Phonetically incomplete neutralization can be phonologically complete: Evidence from Huai'an Mandarin*

Naiyan Du & Karthik Durvasula

The phenomenon of Incomplete neutralization describes a situation where a putative case of categorical phonological neutralization is observed to be phonetically non-neutralizing. This has been argued to be a problem for phonological theories that employ categorical features. Here we use two distinct feeding orders of tone sandhi processes from Huai'an Mandarin to show that incomplete phonetic neutralization is compatible with categorical phonological phenomena. Therefore, incomplete phonetic neutralization does not automatically inform us of gradient phonological representations. We further show that incomplete phonetic neutralization can in fact have a large effect size. Such results are not surprising from a classic generative view of phonology where linguistic performance is argued to be a multi-factorial problem, and linguistic knowledge (i.e., competence) is only one of the many factors involved. Furthermore, our results suggest that the observed incompleteness or gradience may have a source outside phonological knowledge.

1. Introduction

As has long been recognized in discussions of linguistic competence as abstract knowledge, there are multiple potential interacting factors in performance (Chomsky, 1964, 1965; Schütze, 1996; Valian, 1982; inter alia). Similarly, there are also multiple interacting sources that affect speech production, e.g. lexical knowledge, phonological knowledge, memory and processing constraints, etc. (Warner et al., 2004; Whalen, 1991, 1992; Wright, 2004). Consequently, gradience in phonetic manifestations cannot automatically be used as a diagnostic of gradient phonological representations. In this paper, we explore the phenomenon of incomplete neutralization to argue that incomplete phonetic neutralization does not automatically inform us about phonological representations or phonological knowledge, more generally. We present relevant data from Tone 1 and Tone 4 sandhi processes in Huai’an Mandarin (Huai’an hereafter), both of which crucially participate in feeding orders to trigger other tone sandhi processes, to argue that phonetically incomplete neutralization can still be phonologically complete.

Since at least the mid-1980s, the effect of incomplete neutralization has been documented in a variety of languages including Catalan (Dinnsen & Charles-Luce, 1984), Dutch 12/21/2022 7:17:00 PM, Japanese (Braver & Kawahara, 2016), Polish (Slowiaczek & Dinnsen, 1985; Slowiaczek & Szymanska, 1989) and Russian (Dmitrieva, 2005; Kharlamov, 2012; Matsui, 2015). For example, in German it has been described that the phonological voicing contrast for obstruents is neutralized at the right edge of a prosodic word (Wagner, 2002). A rule-based mapping of the relevant phonological process is stated in (1). However, careful phonetic and perceptual

* We are grateful to our two research assistants Zhang Huaying and Yang Chenyuliang for their diligent work in this project. We also thank an associate editor of Phonology and three anonymous reviewers, whose comments and critiques have improved the quality and clarity of this paper greatly. For helpful discussions, we thank Yen-Hwei Lin, Silvina Bongiovanni, Suzanne Wagner and members of the Michigan State University Phonology & Phonetics Group, as well as audiences at the 2020 Berkeley Linguistic Society Workshop and the Annual Meeting Phonology 2020, where the initial experimental results of this paper were presented.

experimentation has shown that the neutralization is incomplete phonetically (Port & O'Dell, 1985; Roettger et al., 2014; inter alia). In other words, *underlying* voiceless stops, *derived* voiceless stops and *underlying* voiced stops all have different phonetic distributions.[1]

(1) [- sonorant] → [-voice]/ __)ω
Note: 'ω' means prosodic word.

