyman and VanBik (2004) as a phonetically-grounded constraint, based on the articulatory evidence provided by Ohala (1978). A similar modality-specific pattern driven by articulatory constraints has been observed in the L2 acquisition of Mandarin tone sandhi. Qin (2022) observed that a sandhi process motivated by articulatory ease was more productive in L2 production than a phonetically arbitrary sandhi; however, this articulatory effect does not appear to exist in L2 perception (Luo, Williams, & Post, 2024). For formal modelling of how articulatory constraints lead to a mismatch between L2 perception and production, interested readers are referred to Zhou and Hamann (2024).

Learners' general difficulty with T2 across positions may be due to its confusability with T3 (Hao, 2012; Pelzl et al., 2019; e.g., So & Best, 2010, 2014), which is supported by the substitution patterns obtained in the current study (see Figure 3). The confusion between T2 and T3 likely arises from their perceptual-acoustic similarity, since both display an initial dip pitch followed by a rising contour (Moore & Jongman, 1997; Pelzl et al., 2019). Moreover, their confusability may also be enhanced by current pedagogical practices. Despite the fact that the most common allotone of T3 is low-level (Duanmu, 2007), its citation form, low-dipping, is often taught in the classroom and appears recurrently in textbooks (He, Wang, & Wayland, 2016; Pelzl et al., 2019; Shi, 2007; J. Zhang, 2014). Thus, the presence of a dipping contour in T2 may be misidentified as a cue to T3, or T3 might be regarded as T2 if its dip is not apparent enough.

Concerning the positional effect for falling tones, it may stem from cross-linguistic interaction. Although European Portuguese does not employ pitch for lexical contrast, it does use pitch movements to delineate the boundary of intonational phrases (Frota, 2000, 2014; Viana, 1987). Specifically, a declarative boundary-final tone manifests as a falling contour, which resembles the Mandarin falling tone to a certain degree. It is therefore plausible that the participants in this study have transferred this L1 prosodic feature to L2 tonal learning, leading to higher identification accuracy of T4 phrase-finally. Relying on L1 phrase-level pitch to acquire Mandarin lexical tones has long been reported (Broselow, Hurtig, & Ringen, 1987; Chunsheng Yang, 2016; C. Yang & Chan, 2010; H. Zhang, 2013). For instance, in a study with native speakers of English, which also displays a declarative-final falling contour, Broselow et al. (1987) have also found that T4 was more accurately identified in the final position than in the non-final position, a result they attributed to positive L1 prosodic transfer. Another potential explanation for the positional effect pertains to the presence of an additional cue for T4 identification, namely duration. Falling tones have the shortest duration in phrase-final position (Ho, 1976). Indeed, in our stimuli, T4 was the shortest tone in the second syllabic position (see Table 1), but longer than T1 in the first position. These hypotheses can be explored in future studies employing synthesised stimuli and controlling for the prosodic contexts in which they occur.

4.3. Are the effects of universals modulated by L2 proficiency?

The divergent findings observed in the tonal production of participants at different stages of L2 learning (Song, 2021; H. Zhang, 2007, 2016) led us to hypothesise that the effects of phonological universals might be modulated by L2 speech proficiency. In particular, we expected that L2 tonal phonology would be less influenced by the OCP and the TMS, with an increase in learners' Mandarin speech proficiency. Contrary to our prediction, the Bayesian analyses support the null effects for such an interaction. Instead, a relatively large effect of pitch acuity on tone identification accuracy was obtained. This finding has two important implications for our current understanding of the relationship between domain-general auditory processing and L2 speech development.

