2025. Proc Ling Soc Amer 10(1). 5886. https://doi.org/10.3765/plsa.v10i1.5886.

# Tone 3 Sandhi in Mandarin: Productivity and acoustic realization in L1 and L2 speakers

Xiao Dong, Chien-Jer Charles Lin*

**Abstract**. This study investigates the productivity and acoustic realization of a highly regular tonal alternation in Mandarin, Tone 3 (T3) sandhi, among native speakers and intermediate-level second language (L2) learners across different lexical contexts. Seventeen native speakers (NSs), fourteen intermediate-low (IL) learners, and sixteen intermediate-high (IH) learners completed a production task involving both real and nonce disyllabic words. Results show that T3 sandhi is highly productive among native speakers and is consistently applied by L2 learners, with IL learners achieving application rates above 77% even in pseudo-word contexts, providing strong support for a computation-based mechanism in sandhi production. Acoustic analysis furthered revealed that IH learners produced more native-like sandhi realizations than IL learners, although differences from NSs remained under pseudo-word conditions. In addition, NSs and IH learners showed similar differences in T3 and Tone 2 (T2) production across lexical conditions, highlighting a lexical effect on tonal production. These findings contribute to our understanding of L2 phonological acquisition, demonstrating that regular phonological alternations such as T3 sandhi can be productively acquired at relatively early stages of L2 acquisition.

**Key words**. Mandarin Tone 3 sandhi; phonological alternations; L2 acquisition; speech production; tonal processing

**1. Introduction**. Phonological alternation, in which the pronunciation of a form changes depending on its phonological context, is prevalent across many languages. These alternations often result in a mismatch between a word's underlying form and its surface realization. Traditional linguistic theories have treated regular alternations as the result of systematic input-output mappings, where an underlying representation is transformed into a surface form based on contextual rules. This mapping is viewed as part of a speaker's competence grammar (Cole & Hualde, 2011; Kenstowicz & Kisseberth, 1979). However, recent research has questioned whether such alternations are computed online during speech production or whether they are stored as surface forms in the mental lexicon (Zhang et al., 2022). This debate has motivated growing interest in the psychological reality of phonological alternations in real-time language use.

Studies on L1 processing suggest that the regularity and productivity of an alternation influence how it is mentally represented and executed. Highly regular and productive alternations are more likely to be computed online, while irregular or low-productivity patterns are often stored as fixed forms (Arndt-Lappe & Ernestus, 2020; Zhang et al., 2022). However, even for highly productive rules, native speaker mastery may not extend uniformly to all levels of speech production. For example, Zhang et al. (2022) found that Mandarin speakers apply the T3 sandhi rule nearly categorically to novel words, but with phonetic realizations that differ from those in familiar words—suggesting incomplete phonetic generalization despite successful rule application.

The acquisition and online implementation of phonological alternations are even more complex for second language (L2) learners. Learners must not only recognize the existence of an

*Authors: Xiao Dong, Indiana University Bloomington (dong1@.edu) & Chien-Jer Charles Lin, Indiana University Bloomington (chiclin@iu.edu).

alternation but also acquire its abstract form and apply it correctly under appropriate linguistic conditions. Despite its theoretical significance, relatively few studies have examined how L2 learners acquire and produce phonological alternations, especially in contexts that require online computation rather than memorization. It remains an open question whether L2 learners form abstract, productive representations of phonological alternations or rely more heavily on lexical memory—particularly when producing novel or nonce forms.

Another critical factor is proficiency (Jin, 2019). Higher proficiency in the L2 may facilitate more native-like abstraction and application of phonological rules, while lower proficiency may result in greater reliance on memorized forms and less stable phonetic realization under increased cognitive load. Understanding how proficiency mediates the production of phonological alternations can offer important insights into the trajectory of L2 phonological development.

This study addresses these questions by investigating the productivity and acoustic realization of Mandarin T3 sandhi, a highly regular and productive tonal alternation. Using a production task involving both real and nonce disyllabic words, we examine how native speakers and intermediate-level L2 learners apply the sandhi rule and realize the resulting tones acoustically. Specifically, we ask: (1) whether native speakers and L2 learners apply the rule productively across lexical contexts; (2) how lexical familiarity and online processing demands influence rule application; (3) how learners' application rates and acoustic realizations differ from those of native speakers; and (4) how learner proficiency affects both the application and realization of the sandhi rule. By examining both rule productivity and phonetic detail, this study aims to contribute to a deeper understanding of phonological processing and acquisition, the phenomenon of incomplete rule implementation, and the cognitive mechanisms underlying phonological alternations in both native and non-native systems.

1.1. LEXICAL TONES IN MANDARIN CHINESE. Mandarin Chinese is a tonal language. It phonemically distinguishes four tones, with Tone 1 (T1) having a high-level pitch, Tone 2 high-rising pitch, Tone 3 low-dipping pitch, and Tone 4 (T4) high-falling pitch (Chao, 1948; see Figure 1). T1–T4 are often transcribed as 55, 35, 214, and 51, respectively, on a scale of 1 to 5, with 1 indicating the lowest pitch and 5 the highest pitch (Chao, 1948). The same segmental context

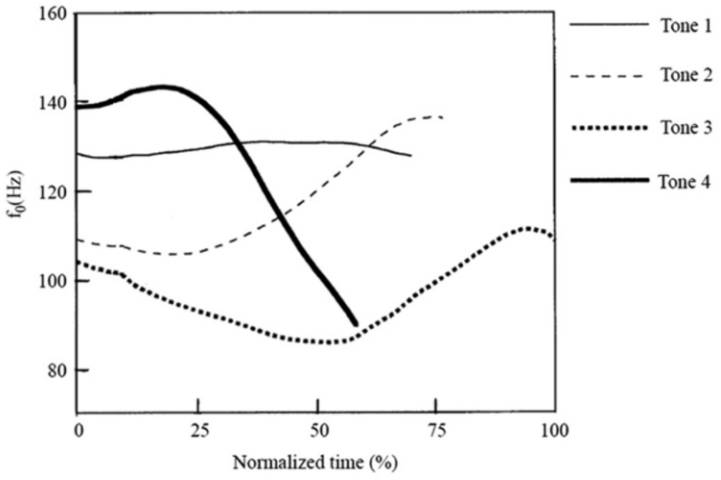


Figure 1: Mandarin Chinese lexical tone pitch contours from Xu (1997)

carries different meanings depending on the tone. For example, the meaning of Mandarin Chinese *ma* with T1 is "mother," the T2 version means "hemp," and the T3 and T4 meanings are "horse" and "scold," respectively.