The observed effect of incomplete neutralization has been argued by many researchers to pose a challenge to traditional formal phonology where categorical phonological representation and modular feed-forward model are assumed (Braver, 2019; Goldrick & Blumstein, 2006; McCollum, 2019; Port & Leary, 2005; Manaster Ramer, 1996; Roettger et al., 2014). It is often assumed that under such a feed-forward framework, the phonological representations are discrete elements that do not contain any gradient phonetic information, and phonetics *only* has access to the output of phonology (Kenstowicz, 1994; Pierrehumbert, 2002). We call this view the *Standard generative view of Phonology*.[2] As a result, underlying representations that undergo phonological neutralization process should not have any consequence on phonetic manifestations. However, in cases of incomplete neutralization, there are traces of the underlying representation in the phonetic manifestation.

One trivial but widely-adopted solution to the puzzle posed by incomplete neutralization has been to simply deny that such an effect can be caused by grammatical knowledge (Dinnsen & Charles-Luce, 1984; Fourakis & Iverson, 1984; Manaster Ramer, 1996; Warner et al., 2004; inter alia). To support this claim, several criticisms have been raised against previous experimental designs as well as the interpretation of the results. One main criticism is whether the observed phenomenon of incomplete neutralization is due to task effects. Among these effects, the most discussed one is orthography. It has been noticed by many researchers (Fourakis & Iverson, 1984; Manster Ramer, 1996; inter alia) that in the seminal work of Port and O'Dell (1985), participants were presented stimuli orthographically where minimal pairs were always in contrast. Native speakers of German may have hypercorrected and produced unnatural speech to match the forms of orthography. This suspicion becomes especially disturbing when Warner et al. (2004) showed a significant production difference in words that are identical in underlying representations but differ in orthography in Dutch.[3] To circumvent the interference of orthography, two methods have been employed, namely changing the experimental paradigm and looking at languages where the relevant phonological contrast is not reflected orthographically.

In line with changing the experimental paradigm, Fourakis and Iverson (1984) employed a unique strategy aimed at concealing the morphological forms that native speakers of German are supposed to produce, so the influence of orthography is expected to decrease. The participants were instead presented auditorily with the conjugated form where the contrast is maintained in both underlying and surface representations and asked to decompose and produce the bare form

---

[1] To be succinct, here we use '*underlying* voiceless stop' to mean surface voiceless stops that map from *underlying* voiceless stops, '*derived* voiceless stop' to mean surface voiceless stops that map from *underlying* voiced stops, and '*underlying* voiced stops' to mean surface voiced stops that map from *underlying* voiced stops. We will use the terms '*underlying*' and '*derived*' in the same fashion to describe stops and tones for the rest of the paper.

[2] While this conception of phonology is relatively standard in our opinion, it is actually quite different from the *Classic generative view of Phonology* where phonology is seen as *knowledge*. We return to this issue in the Conclusion, where we point out that the latter notion has no trouble in accounting for facts related to incomplete neutralization.

[3] In support of our larger point in this paper, we would like to point out that their observation in fact shows how powerful performance (i.e., non-phonological) factors can be in accounting for phonetic manifestations.

where the incomplete neutralization is expected to happen. Through this paradigm, Fourakis and Iverson camouflaged the task as a morphological exercise to distract the participants to elicit more natural pronunciations. Interestingly, the effect of incomplete neutralization was not observed, and Fourakis and Iverson concluded that the previously found incomplete neutralization was actually a task effect. Using a different strategy, Jassem and Richter (1989) asked participants to answer questions designed to elicit the target words in Polish, and observed no evidence of incomplete neutralization. However, by implementing the same strategy as Fourakis and Iverson (1984) and increasing the statistical power with more speakers and more test minimal pairs, Roettger et al. (2014) found an effect of incomplete neutralization. However, it is worth noting that, as Roettger et al. (2014) themselves pointed out, the strategy employed by Fourakis and Iverson can incur a potential artifact of phonetic accommodation. In the experimental paradigm, the participants hear the conjugated form where neutralization cannot happen and the voicing contrast is present, and have to produce the form where neutralization does happen. In such a paradigm, the observed effect of incomplete neutralization may be due to the participants mirroring vowel duration differences in the stimulus recordings they heard of the conjugated forms. When Roettger et al. (2014) controlled for this confound in one of their experiments, they only found a very small, non-significant effect (< 3 ms) in the right direction. This suggests that there might indeed be no clear evidence for incomplete neutralization even in their well-powered study. To sum up, this general strategy to solve the problem of orthography by changing the task performed by the participants leads to very weak evidence (if that) for the presence of incomplete neutralization.