On the one hand, the observed effect sheds light on how different constructs of pitch acuity are related to L2 tonal perception. A listener's pitch aptitude can be inferred from their ability to perceive relative pitch differences in either linguistic (over speech units, including isolated vowels and monosyllabic Mandarin minimal pairs) or non-linguistic/domain-general (over sine waves) contexts. While many studies have found a positive correlation between linguistic pitch acuity and L2 tonal perception (Chandrasekaran, Sampath, & Wong, 2010; Perrachione, Lee, Ha, & Wong, 2011; e.g., Wong & Perrachione, 2007), domain-general pitch acuity was shown to only underlie the generalisation of tone learning to stimuli produced by novel speakers (Bowles, Chang, & Karuzis, 2016). Extending these previous findings, we demonstrated that domain-general pitch acuity plays an important role in L2 tonal perception, consistent with the Auditory Precision Hypothesis-L2 (Mueller, Friederici, & Männel, 2012; Saito, 2023). In particular, L2 learners with better auditory acuity exhibited heightened sensitivity to F0 differences in the Mandarin input, thus acquiring the mapping of specific F0 height/movement to the corresponding Mandarin tone category more effectively.

On the other hand, the current findings on pitch acuity help generalise the predictive power of domain-general auditory processing to different L2 learning settings. In prior research, the positive relationship between auditory acuity and phonological proficiency was only robustly observed in naturalistic L2 learning (Saito, Sun, et al., 2022; Zheng, Saito, & Tierney, 2022), but not in

[2]An anonymous reviewer pointed out yet another account for the lack of OCP effects, which is that the OCP might apply to 'constituent tones', rather than holistic units (e.g., T4 might be represented as H and L). An analysis in which we recoded our pairs to assess this hypothesis provided evidence against it (see Appendix S2).

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press

classroom settings (Saito, Suzukida, Tran, & Tierney, 2021). Saito et al. (2021) posited that L2 classroom learners cannot entirely benefit from auditory acuity because they typically receive and process a relatively limited amount of aural input in L2 classroom contexts. The results of the present study, however, suggest otherwise. Though our participants mainly received Mandarin input through formal instruction, having 'good ears' (domain-general pitch acuity) does lead to better performance with the perception of Mandarin tones. Note that the comparison between studies should be made with caution due to methodological differences. The outcome variable phonological accuracy in Saito et al. (2021) was a broad measure, namely the average native-like ratings of three different constructs (segments, word stress and intonation), and the predictor (auditory acuity) was likewise assessed by collapsing across various acoustic dimensions, such as duration, pitch and formant. Given the multifaceted nature of these two variables, their correlation might be rather loose. In the current study, the relationship between the predictor and the outcome variable was arguably much tighter, as pitch acuity is more strongly related to the identification of lexical tones, which are reliably distinguished by pitch differences. Further support for the robust effects of auditory processing at the dimension-specific level can be found in Saito, Kachlicka, et al. (2022), where F3 sensitivity was shown to predict the mastery of the English lateral-rhotic contrast (which is cued mainly by F3 differences) in both classroom and immersion learning settings.

The other predictor, Mandarin vocabulary knowledge—which has been considered a good indicator of L2 Mandarin speech proficiency—did not relate to the accuracy of L2 tonal perception. We speculate that this null result might be ascribed to the specific characteristics of the proficiency measure used in this study. First, the LEXTALE_CH vocabulary test (Chan & Chang, 2018) assesses knowledge of written characters (tonal knowledge is not explicitly required), and thus may be less predictive of tone perception. Second, all of the LEXTALE_CH test items are single characters (monosyllabic), which might compromise the validity of the test, given that most Chinese words are disyllabic (Duanmu, 2007). Future research interested in the lexical effects in L2 tonal perception is encouraged to consider a more recent version of the Chinese vocabulary test with two-character items, LexCHI (Wen et al., 2023), which revealed higher correlations with a cloze test and a translation task than the single-character version.

5. Conclusion

This study set out to investigate whether the L2 perception of Mandarin tone sequences is governed by phonological universals —namely, the Obligatory Contour Principle (OCP) and the Tonal Markedness Scale (TMS)—and whether the potential effects of these universals are modulated by individual difference predictors, such as vocabulary size and pitch acuity.

Contrary to predictions derived from previous production studies, Bayesian mixed-effects analyses revealed null effects for both phonological universals in L2 perception. These findings point to a possible perception-production asymmetry in L2 tonal acquisition, calling for future studies to entertain the precise nature of these phonological universals and explicitly model how the two speech modalities interact in L2 phonological acquisition.