1.2. MANDARIN TONE 3 SANDHI. Mandarin T3 sandhi is an obligatory phonological alternation rule in Mandarin Chinese. In T3 sandhi, an initial T3 (the low-dipping tone, 214) syllable changes into a T2 (the high-rising tone, 35) when it is followed by another T3 syllable. For example, in the disyllabic compound /mei3-xau3/ "wonderful," which combines /mei3/ "beautiful" and /xau3/ "good," the first syllable /mei3/ surfaces as [mei2] due to sandhi, resulting in the pronunciation [mei2-xau3]. This creates a mismatch between the word's underlying tones and its surface realization.

Both T2 and T3 are contour tones. Most production and perception studies highlight the F0 turning point and ΔF0 (i.e., the average F0 change) as the key features differentiating these tones (Chang, 2011). T2 typically has an earlier F0 turning point and a higher ΔF0, reflecting its more pronounced rising contour. Some research has also suggested that duration may serve as a distinguishing factor, with T3 generally being longer than T2 when produced in isolation (Shen, 1990). However, Shih (2007) challenges the reliability of duration as a consistent cue, noting that T3 is longer than T2 only in isolated settings. In conversational speech, T3 often becomes the shortest of the four Mandarin tones because its rising tail is usually truncated, undermining the use of duration as a distinguishing feature.

1.3. PRODUCTIVITY AND ACOUSTIC REALIZATION OF TONE 3 SANDHI AMONG L1 ADULTS. The wug test has been widely used to test the productivity of Mandarin T3 sandhi and other phonological rules. It is a classic experimental tool in linguistic research designed to assess whether speakers can generalize known rules to new, unfamiliar contexts. It involves presenting participants with made-up or nonce words (e.g., "wug") and asking them to apply a linguistic rule, such as pluralization or tone sandhi, to see if they can extend it to these novel forms. Successful application of the rule to nonce words suggests that speakers have internalized the abstract phonological rule, rather than relying on rote memorization of specific word pairs.

Using this approach, Zhang and Lai (2010) and Zhang and Peng (2013) found that native Mandarin speakers applied the T3 sandhi rule 100% of the time to nonce words, confirming its productivity. These results indicate that native speakers dynamically process the rule in real time, applying it flexibly rather than simply recalling memorized forms, providing support for the computation mechanism.

However, acoustic analysis in Zhang and Lai (2010) discovered differences in the acoustic realization of sandhi T3 between real words and wug words: the turning point of sandhi T3 was significantly lower and occurred later in wug words than in real words. Since this difference was not due to a failure to apply the sandhi rule, the authors suggested that the sandhi was incompletely applied in many cases involving wug words. They proposed that this might be because T3 sandhi lacks a strong phonetic motivation, making its application more variable. Although the rule was still applied to wug words, the process appeared to be incomplete, resulting in productions that more closely resembled the underlying T3.

1.4. PRODUCTIVITY OF MANDARIN TONE 3 SANDHI IN L2 LEARNERS. Several studies have examined how L2 learners acquire and apply Mandarin T3 sandhi. Early work by Yang (2016) and Zhang (2016) investigated sandhi production using reading tasks. Yang (2016) found that intermediate-level learners applied T3 sandhi at rates of 67.5% for familiar words and 51.25% for nonce words under normal speech. Zhang's (2016) study, which included learners at

beginning, intermediate, and advanced levels and focused exclusively on pseudo-words, found that sandhi application increased with proficiency, from 44% in beginners to 80% in advanced learners. These results suggest that L2 learners develop partial rule knowledge and can generalize the alternation beyond memorized chunks. However, small sample sizes and reliance on reading tasks limit the generalizability of these findings. In reading tasks, participants must perform two tasks simultaneously: retrieving the tonal contours of T3 and applying the sandhi rule. When participants fail to apply the sandhi rule correctly, it is unclear whether the difficulty arises from retrieving the correct tone or from the application of the rule itself.

More recent studies have shifted to repetition tasks to reduce lexical retrieval demands and better isolate online rule application (Chen et al., 2019; Qin, 2022). In this approach, participants first listen to the two syllables of the disyllabic target words separately and then repeat the full target word. Chen et al. (2019) studied 23 intermediate-level L2 learners (Cantonese and English speakers) and 12 native Mandarin speakers, comparing F0 contours across real and pseudo-words. They found no significant acoustic differences between conditions and interpreted this as evidence that L2 learners had developed an abstract representation of the T3 category and used a computation mechanism to dynamically process and apply the sandhi rule across lexical conditions. Recruiting 9 high-proficiency and 10 low-proficiency native Korean-speaking learners along with 12 native northern Mandarin speakers, Qin (2022) observed that both learners and native speakers exhibited a smaller rising slope (more T3-like) for wug words than for real words. Qin interpreted this discrepancy as evidence of low productivity among the L2 learner groups, indicating that their ability to apply the sandhi rule was limited. Furthermore, proficiency level did not significantly alter this pattern.

A broader issue is the lack of consistency in how productivity is defined and measured. While some studies focus on categorical application rates (e.g., Yang, 2016; Zhang, 2016), others emphasize acoustic features (e.g., Chen et al., 2019; Qin, 2022), leading to divergent interpretations. These inconsistencies underscore the need for more methodologically unified research that combines both application rates and acoustic analyses to provide a more comprehensive understanding of how L2 learners apply and realize T3 sandhi.

**2. Method**. Our study adopts a wug test paradigm. To make sure learner participants had developed the knowledge of Mandarin lexical tonal production and acquired all the real test words, we administered two pre-tests for learner participants: (1) a monosyllabic word repetition test and (2) a word test.