A second method employed to overcome task effects related to orthography has been to use a language where the crucial contrast is not marked in the orthography. For example, Catalan has been claimed to be a language that has a devoicing process but does not reflect an underlying voicing contrast orthographically under any phonological conditions, and Dinnsen and Charles-Luce (1984) did not observe any evidence of incomplete neutralization in the devoicing process of Catalan. However, later Charles-Luce and Dinnsen (1987) reanalyzed their data and found incomplete neutralization in the cue of voicing into closure. Here, it is worth noting that in Catalan, quite a few words actually maintain the underlying voicing contrast in orthography, so the real situation is more complicated and Catalan cannot simply be treated as a language that does not mark underlying voicing contrast in orthography (Badia Margarit, 1962; Manaster Ramer, 1996).

In another case, Braver and Kawahara (2016) observed a putative case of incomplete neutralization in Japanese. Most of their stimuli were presented in Chinese characters (Kanji), which is an orthographical system that is commonly used in Japanese but only has a very weak connection with pronunciation.[4] Although most Chinese characters were originally created by combining a part that indicates pronunciation and a part that indicates meaning (Yang, 1995), the connection between characters and pronunciation is largely obscured by historical sound change and character change (Huang & Liao, 2017). In Japanese, most Chinese characters are used to represent both borrowed words from China (Sino-Japanese lexicon) and words that are originated in Japan (Yamato lexicon) (Itô & Mester, 1999; Japan Broadcasting Corporation, 1998), and the resulting multiple pronunciations (onyomi and kunyomi) of many Chinese characters can only further weaken the connection between Chinese characters and pronunciations. So, it is hard to

[4] In Braver and Kawahara's experiment, 10 out of 13 sets of stimuli were presented only in Chinese characters while the other 3 sets partially contained Kana (2 in Katakana and 1 in Hiragana). In current usage, Kana refers to two syllabaries where each character represents a mora, which is an onset-vowel combination or a coda or the second half of a long vowel in Japanese phonology. It is also worth noting that geminates and some long vowels are marked with diacritics in Kana system. Braver and Kawahara observed incomplete phonetic neutralization for each set of stimuli.

imagine that Japanese speakers hypercorrected based on Chinese characters, and Braver and Kawahara still appeared to observe incomplete neutralization in monomoraic prosodic word lengthening process.

To sum up, although the case of Catalan is controversial, the case of Japanese provides good evidence that at least in some languages, the observed incomplete neutralization is not caused by orthographic knowledge.

Another source of criticism of incomplete neutralization is that the observed effect size is typically quite small. As a consequence, such a small effect size has been argued to likely not be functionally significant and therefore not in need of a grammatical explanation (Dinnsen & Charles-Luce, 1984; Mascaró, 1987; Warner et al., 2004).[5] For example, among the phonetic cues examined by Port and O'Dell (1985), preceding vowel duration before underlying voiced stops was only about 15 ms longer than that before underlying voiceless stops, voicing into closure of derived voiceless stops was only 5 ms longer than that of underlying voiceless stops, and duration of aspiration noise before underlying voiceless stops was only 15 ms longer than that of derived voiceless stops. Similar effect sizes were also found in Polish (Jassem & Richter, 1989; Slowiaczek & Dinnsen, 1985), Dutch (Warner et al., 2004), and two other studies on German (Piroth & Janker, 2004; Roettger et al., 2014). To summarize the discussion on the criticisms on incomplete neutralization, the debate on the existence of incomplete neutralization is still pretty much ongoing, especially with respect to the issue of effect size.