Notably, while vocabulary size—a common proxy for L2 speech proficiency—was unrelated to tone identification performance, learners' pitch acuity turned out to be a robust predictor of learning success. This result not only provides support for the Auditory Precision Hypothesis, underscoring the role of domain-general auditory processing in L2 phonological acquisition, but also extends its applicability to classroom-based learning contexts.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728925100114.

Data availability statement. All materials, data and code are openly available in the OSF repository at https://osf.io/ezadw.

Acknowledgements. We thank Ling Li for help in recording the stimuli, and the audiences of the Architectures and Mechanisms for Language Processing (AMLaP) 2023, Manchester Phonology Meeting (mfm) 2023, and Encontro Nacional da Associação Portuguesa de Linguística (ENAPL) 2022 conferences for helpful discussions. We are also grateful to the three anonymous reviewers for their insightful feedback on earlier versions of this paper.

Funding statement. This work has been funded by the Fundação para a Ciência e a Tecnologia (FCT, Foundation for Science and Technology), grant UID/00214: Center of Linguistics of the University of Lisbon.

Competing interests. The authors declare none.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, **52**(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Benus, S., Smorodinsky, I., & Gafos, A. (2004). Gestural coordination and the distribution of English "geminates. In S. Arunachalam & T. Scheffler (Eds.), *Proceedings of the twenty-seventh penn linguistics colloquium* (pp. 33–46). University of Pennsylvania.

Berent, I., Everett, D. L., & Shimron, J. (2001). Do phonological representations specify variables? Evidence from the obligatory contour principle. *Cognitive Psychology*, **42**(1), 1–60. https://doi.org/10.1006/cogp.2000.0742

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, **64**(1), 39–72. https://doi.org/10.1016/S0010-0277(97)00016-4

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives* (PhD thesis). University of Amsterdam.

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz & J. Mattausch (Eds.), *Linguistik Aktuell/linguistics today* (Vol. **180**, pp. 33–72). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/la.180.02boe

Boersma, P., & Hamann, S. (2009a). Introduction: Models of phonology in perception. In P. Boersma & S. Hamann (Eds.), *Phonology in perception* (pp. 1–24). DE GRUYTER MOUTON. https://doi.org/10.1515/9783110219234.1

Boersma, P., & Hamann, S. (2009b). Loanword adaptation as first-language phonological perception. In A. Calabrese & W. L. Wetzels (Eds.), *Current issues in linguistic theory* (Vol. **307**, pp. 11–58). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.307.02boe

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer [computer program]*. Retrieved from http://www.praat.org/

Boll-Avetisyan, N. (2012). Phonotactics and its acquisition, representation, and use: An experimental-phonological study (Doctoral dissertation). Universiteit Utrecht.

Boll-Avetisyan, N., & Kager, R. (2014). OCP- PLACE in speech segmentation. *Language and Speech*, **57**(3), 394–421. https://doi.org/10.1177/0023830913508074

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, **66**(4), 774–808. https://doi.org/10.1111/lang.12159

Broselow, E., Hurtig, R., & Ringen, C. (1987). The perception of second language prosody. In G. Ioup & S. Weinberger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system*. Cambridge, MA: Newbury House.

Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., & Tyler, M. D. (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied PsychoLinguistics*, **33**(3), 643–664. https://doi.org/10.1017/S0142716411000518

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, **33**(3), 433–461. https://doi.org/10.1017/S0272263111000040

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, **80**(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, **5**(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Cavirani, E., & Hamann, S. (2022). Formalising phonological perception: The role of voicing assimilation in consonant cluster perception in Emilian dialects. *Journal of Linguistics*, 1–32. https://doi.org/10.1017/S0022226722000457

Chan, I. L., & Chang, C. B. (2018). LEXTALE_CH: A quick, character-based proficiency test for mandarin Chinese. In A. B. Bertolini & M. J. Kaplan (Eds.), *Proceedings of the 42nd annual Boston University conference on language development* (Vol. **1**, pp. 114–130). Cascadilla Press.