2.1. PARTICIPANTS. A total of 46 participants took part in the study, including 17 native Mandarin speakers, 13 intermediate-low learners, and 16 intermediate-high learners. One native speaker was excluded from the analysis due to an extended period of residence in southern China during childhood. All other native speakers were born and raised in northern China and were enrolled in degree programs at a university in the United States. The L2 learners were recruited from the same university's Chinese language program, with intermediate-low learners drawn from second-year classes and intermediate-high learners from third-year classes. All participants reported normal hearing and no history of language disorders. None of the learner participants were heritage speakers of Mandarin.

At the end of the experiment, all participants completed a background questionnaire that collected demographic information and language experience, including self-rated Mandarin proficiency. In addition, they completed a cloze test adapted from Qin (2022) to provide an independent measure of language proficiency. As shown in Table 1, all native speakers acquired Mandarin as their first language and reported high self-rated proficiency (mean = 6.6/7), with an

average cloze test score of 39.7/40. Intermediate-low learners began learning Mandarin at an average age of 16.5 and had approximately 15 months of formal instruction. They reported an average self-rated proficiency of 2 out of 7 and scored an average of 12 out of 40 on the cloze test. Intermediate-high learners began learning Mandarin at an average age of 15 and had studied for approximately 57 months. Their average self-rating was 4 out of 7, and their average cloze score was 20 out of 40. Statistical analysis revealed significant differences among the three groups in both self-rated proficiency and cloze test performance (self-rated proficiency: NS vs IL: $p<0.001$, NS vs. IH: $p<0.001$, IH vs. IL: $p< 0.001$; cloze test: NS vs IL: $p<0.001$, NS vs. IH: $p<0.001$, IH vs. IL: $p<0.005$).

| | Age | Native Language | First Exposure to Mandarin | Months of Mandarin Instruction | Self-rated Proficiency | Cloze Test |
|---|---|---|---|---|---|---|
| IH learners (n=16, F=10) | 20 | English | 15 years old | 57 | 4/7 | 20/40 |
| IL learners (n=11, F=5) | 19 | English | 16.5 years old | 15 | 2/7 | 12/40 |
| NS group (n=15, F=8) | 26 | Mandarin Chinese | 0 years old | - | 6.6/7 | 39.7/40 |

Table 1: Background and proficiency information of participants

2.2. STIMULI. The study includes two sets of words: one pre-test set, and one test set. The pre-test set comprises 16 monosyllabic words, consisting of 8 real words and 8 novel words ((2 real + 2 novel) × 4 lexical tones). All the real words were high-frequency words selected from the 150 Essential Vocabulary Words for Level 1 of HSK (a standardized national test for non-native learners of Mandarin Chinese). The pseudo-words were created by combining non-existing syllables with occurring tones. For example, /tei2/ is composed of the non-occurring syllable *tei* and lexical T2. Such combinations do not exist in Mandarin Chinese and were used to better control for lexical influence, as suggested by Zhang & Peng (2013).

The set of words for the main test consists of twelve minimal pairs of disyllabic words with T2 + T3 and T3 + T3 combinations. Six of these pairs were real words (e.g., *qi2ma3* 'ride a horse' vs. *qi3ma3* 'at least'), and the remaining six were pseudo-words constructed using the pseudo-syllables introduced above (e.g., *tei2bua3* vs. *tei3bua3*). All real-word minimal pairs were adapted from Zhang and Peng (2013) and matched for frequency and syntactic structure. An additional 28 disyllabic sequences were included as fillers, containing all possible tonal combinations except T3T3 and balanced for real and pseudo-word conditions.

Each monosyllable in both the test items and fillers was recorded by a female native Mandarin speaker from northern China. She was instructed to articulate T3 in its full phonetic form, i.e., 213 (as per Zhang & Peng, 2013). These recordings were used to prompt participants during the production tasks.

2.3. PROCEDURE. Two days before the experiment, learner participants were emailed a word list containing all real test and filler items. They were instructed to review the list to ensure familiarity with the vocabulary.

The first pre-test (Pre-Test 1) was a word task, in which participants were given a worksheet listing all 12 real test words and 12 real filler words, with their meanings removed. Participants were asked to write down the meanings. Those who did not achieve full accuracy on the first attempt were given time to review the meanings and then retook the task. All participants eventually achieved 100% accuracy before proceeding to the next task.

The second pre-test (Pre-Test 2) was a monosyllabic repetition task designed to evaluate learners' production of the four lexical tones. Participants listened to each monosyllable from the pre-test set and repeated it as naturally as possible. Stimuli were presented using E-Prime 3, with corresponding pinyin (for all words) and Chinese characters (for real words only) displayed on-screen.

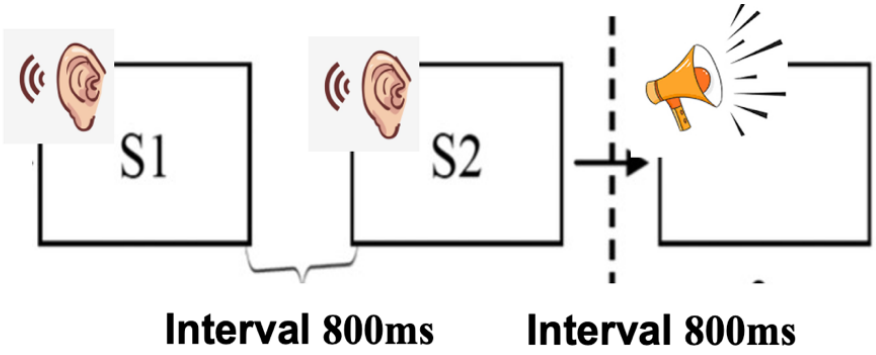


Figure 2: Main test procedure

The main test utilized the wug task, which also employed the repetition task paradigm. Both learner participants and native speakers participated in the main test. In this phase, participants heard the two monosyllables of each disyllabic test unit (i.e., words/sequences in word set 2) separately with an 800 ms interval. They were instructed to produce the two syllables together as a disyllabic word at normal speed after hearing the second syllable (as shown in Figure 2). Again, the pinyin and characters (when available) were presented along with sounds, and practice sessions with instructions were offered before the experimental session.

For both repetition tasks, participants were allowed to correct themselves during the task, and all productions were recorded using Praat. When a word was produced multiple times, the final attempt was selected for analysis.