In this paper, as mentioned above, we will argue using data from Huai'an that incomplete phonetic neutralization can stem from phonologically complete neutralization. By using Huai'an, we avoid the orthographic confound discussed above as the stimuli can be presented in Chinese characters, an orthographical system that only has a weak connection with pronunciation (note, this is similar to the Japanese case we discussed above.). Furthermore, the language allows us to argue that effect sizes are tangential to the issue of phonological neutralization. Anticipating our results, we show that despite there being a rather large phonetic difference with respect to incomplete phonetic neutralization, there is clear evidence that the relevant processes are phonologically categorically neutralizing as evidenced by the fact that their outputs feed other sandhi processes.

2. The issue of phonological neutralization versus phonetic implementation

As introduced in Section 1, the definition of incomplete neutralization is two-fold, which includes phonological neutralization and phonetically incomplete neutralization. An issue of many previous studies of incomplete neutralization is that researchers do not typically show evidence that the examined processes are truly phonological neutralization, as opposed to phonetic implementation (Cohn, 1993; Dunbar, 2013). Under the categorical view of phonological representations, phonological neutralization entails there is a change from one phonological category to another phonological category, while phonetic implementation does not result in any categorical changes. To give an example, it is assumed by Port and O'Dell (1985) and other previous studies on German that the devoicing process results in a voiceless obstruent category in

[5] To be clear, we are not claiming that the effect size must be large for incomplete neutralization that is caused by grammatical knowledge. We are only recognizing here that it is a reasonable concern that incomplete neutralization with a small effect size may not even be captured by grammatical knowledge and therefore may not be able to pose any challenges to traditional formal phonology.

the phonological surface form.[6] However, there is no clear evidence, especially evidence from phonological behavior, that shows a derived voiceless obstruent is actually neutralized with the underlying voiceless obstruent in the phonology. If Dunbar's (2013) suspicion that the word-final devoicing in German is actually a phonetic implementational process turns out to be valid, then the so-called 'devoiced obstruent' at the right edge of prosodic word remains phonologically unchanged and still belongs to the same 'voiced' category with voiced obstruents in other positions. As a result, it would not be surprising according to the *Standard generative view of Phonology* that the so-called 'devoiced obstruent' is phonetically different from an underlying voiceless obstruent since they are phonologically different, i.e., different in the surface representations.

To the best of our knowledge, the only careful previous study that attempted to establish phonological neutralization using evidence from phonological behavior is Braver and Kawahara's study on the lengthening of prosodic words in Japanese (Braver & Kawahara, 2016). In Japanese, since a prosodic word has been argued to be at least bimoraic (Itô, 1990; Itô & Mester, 2003; Mester, 1990; Mori, 2002; Poser, 1990), an *underlying* monomoraic prosodic word has been argued to lengthen to be bimoraic. In relation to this, Braver and Kawahara (2016) showed that this neutralization is incomplete phonetically, i.e. a derived bimoraic prosodic word is still shorter than an underlying bimoraic prosodic word.

The current paper utilizes a different strategy of examining rules in feeding orders to establish phonological neutralization. The fact that the output of a process can trigger another process provides evidence that the first process results in complete neutralization in the phonology. Yet despite there being complete neutralization in the phonology, we will show that there is incomplete neutralization in the phonetics for each of the feeding processes in Huai'an tone sandhi processes.[7] We will elaborate the feeding orders in Huai'an in Section 3 with more background information. And Section 4 and Section 5 will present the two experiments we have run based on two feeding orders in Huai'an.