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

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America*, **138**(6), 3703–3716. https://doi.org/10.1121/1.4937612

Coetzee, A. (2005). The OCP in the perception of English. In S. Frota, M. Vigário, & M. J. Freitas (Eds.), *Prosodies* (pp. 223–245). De Gruyter Mouton.

Coetzee, A. (2008). Grammar is both categorical and gradient. In S. Parker (Ed.), *Phonological argumentation*. Equinox.

Daidone, D., & Darcy, I. (2021). Vocabulary size is a key factor in predicting second language lexical encoding accuracy. *Frontiers in Psychology*, **12**, 688356. https://doi.org/10.3389/fpsyg.2021.688356

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, **41**(1), 214–226. https://doi.org/10.1214/aoms/1177697203

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, **25**(1), 207–218. https://doi.org/10.3758/s13423-017-1266-z.

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). Oxford University Press.

Elliot, C. E. (1991). The relationship between the perception and production of Mandarin tones: An exploratory study. *University of Hawai'i Working Papers in ESL*, **10**(2), 177–204. Retrieved from http://hdl.handle.net/10125/38573

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-place in Arabic. *Language*, **77**(1), 91–106. https://doi.org/10.1353/lan.2001.0014

Frota, S. (2000). *Prosody and focus in European Portuguese: Phonological phrasing and intonation*. Routledge. https://doi.org/10.4324/9781315054384.

Frota, S. (2014). The intonational phonology of European Portuguese. In S.-A. Jun (Ed.), *Prosodic typology II: The phonology of intonation and phrasing* (1st ed., pp. 6–42). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567300.003.0002.

Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language and Linguistic Theory*, **20**(2), 269–337. https://doi.org/10.1023/a:1014942312445.

Gafos, A. I. (2006). Dynamics in grammar. In M. L. Goldstein, D. H. Whalen, & C. Best (Eds.), *Laboratory phonology 8: Varieties of phonological competence* (pp. 51–79). De Gruyter Mouton.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**(4). https://doi.org/10.1214/08-AOAS191

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, **13**(2), 359–383. https://doi.org/10.1214/17-BA1051

Goldsmith, J. (1976). *Autosegmental phonology* [PhD thesis]. MIT.

Gong, S. (2022). *The Obligatory Contour Principle effects in phonological learning* [PhD thesis]. University of Kansas.

Graff, P. (2012). *Communicative efficiency in the lexicon* [PhD thesis]. MIT.

Greenberg, J. (1950). The patterning of root morphemes in Semitic. *Word*, **5**, 162–181.

Gussenhoven, C., & Chen, A. (Eds.) (2020). In *The Oxford handbook of language prosody* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.001.0001

Hao, Y.-C. (2012). Second language acquisition of mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, **40**(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

He, Y., Wang, Q., & Wayland, R. (2016). Effects of different teaching methods on the production of mandarin tone 3 by English speaking learners. *Chinese as a Second Language*, **51**(3), 252–265. https://doi.org/10.1075/csl.51.3.02he

Ho, A. T. (1976). The acoustic variation of mandarin tones. *Phonetica*, **33**(5), 353–367. https://doi.org/10.1159/000259792

Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language*, **27**(1), 3–42. https://doi.org/10.1017/s030500099900402x

Huang, T., & Johnson, K. (2011). Language specificity in speech perception: Perception of mandarin tones by native and nonnative listeners. *Phonetica*, **67**(4), 243–267. https://doi.org/10.1159/000327392

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, **15**(2), 422–433. https://doi.org/10.1017/S1366728911000678

Hyman, L. M., & VanBik, K. (2004). Directional rule application and output problems in Hakha Lai tone. *Language and Linguistics*, **5**, 821–861.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, **59**(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, **192**, 15–24. https://doi.org/10.1016/j.bandl.2019.02.004

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Leben, W. (1973). *Suprasegmental phonology* [PhD thesis]. MIT.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented mandarin tones by native and non-native listeners. *Journal of Phonetics*, **37**(1), 1–15. https://doi.org/10.1016/j.wocn.2008.08.001

Llompart, M. (2021). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and Cognition*, **24**(3), 481–496. https://doi.org/10.1017/S1366728920000656.