All tests were done in a quiet lab at a university in the United States. At the end of the experiment, all participants completed the cloze test and the background questionnaire. Each participant received $12 as compensation for their participation, with the entire experiment lasting approximately 40 minutes.

2.4. ANNOTATION AND DATA ANALYSIS. The experiment elicited a total of 720 productions for Pre-Test 2 (16 monosyllabic items × 45 participants) and 1,080 target productions for the main test (12 disyllabic T2+T3 and T3+T3 minimal pairs × 45 participants).

In the first step of analysis, two trained annotators independently listened to and coded each production. For each item, they identified which tone was produced and noted whether the participant inserted a pause between the two syllables. The inter-annotator agreement rate was 89.1%. All discrepancies were resolved through discussion.

As a first step in data cleaning, we excluded productions with clear pauses between the two syllables, following Zhang & Lai (2010), Zhang & Liu (2016), and Qin (2022). Two speakers, one native speaker and one intermediate-low learner, frequently separated the syllables and were excluded from the analysis. For the remaining participants, individual tokens were removed if they exhibited obvious intervals between the first and second syllables.

We then evaluated learners' performance in Pre-Test 2 (the monosyllabic repetition task). One learner's data was excluded due to consistently producing both T2 and T3 as a low, flat tone that lacked tonal contrast. All remaining learners achieved 93.75% or greater accuracy, indicating reliable perception and production of the four Mandarin tones at the monosyllabic level. These participants were included in the main analysis.

Using the annotated data, we then calculated each speaker's sandhi application rate, defined by whether the first syllable of each sandhi sequence was produced as T2.

For the acoustic analysis, vowel boundaries of the first syllables in the target T2+T3 and T3+T3 words were manually annotated in Praat by the first author. F0 values were then extracted at 10 equidistant points within each annotated vowel (Xu, 2013). To control for inter-speaker pitch variation, the F0 values were standardized using z-scores, calculated based on each speaker's mean and standard deviation:

$$\text{Normalized F0} = (\text{Observed F0} - \text{Speaker Mean}) / \text{Speaker SD} \qquad \text{(Qin, 2022)}$$

The resulting normalized pitch contours for T2 and sandhi T3 were analyzed using Generalized Additive Mixed Models (GAMMs), implemented via the bam() function in the *mgcv* package (Wood, 2011, 2017). GAMMs were selected for their ability to model dynamic time-series data, allowing us to track F0 contour evolution over time and identify when and where contours diverge across conditions and groups.

Two models were constructed. The first model examines group differences in the realization of sandhi T3 across word conditions (real vs. pseudo). The dependent variable was the normalized F0 across 10 time points of the first syllable of sandhi words. Fixed effects included Group, Word Condition, and their interaction. Participant and Word are included as non-linear random effects. The second model used the same structure to analyze T2 production in T2+T3 sequences. Autocorrelation structures were included to reduce temporal dependencies in the data, and model criticism procedures were carried out to identify and address potential misfits, ensuring the robustness and interpretability of the results.

## 3. Results

3.1. SANDHI APPLICATION RATES. Under the real word condition, native speakers achieved a 100% application rate, the IL group achieved 85%, and the IH group achieved 89%. In the pseudo-word condition, native speakers again achieved 100%, while the IL group reached 77% and the IH group 80%. Statistical analysis revealed a significant effect of Group: both learner groups had significantly lower sandhi application rates than did native speakers (NS vs IL: $p<0.001$; NS vs. IH: $p<0.001$). However, no significant difference was found between the two learner groups. Additionally, no significant effects of Word Condition or the interaction between Group and Word Condition were observed.

3.2. ACOUSTIC REALIZATIONS. This section presents the results of acoustic analysis for both sandhi T3 and T2 in T2+T3 combinations.

3.2.1. GROUP DIFFERENCES IN SANDHI REALIZATION. Figure 3 illustrates the F0 contours of sandhi T3 production by different groups across word conditions. A visual comparison

reveals that NSs tend to exhibit an earlier F0 turning point and a larger ΔF0. In other words, they show more pronounced rising contour. IH learners' sandhi productions more closely resemble those of NSs under the real word condition compared to the pseudo-word condition. In contrast, IL learners' acoustic realizations of sandhi differ from those of NSs in both conditions.

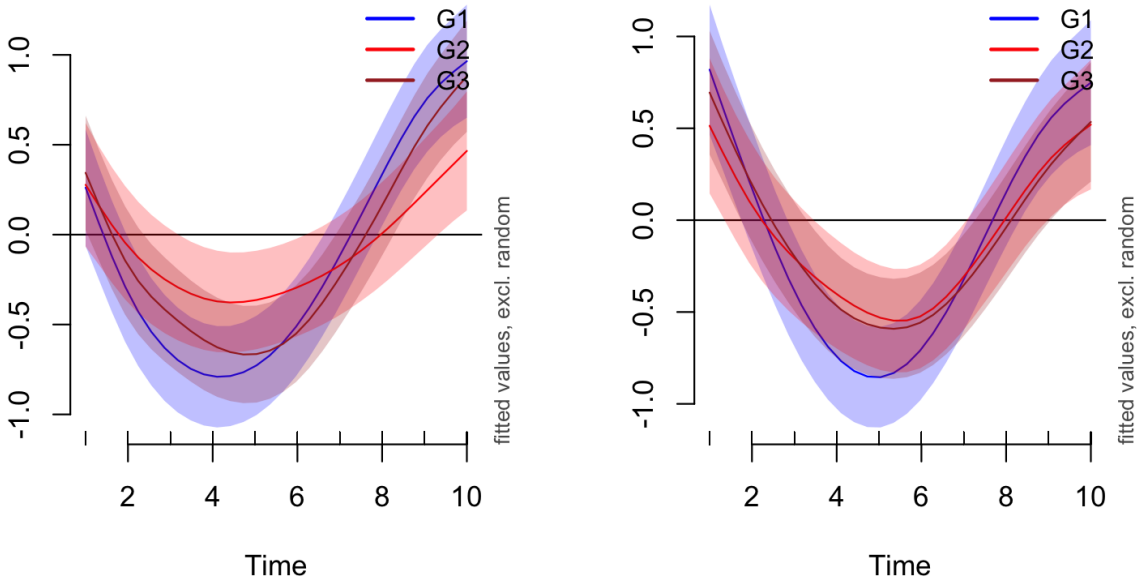


Figure 3: Sandhi contours of different groups (left: real words; right: pseudo-words; G1–G3 = NS, IL, IH groups, respectively)