3. Background

Huai'an belongs to the Jianghuai Guanhua Group (Lower Yangtze Mandarin) of the Mandarin language family. Native speakers are mainly from Huai'an city, which is located in the northern part of Jiangsu Province (Li, 1989). Huai'an has four phonemic tones, labelled as Tone 1, Tone 2, Tone 3 and Tone 4 (Jiao, 2004; Y. Wang & Kang, 2012). Following the tradition of tone description in Chinese languages, in Table *1*, the four tones are given in tone letters using a scale of 1 to 5 where 1 is the lowest f0 and 5 is the highest f0 and followed by a contour description in

[6] To be fair to them, this is likely what they assumed since many phonologists have claimed as much in the extant literature.

[7] As pointed out by an anonymous reviewer, Ernestus et al. (2006) showed that Dutch has an optional progressive voice assimilation process across word boundary for obstruents. A word-final devoicing process feeds into the assimilation process and causes the initial obstruent of the next word to become voiceless, which suggests that the devoiced obstruent in the preceding word belongs to the same phonological category as its underlyingly voiceless counterpart. In relation to this, Ernestus and Baayen (2006) showed in a separate experiment that there is incomplete phonetic neutralization in word-final devoicing process in Dutch (see also Warner et al., 2004). Further study is needed to show that a devoiced obstruent that can trigger another devoicing process is actually incompletely neutralized in the phonetics. If this is indeed the case, Dutch will serve as another clear case of incomplete phonetic neutralization under the condition of complete phonological neutralization.

words (Chao, 1930).[8] The tonal contours of phonemic tones in isolation are given in Figure *1*. The speaker (male, age: 53) pronounced 4 repetitions of four monosyllabic morphemes that share the same segmental content ([sɔ]) and only contrast in the tone on the vowel. f0 was extracted only from the vowel at 5% steps with a script in Praat (Boersma & Weenink, 2021). However, it is worth noting that the tonal contours in isolation for Mandarin tones have been noticed to be quite different when compared with their counterparts in context (Jongman et al., 2006; Shen, 1990; Xu, 1994, 1997; inter alia). So, we expected the same kind of differences in our experiments where tones are pronounced in sentences.

Phonemic tone	Tone letter	Contour description
Tone 1	42	high falling
Tone 2	24	high rising
Tone 3	312	low/low rising
Tone 4	55	high level

Table 1: Descriptions of phonemic tones in Huai'an

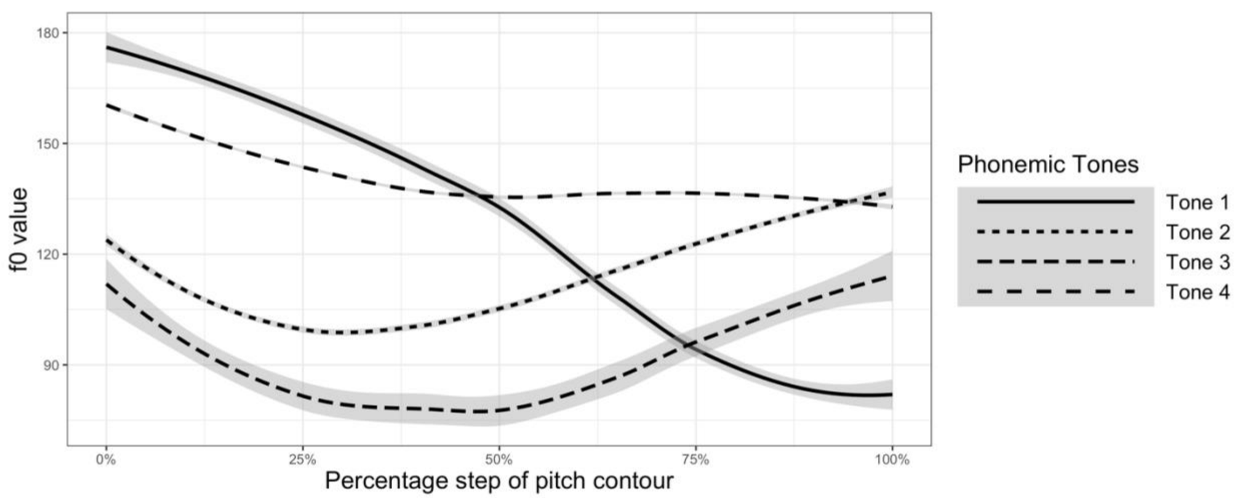


Figure 1: Tonal Contour of Phonemic Tones in Huai'an

For the examples in the rest of the paper, to make it easier on the reader, we will only use the T plus tone number to refer to tones. For example, 'T3' refers to Tone 3. We will however continue to use the full form Tone 3 in the text.