Luo, X., Williams, J., & Post, B. (2024, November 12). Incidental learning of phonetically (un)motivated tone sandhi patterns by tonal and non-tonal L1 speakers. https://doi.org/10.33774/coe-2024-xhswk

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology* (0th ed.). Routledge. https://doi.org/10.4324/9781410604293

Mayer, T., Rohrdantz, C., & Plank, F. (2010). Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. *Proceedings of the ACL 2010 workshop on NLP and linguistics: Finding the common ground*, 67–75. Uppsala, Sweden.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of mandarin Chinese tones. *The Journal of the Acoustical Society of America*, **102**(3), 1864–1877. https://doi.org/10.1121/1.420092

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences*, **109**(39), 15953–15958. https://doi.org/10.1073/pnas.1204319109

Nagano-Madsen, Y., & **Wan, X.** (2017). Perception and production of L2 mandarin tones by Swedish learners. *2017 25th European signal processing conference (EUSIPCO)*, 578–582. Kos, Greece: IEEE. https://doi.org/10.23919/EUSIPCO.2017.8081273

Nagle, C. L., & **Baese-Berk, M. M.** (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, **44**(2), 580–605. https://doi.org/10.1017/S0272263121000371

Neergaard, K. D., **Xu, H.**, **German, J. S.**, & **Huang, C.-R.** (2022). Database of word-level statistics for mandarin Chinese (DoWLS-MAN). *Behavior Research Methods*, **54**(2), 987–1009. https://doi.org/10.3758/s13428-021-01620-7

Oberauer, K. (2022). The importance of random slopes in mixed models for Bayesian hypothesis testing. *Psychological Science*, **33**(4), 648–665. https://doi.org/10.1177/09567976211046884

Odden, D. (1986). On the role of the obligatory contour principle in phonological theory. *Language*, **62**(2), 353. https://doi.org/10.2307/414677

Odden, D. (1988). Anti antigemination and the OCP. *Linguistic Inquiry*, **19**(3), 451–475. Retrieved from http://www.jstor.org/stable/25164904

Ohala, J. J. (1973). The physiology of tone. *Southern California Occasional Papers in Linguistics*, **1**, 1–14.

Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 3–39). Academic Press.

Pelzl, E. (2019). What makes second language perception of mandarin tones hard?: A non-technical review of evidence from psycholinguistic research. *Chinese as a Second Language*, **54**(1), 51–78. https://doi.org/10.1075/csl.18009.pel

Pelzl, E., **Lau, E. F.**, **Guo, T.**, & **DeKeyser, R.** (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, **41**(1), 59–86. https://doi.org/10.1017/S0272263117000444

Pelzl, E., **Lau, E. F.**, **Guo, T.**, & **DeKeyser, R.** (2021a). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in mandarin: Findings from behavioral and event-related potentials experiments. *Studies in Second Language Acquisition*, **43**(2), 268–296. https://doi.org/10.1017/S027226312000039X

Pelzl, E., **Lau, E. F.**, **Guo, T.**, & **DeKeyser, R. M.** (2021b). Advanced second language learners of mandarin show persistent deficits for lexical tone encoding in picture-to-word form matching. *Frontiers in Communication*, **6**, 689423. https://doi.org/10.3389/fcomm.2021.689423

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

Pozdniakov, K., & **Segerer, G.** (2007). Similar place avoidance: A statistical universal. *Linguistic Typology*, **11**(2). https://doi.org/10.1515/LINGTY.2007.025.

Prince, A., & **Smolensky, P.** (1993). *Optimality theory: Constraint interaction in generative grammar* (Technical Report No. RuCCS-TR-2). New Brunswick, NJ: Rutgers University Center for Cognitive Science.

Prince, A., & **Smolensky, P.** (2004). *Optimality theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell Pub.