To assess whether these F0 differences are statistically significant, we examined the difference curves using GAMMS. Figure 4 displays the estimated difference curves for sandhi T3 between groups under the real word condition. The 95% confidence interval is represented by a shaded band. Significant differences, where the confidence band does not overlap with the x-axis (i.e., the value is significantly different from zero), are marked by red lines along the x-axis and vertical dotted lines. The results show that under the real word condition, IH learners' acoustic

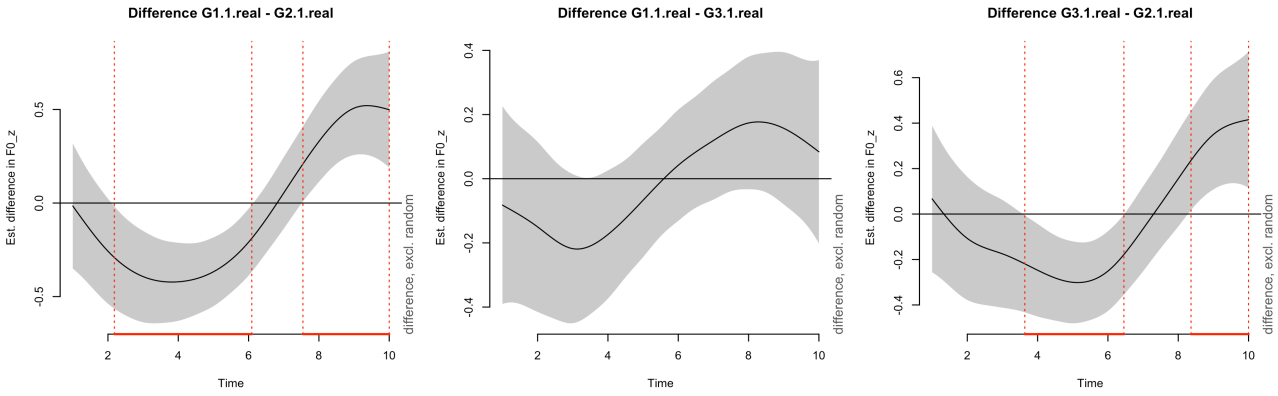


Figure 4: Estimated differences in normalized F0 (z-scored) trajectories for sandhi T3 between groups under the real word condition (left to right: NS vs. IL, NS vs. IH, IH vs. IL)

realization of sandhi T3 did not significantly differ from that of NSs. However, significant differences were observed between NSs and IL learners, specifically from timepoints 2.5 to 6 and 8 to 10. NSs exhibited significantly lower pitch values from timepoints 2.5 to 6 and significantly higher pitch values from timepoints 8 to 10 than did IL learners. Additionally, IH learners also showed significantly lower pitch in the middle portion and higher pitch toward the end of the contour relative to IL learners.

These results are consistent with the visual patterns observed in Figure 3: both NSs and IH learners produced a more pronounced rising contour than IL learners under the real word condition, and NSs showed a steeper rise than both IH and IL learners under the pseudo-word condition.

Under the pseudo-word condition, both intermediate-low and intermediate-high learners produced contours that significantly differed from native speakers, but the duration of the difference was slightly shorter for intermediate-high learners (see Figure 5). No significant difference was found between the two learner groups.

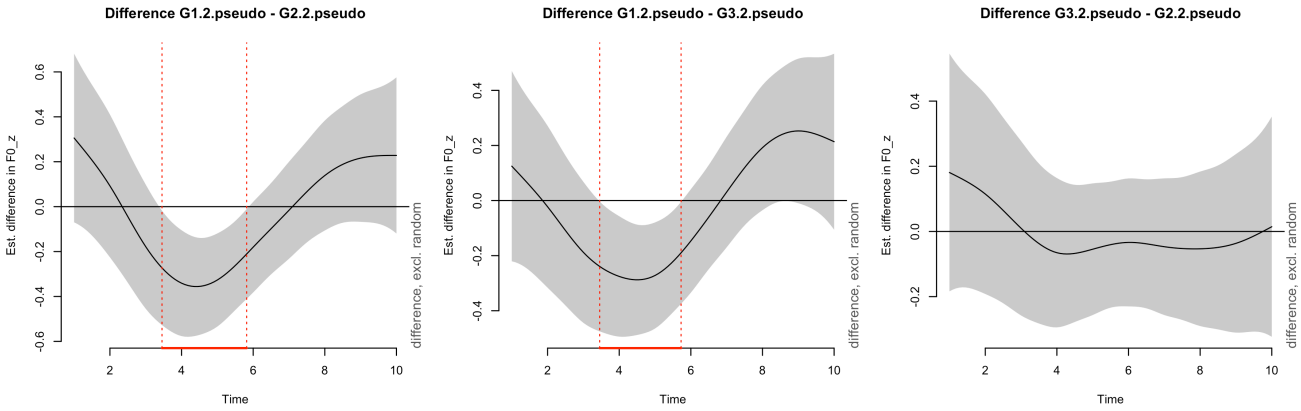


Figure 5: Estimated differences in normalized F0 (z-scored) trajectories for sandhi T3 between groups under the pseudo-word condition (left to right: NS vs. IL, NS vs. IH, IH vs. IL)

3.2.2. WORD CONDITION DIFFERENCES IN SANDHI REALIZATION. GAMM analysis also revealed word condition effects within groups (see Figure 6). The NS and IH groups showed a similar pattern: significantly lower pitch at the beginning of the contour (timepoints 0 to 3) and