The three tone sandhi rules that are related to the current paper are shown in (2). At the post-lexical level, the low-register Tone 3 sandhi is mandatory in some contexts and optional in others (we will elaborate on this at a later point in this paper). In contrast, the high-register Tone 1 and Tone 4 sandhis are always optional. Furthermore, Tone 3 undergoes tone sandhi to become Tone 2, and this Tone 3 sandhi process can only happen when immediately preceding Tone 3 (underlying or derived).[9] As dissimilation processes, the tone sandhis in Huai'an can be straightforwardly

[8] Note, Huai'an phonemic tones are different from those in Standard Mandarin. In Huai'an, Tone 1 is a high falling tone and Tone 4 is a high level tone. While in Standard Mandarin, Tone 1 is a high level tone (tone letter: 55) and Tone 4 is a falling tone (tone letter: 51). Tone 2s and Tone 3s in Huai'an and Standard Mandarin are largely similar.

[9] We reiterate the usage of 'underlying' and 'derived' in this paper: Consistent with other places in this paper, 'underlying Tone 3' means surface Tone 3 that maps from underlying Tone 3, and 'derived Tone 3' means surface Tone 3 that maps from underlying Tone 1 or Tone 4.

explained by the Obligatory Contour Principle (Leben, 1973; McCarthy, 1986; Yip, 2002; inter alia). However, some researchers reject the use of the Obligatory Contour Principle as the motivation of tone sandhi processes in Mandarin languages (Duanmu, 1994, 2007; inter alia). We will not address this debate since it is tangential to the main argument of this paper.[10]

(2) Relevant tone sandhi rules in Huai'an Mandarin (Y. Wang & Kang, 2012)

a. Low-register tone sandhi[11]
   i. Tone 3 sandhi: T3 + T3 → T2 + T3

b. High-register tone sandhi [optional processes]
   i. Tone 1 sandhi: T1 + T1 → T3 + T1
   ii. Tone 4 sandhi: T4 + T4 → T3 + T4

Crucially, the Tone 3 output of the high-register tone sandhi processes feeds the low-register Tone 3 sandhi process as in (3). Since high-register tone sandhis are optional and Tone 3 sandhi is also optional for trisyllabic utterances in (3), multiple surface representations are possible for both examples.

(3) Feeding Order in Huai’an Mandarin (boldface represents the locus of a potential tonal change due to the relevant tone sandhi process; the data is from the authors)

a. Tone 1 sandhi feeds Tone 3 sandhi

u ku fən
Mr. Wu estimate score
'Mr. Wu estimates scores.'

UR	T3 T1 T1		
Tone 1 sandhi	T3 **T3** T1		(or) T3 **T1** T1
Tone 3 sandhi	**T2** T3 T1	(or) **T3** T3 T1	**T3** T1 T1
SR	T2 T3 T1	(or) T3 T3 T1	(or) T3 T1 T1

[10] Depending on different proposed representations of the tones (M. Y. Chen, 2000; Duanmu, 2007; Yip, 2002; to name a few), the motivation of tone sandhis can be different. We will discuss the assumed tonal representation for this paper in Section 6.1 that a Mandarin tone is a single phonological unit.