Qin, Z. (2022). The second-language productivity of two mandarin tone sandhi patterns. *Speech Communication*, **138**, 98–109. https://doi.org/10.1016/j.specom.2022.02.009

Quené, H., & **van den Bergh, H.** (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, **59**(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ramus, P., **Christophe, A.**, **Jacquemot, C.**, **Kouider, S.**, & **Dupoux, E.** (2010). A psycholinguistic perspective on the acquisition of phonology. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 311–342). De Gruyter Mouton. https://doi.org/10.1515/9783110224917.3.311

Rouder, J. N., **Speckman, P. L.**, **Sun, D.**, **Morey, R. D.**, & **Iverson, G.** (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, **16**(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Saito, K. (2023). How does having a good ear promote successful second language speech acquisition in adulthood? Introducing auditory precision hypothesis-L2. *Language Teaching*, **56**(4), 522–538. https://doi.org/10.1017/S0261444822000453

Saito, K., **Kachlicka, M.**, **Suzukida, Y.**, **Petrova, K.**, **Lee, B. J.**, & **Tierney, A.** (2022). Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning. *Cognition*, **229**, 105236. https://doi.org/10.1016/j.cognition.2022.105236

Saito, K., **Sun, H.**, **Kachlicka, M.**, **Alayo, J. R. C.**, **Nakata, T.**, & **Tierney, A.** (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, **44**(1), 57–86. https://doi.org/10.1017/S0272263120000467

Saito, K., **Suzukida, Y.**, **Tran, M.**, & **Tierney, A.** (2021). Domain-general auditory processing partially explains second language speech learning in classroom settings: A review and generalization study. *Language Learning*, **71**(3), 669–715. https://doi.org/10.1111/lang.12447

Sandlund, E., **Sundqvist, P.**, & **Nyroos, L.** (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass*, **10**(1), 14–29. https://doi.org/10.1111/lnc3.12174

Schad, D. J., **Betancourt, M.**, & **Vasishth, S.** (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, **26**(1), 103–126. https://doi.org/10.1037/met0000275

Schad, D. J., **Nicenboim, B.**, **Bürkner, P.-C.**, **Betancourt, M.**, & **Vasishth, S.** (2022). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*. https://doi.org/10.1037/met0000472

Schad, D. J., **Vasishth, S.**, **Hohenstein, S.**, & **Kliegl, R.** (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, **110**, 104038. https://doi.org/10.1016/j.jml.2019.104038

Schmalz, X., **Biurrun Manresa, J.**, & **Zhang, L.** (2023). What is a Bayes factor? *Psychological Methods*, **28**(3), 705–718. https://doi.org/10.1037/met0000421

Shatzman, K. B., & **Kager, R.** (2007). A role for phonotactic constraints in speech perception. *Proceedings of ICPhS XVI*, 1409–1412.

Shi, J. (2007). On teaching tone three in mandarin. *Journal of the Chinese Language Teachers Association*, **42**(2), 1–10.

Silpachai, A. (2020). The role of talker variability in the perceptual learning of mandarin tones by American English listeners. *Journal of Second Language Pronunciation*, **6**(2), 209–235. https://doi.org/10.1075/jslp.19010.sil

Sinharay, S., & **Stern, H. S.** (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, **56**(3), 196–201. https://doi.org/10.1198/000313002137

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, **27**(4), 720–731. Retrieved from http://www.jstor.org/stable/4178959

So, C. K., & **Best, C. T.** (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, **53**(2), 273–293. https://doi.org/10.1177/0023830909357156

So, C. K., & **Best, C. T.** (2014). Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, **36**(2), 195–221. https://doi.org/10.1017/S0272263114000047

Song, C. (2021). What is in the final stage of inter-language? Tone errors and phonological constraints in spontaneous speech in very advanced learners of mandarin. In C. Yang (Ed.), *The acquisition of Chinese as a second language pronunciation* (pp. 21–54). Springer Singapore. https://doi.org/10.1007/978-981-15-3809-4_2