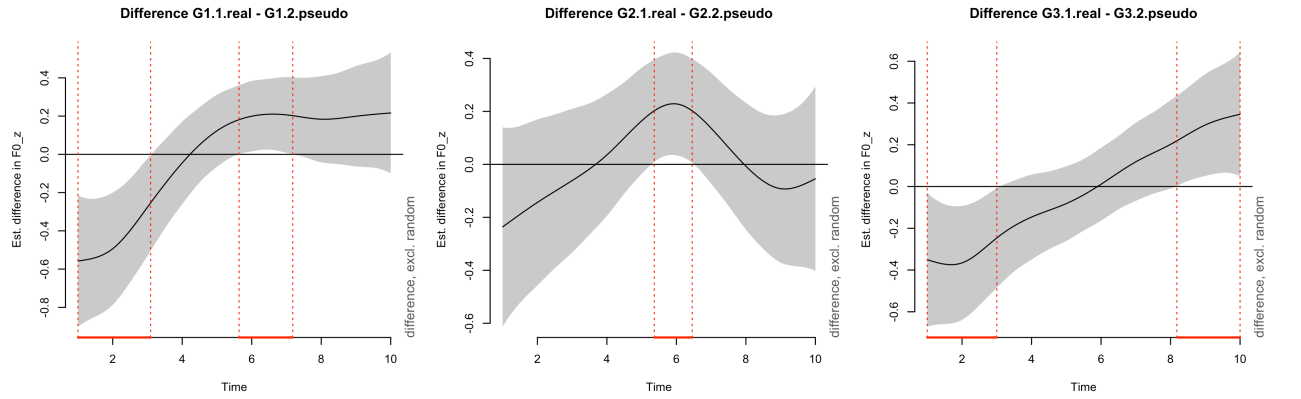


Figure 6: Estimated differences in normalized F0 (z-scored) trajectories for sandhi T3 between word conditions for each group (left: NS, middle: IL, right: IH)

significantly higher pitch later in the contour, timepoints 4.5 to 7 for NSs and 8 to 10 for IH learners, under real word condition. This pattern suggests that both groups produced a more pronounced rising contour in the real word condition. For IL learners, significant differences emerged around timepoints 5.5 to 6; their pitch was significantly higher within this short time span in the real word condition.

3.2.3. GROUP DIFFERENCES IN T2 REALIZATION. To determine whether the observed effects were specific to sandhi application or reflected more general differences in tonal production, a second model analyzed the production of T2 across groups and word conditions.

As shown by Figure 7, under the real word condition, native speakers and intermediate-low learners significantly differed in their T2 realizations from timepoints 0 to 2. Native speakers produced significantly lower pitch than did intermediate-low learners. No significant differences were found between native speakers and intermediate-high learners, nor between intermediate-low and intermediate-high learners under this condition.

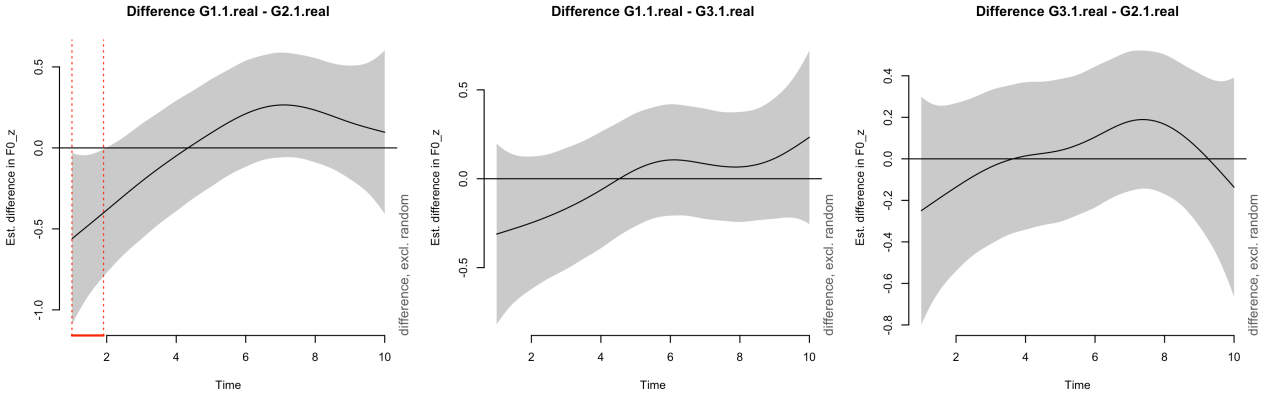


Figure 7: Estimated differences in normalized F0 (z-scored) trajectories for T2 between groups under the real word condition (left to right: NS vs. IL, NS vs. IH, IH vs. IL)

Under the pseudo-word condition, both learner groups significantly differed from native speakers. Native speakers exhibited significantly lower pitch at the beginning of the contour compared to both learner groups, and significantly higher pitch at the end than intermediate-low learners

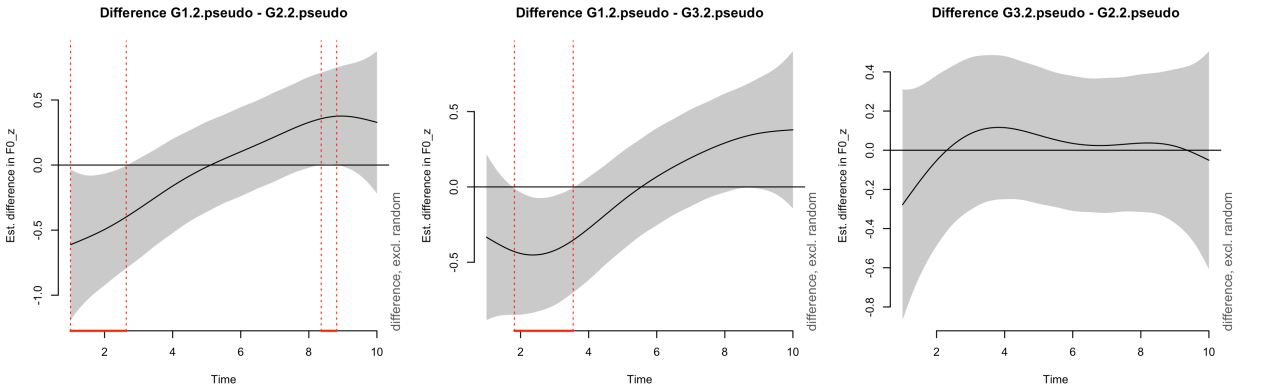


Figure 8: Estimated differences in normalized F0 (z-scored) trajectories for T2 between groups under the pseudo-word condition (left to right: NS vs. IL, NS vs. IH, IH vs. IL)

(see Figure 8). No significant difference was found between the two learner groups under the pseudo-word condition.

3.2.4. WORD CONDITION DIFFERENCES IN T2 REALIZATION. Word condition effects for T2 production were found across all three groups. Across groups, speakers tended to exhibit lower pitch at the beginning and higher pitch at the end of the contour in the real word condition than in the pseudo-word condition (see Figure 9).

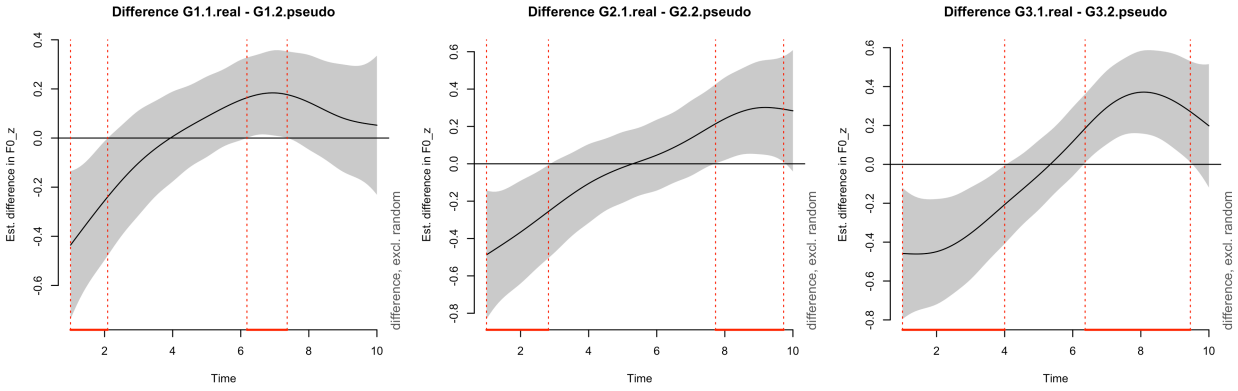


Figure 9: Estimated differences in normalized F0 (z-scored) trajectories for T2 between word conditions for each group (left: NS, middle: IL, right: IH)

**4. Discussion**. This study investigated four main questions: (1) whether native speakers and L2 learners apply the Mandarin T3 sandhi rule productively across lexical contexts; (2) how lexical familiarity and online processing demands influence rule application; (3) how learners' sandhi application rates and acoustic realizations differ from those of native speakers; and (4) how learner proficiency level affects both sandhi application and realization.

First, our findings showed that native speakers applied T3 sandhi at a 100% rate even to nonce words, confirming the high productivity of the rule among native speakers (Zhang & Lai, 2010; Zhang & Peng, 2013). Both learner groups applied the sandhi rule to nonce words over 77% of the time. This high application rate suggests that even intermediate-low level L2 learners have internalized the sandhi rule and were able to compute sandhi forms, rather than simply memorizing the tones of sandhi words as fixed chunks. This supports the view that L2 learners can acquire abstract phonological rules in the target language (Shea & Curtin, 2006) and is consistent with findings in L1 phonological acquisition, where regular and transparent alternations are learned relatively easily and can be applied online (Kerkhoff, 2007).

It is important to note that the application rate of T3 sandhi among L2 learners in our study was much higher than the rates reported by Yang (2016) and Zhang (2016). Yang (2016) found an application rate of 51.25% for regular speech and 60% for fast speech when applying the sandhi rule to nonce words among intermediate-level learners. Zhang (2016) reported a 68% accuracy for sandhi T3 in intermediate learners. The difference may be attributed to the different tasks used in their studies compared to ours. Both Zhang (2016) and Yang (2016) used word reading tasks, which may add complexity to sandhi production, as learners first need to retrieve the contour of T3 before applying the sandhi rule. Additionally, word reading tasks may prompt learners to articulate each syllable more clearly and pay more attention to the underlying T3 contour, leading to the production of the underlying T3 rather than the sandhi T3 even in sandhi contexts. In contrast,

our study employed a repetition task in which participants heard the target syllables and simply repeated them, reducing the cognitive load associated with tone retrieval and potentially facilitating more accurate and spontaneous sandhi application.

Our findings also shed light on the cognitive mechanisms underlying sandhi production. Two mechanisms have been debated in the literature: the computation-based mechanism, where surface forms are derived based on phonological context, and the storage-based lexical mechanism, where context-appropriate forms are retrieved from memory. The high sandhi application rates for nonce words and the absence of application rate differences between real and pseudo-words strongly support the computation-based mechanism. However, acoustic analyses revealed that native speakers and intermediate-high learners exhibited significant differences in sandhi realization between real and pseudo-word conditions: both groups produced lower pitch at the beginning and higher pitch at the end under the real word condition, and a similar pattern was observed for T2 production. These results suggest that computation governs the initial application of T3 sandhi, while storage and lexical memory shape the fine-grained phonetic realization for familiar words at a later stage. Intermediate-low learners also showed differences in T2 production across lexical conditions but exhibited only a minimal difference for sandhi realization at a narrow time window (around timepoint 5.5–6). This developmental pattern indicates that computation-based processing of T3 sandhi emerges early in L2 acquisition, while lexical effects on tonal realization develop more selectively—emerging earlier for base tones like T2, and more gradually for derived forms like sandhi T3, likely due to its added processing complexity.

These findings align with Chen et al. (2019) in supporting the computation-based mechanism for both native speakers and L2 learners. However, whereas Chen et al. found no statistical differences in sandhi realization between real and nonce words, our study revealed consistent acoustic differences. One possible explanation is that the real words used in our study were highly familiar, which may have strengthened the lexical effect on acoustic realization.

In addition, our results offer new insight into an interpretation from Zhang and Lai (2010). They previously attributed real-pseudo word differences to incomplete sandhi application for nonce forms. However, we observed real-pseudo differences not only for sandhi T3, but also for T2, which does not undergo sandhi. This pattern indicates that the differences are more likely to reflect general lexical effects on tonal production rather than incomplete application of the sandhi rule. In other words, while the sandhi rule is successfully applied across both real and pseudo contexts, lexical familiarity subtly shapes the fine-grained phonetic realization of tones.