Spinelli, G., **Forti, L.**, & **Jared, D.** (2021). Learning to assign stress in a second language: The role of second-language vocabulary size and transfer from the native language in second-language readers of Italian. *Bilingualism: Language and Cognition*, **24**(1), 124–136. https://doi.org/10.1017/S1366728920000243

Vasishth, S., **Nicenboim, B.**, **Beckman, M. E.**, **Li, F.**, & **Kong, E. J.** (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, **71**, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Veríssimo, J. (2025). A gentle introduction to Bayesian statistics, with applications to bilingualism research. *Linguistic Approaches to Bilingualism*. https://doi.org/10.1075/lab.24027.ver

Veríssimo, J., & **Clahsen, H.** (2014). Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese. *Journal of Memory and Language*, **76**, 61–79. https://doi.org/10.1016/j.jml.2014.06.001

Viana, M. C. (1987). *Para a síntese da entoação do português [For a synthesis of Portuguese intonation]* [Dissertation for access to Assistant Researcher]. CLUL-INIC.

Wagenmakers, E.-J., **Lodewyckx, T.**, **Kuriyal, H.**, & **Grasman, R.** (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–Dickey method. *Cognitive Psychology*, **60**(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

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

Wen, Y., **Qiu, Y.**, **Leong, C. X. R.**, & **Van Heuven, W. J. B.** (2023). LexCHI: A quick lexical test for estimating language proficiency in Chinese. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02151-z

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

Xu, Y. (2002). Articulatory constraints and tonal alignment. *Proceedings of the 1st international conference on speech prosody*, 91–100. Aix-en-Provence, France.

Xu, Y., & **Sun, X.** (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America*, **111**(3), 1399–1413. https://doi.org/10.1121/1.1445789

Yang, C. (2016). *The acquisition of L2 mandarin prosody: From experimental studies to pedagogical practice*. Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.1

Yang, C., & **Chan, M. K. M.** (2010). The perception of mandarin Chinese tones and intonation. *Journal of the Chinese Language Teachers Association*, **45**(1), 7–36.

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

Zhang, H. (2007). *A phonological study of second language acquisition of Mandarin Chinese tones* [Master's thesis]. University of North Carolina at Chapel Hill.

Zhang, H. (2013). *The second language acquisition of Mandarin Chinese tones by English, Japanese and Korean speakers* [PhD thesis]. University of North Carolina at Chapel Hill.

Zhang, H. (2016). Dissimilation in the second language acquisition of mandarin Chinese tones. *Second Language Research*, **32**(3), 427–451. https://doi.org/10.1177/0267658316644293.

Zhang, J. (2002). *The effects of duration and sonority on contour tone distribution: A typological survey and formal analysis*. Routledge.

Zhang, J. (2004). The role of contrast-specific and language-specific phonetics in contour tone distribution. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 157–190). Cambridge University Press.

Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. In C.-T. J. Huang, Y.-H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 443–464). Wiley-Blackwell.

Zheng, C., **Saito, K.**, & **Tierney, A.** (2022). Successful second language pronunciation learning is linked to domain-general auditory processing rather than music aptitude. *Second Language Research*, **38**(3), 477–497. https://doi.org/10.1177/0267658320978493

Zhou, C., & **Hamann, S.** (2024). Modelling the acquisition of the Portuguese tap by L1-mandarin learners: A BiPhon-HG account for individual differences, syllable-position effects and orthographic influences in L2 speech. *Glossa*, **9**(1). https://doi.org/10.16995/glossa.9692

Zhou, C., & **Li, X.** (2021). LextPT: A reliable and efficient vocabulary size test for L2 Portuguese proficiency. *Behavior Research Methods*, **54**(6), 2625–2639. https://doi.org/10.3758/s13428-021-01731-1

Zou, T., **Chen, Y.**, & **Caspers, J.** (2017). The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of mandarin tones. *Bilingualism: Language and Cognition*, **20**(5), 1017–1029. https://doi.org/10.1017/S1366728916000791

https://doi.org/10.1017/S1366728925100114 Published online by Cambridge University Press