Proficiency effects were also evident. Although intermediate-high learners did not significantly outperform intermediate-low learners in sandhi application rates, their acoustic realizations of both sandhi T3 and T2 were more native-like. Intermediate-high learners showed no significant difference from native speakers in T2 production under real word conditions, whereas intermediate-low learners did, suggesting more stable tonal production among higher-proficiency learners. These findings support Jin's (2019) concept of a learning effect, where greater exposure and practice lead to more native-like phonological competence over time. However, under pseudo-word conditions, both groups showed deviations from native norms, indicating that increased cognitive load still challenges tonal stability at intermediate proficiency levels.

**5. Conclusion**. This study examined the production of Mandarin T3 sandhi among native speakers and L2 learners, analyzing both application rates and acoustic realizations. The results confirmed the high productivity of the sandhi rule among native speakers and demonstrated that even intermediate-low learners could apply the rule productively to novel sequences. These findings highlight that regular and transparent phonological rules, such as T3 sandhi, can be acquired

relatively early (at intermediate low level) in L2 learning. In addition, high productivity in the pseudo-word condition and the absence of application rate differences across word conditions provide strong support for a computation-based mechanism underlying sandhi application.

Acoustic analyses further revealed that intermediate-high learners demonstrated more native-like sandhi realizations than intermediate-low learners, though differences from native speakers persisted under pseudo-word conditions. This suggests that sandhi application becomes increasingly reliable with higher proficiency but continues to be influenced by task demands and cognitive load.

The study also found that both native speakers and intermediate-high learners produced sandhi T3 differently in real versus pseudo-word contexts, but these differences likely reflect general lexical effects on tonal production, rather than incomplete application of the sandhi rule. In contrast, intermediate-low learners showed minimal real-pseudo differences in sandhi realization, suggesting that early stages of L2 acquisition primarily rely on computation-based processing, with lexical effects on tonal realization emerging only at later stages of development.

Future research could expand on these findings by examining real words of varying lexical frequencies or by incorporating neurolinguistic methods to further explore the interplay between computation and lexical storage mechanisms in phonological production.

This study contributes to the fields of L1 phonology and L2 phonological acquisition by providing empirical evidence that real-pseudo differences in native sandhi production stem from lexical effects rather than incomplete rule application. It also demonstrates that L2 learners are capable of generalizing phonological rules to novel contexts and offers a comprehensive view of how both rule productivity and acoustic realization develop across different proficiency levels in second language acquisition.

## References


Arndt-Lappe, Sabine & Mirjam Ernestus. 2020. Morpho-phonological alternations: The role of lexical storage. In Vito Pirrelli, Ingo Plag & Wolfgang Dressler (eds.), *Word Knowledge and Word Usage: A Cross-disciplinary Guide to the Mental Lexicon*, 185–220. De Gruyter.

Chang, Yung-hsiang Shawn. 2011. Distinction between Mandarin Tones 2 and 3 for L1 and L2 listeners. *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, *1*, 84–96.

Chao, Yuan Ren. 1948. *Mandarin primer*. Cambridge: Harvard University Press.

Chen, Si, Yunjuan He, Ratree Wayland, Yike Yang, Bei Li & Chun Wah Yuen. 2019. Mechanisms of tone sandhi rule application by tonal and non-tonal non-native speakers. *Speech Communication*, *115*, 67–77. doi:10.1016/j.specom.2019.10.008

Cole, Jennifer & José Ignacio Hualde. 2011. Underlying representations. In *the blackwell companion to phonology*, 1–26. https://doi.org/10.1002/9781444335262.wbctp0001

Jin, Wenhua. 2019. Acquisition of tone sandhis by English speaking learners of Chinese. *Journal of National Council of Less Commonly Taught Languages*, *25*(1), 67–107.

Kenstowicz, Michael & Charles Kisseberth. 1979. *Generative phonology: Description and Theory*. Academic Press.

Kerkhoff, Annemarie Odilia. 2007. *The acquisition of morpho-phonology: The Dutch voicing alternation.* Unpublished doctoral dissertation, Utrecht University.

Qin, Zhen. 2022. The second-language productivity of two Mandarin tone sandhi patterns. *Speech Communication*, *138*, 98–109.

Shea, Christine & Suzanne Curtin. 2006. Learning allophonic alternations in a second language: Phonetics, phonology and grammatical change. In *Proceedings of the 8th Conference on Generative Approaches to Second Language Acquisition*. Banff, Alberta.

Shen, Xiaonan Susan. 1990. Tonal coarticulation in Mandarin. *Journal of Phonetics*, *18*(2), 281–295.

Shih, Chilin. 2007. *Prosody Learning and Generation*. Berlin: Springer.

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61–83.

Xu, Yi. 2013. ProsodyPro - A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 7–10.

Yang, Chunsheng. 2016. *The acquisition of second language Mandarin prosody: From experimental studies to pedagogical practice*. John Benjamins Publishing Company, Amsterdam.

Zhang, Caicai & Gang Peng. 2013. Productivity of Mandarin third tone sandhi: A wug test. In G. Peng & F. Shi (eds.), *Eastward Flows the Great River: Festschrift in Honor of Prof. William S-Y. Wang on his 80th Birthday*, 256–282. City University of Hong Kong Press, Hong Kong.

Zhang, Jie. 2016. Using nonce-probe tests and auditory priming to investigate speakers' phonological knowledge of tone sandhi. *Proceedings of the 5th International Symposium on Tonal Aspects of Languages,* 12–18.

Zhang, Jie & Yuwen Lai. 2010. Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(1), 153–201. http://www.jstor.org/stable/40783035

Zhang, Jie & Jiang Liu. 2016. The productivity of variable disyllabic tone sandhi in Tianjin Chinese. *Journal of East Asian Linguistics*, *25*, 1–35. https://doi.org/10.1007/s10831-015-9135-0.

Zhang, Jie, Caicai Zhang, Stephen Politzer-Ahles, Ziyi Pan, Xunan Huang, Chang Wang, Gang Peng & Yuyu Zeng. 2022. The neural encoding of productive phonological alternation in speech production: Evidence from Mandarin Tone 3 sandhi. *Journal of Neurolinguistics*, *62*, 101060.