[11] Register is a tonal feature first proposed by Moira Yip (1980) and then widely adopted in the literature of Chinese tonal phonology. Here we simply use the feature to distinguish Tone 3 sandhi from other tone sandhi processes in Huai’an.

b. Tone 4 sandhi feeds Tone 3 sandhi

u	to	ʐəɯ
Mr. Wu	chop	meat

'Mr. Wu chops meat.'

UR	T3 T4 T4		
Tone 4 sandhi	T3 **T3** T4		(or) T3 **T4** T4
Tone 3 sandhi	**T2** T3 T4	(or) **T3** T3 T4	**T3** T4 T4
SR	T2 T3 T4	(or) T3 T3 T4	(or) T3 T4 T4

The feeding relationships between each of the high-register tone sandhis and Tone 3 sandhi suggest that the high-register tone sandhis result in a Tone 3 category that is phonologically the same as an underlying Tone 3. This interpretation of the data remains the same given a parallel approach to phonology such as Optimality Theory (Prince & Smolensky, 1993). A usually employed markedness constraint for low-register tone sandhi in Mandarin languages is '*Tone3Tone3', which is based on the Obligatory Contour Principle. This constraint prevents adjacent Tone 3 syllables (C.-Y. Wang & Lin, 2011; Zhang, 1997; see also M. Y. Chen, 2000 for using this constraint in an implicit fashion). For this constraint to trigger the structural change in the first syllable (namely, Tone 3 → Tone 2), the second syllable in /Tone 3 Tone 4 Tone 4/ or /Tone 3 Tone 1 Tone 1/ must surface as Tone 3. Consequently, an Optimality Theory analysis would also maintain the crucial categorical aspects of the feeding order that are focal for the current paper.

With regard to this interpretation of phonological identity between derived and underlying Tone 3s, concerns may be raised about application rates, especially when an underlying Tone 3 mandatorily triggers Tone 3 sandhi while a derived Tone 3 can only optionally trigger Tone 3 sandhi when the two types of Tone 3 syllables are the middle syllable of a trisyllabic utterance. The comparison is shown in (3) and (4). And some researchers may want to ascribe the difference in application rates to a difference between derived Tone 3 and underlying Tone 3 in the phonology, either as different phonological representations or as the same representations that are indexed to different variable processes. However, intervening factors are not controlled for when compared in this way. For Tone 3 sandhi to mandatorily apply before an underlying Tone 3, the established planning window only needs to include the two Tone 3 syllables. Therefore, in (4), the established planning window only needs to include the first two syllables. In contrast, for Tone 3 sandhi to apply before a derived Tone 3, the established planning window needs to include at least three syllables to ensure both high-register Tone 1/Tone 4 sandhi and low-register Tone 3 sandhi occur. It is well recognized in the previous literature that a larger planning window has more planning difficulty and is therefore less likely (Ferreira & Swets, 2002; Kilbourn-Ceron & Goldrick, 2021; Wagner et al., 2010; inter alia). The reason is the increasing burden on working memory, which can lead to speech errors or delays. Huai'an turns out to not be an exception. Previous experimental study on Tone 3 sandhi in Huai'an does support the existence of the effect of planning difficulty (Du & Lin, 2021). Due to such an effect, a planning window that extends three syllables long is less likely to be established in (3), which means Tone 3 sandhi is less likely to apply before a derived Tone 3. Overall, the difference in application rates comes naturally from the planning difficulty effect and does not need to be accounted for in the phonology.[12]

[12] Note, it has been noted in previous literature that tone sandhi patterns in many Mandarin languages are sensitive to prosodic structures (M. Y. Chen, 2000; Duanmu, 2007; C.-Y. Wang & Lin, 2011; Zhang, 1997; inter alia).

As pointed out by two anonymous reviewers, proponents of gradient phonological representation may argue that although both underlying and derived Tone 3s can trigger Tone 3 sandhi, they may still have different phonological representations. By this analysis, the difference in application rates would be explained by the difference in the phonological representations. First, we would like to point out that any analysis that predicts application rates based on gradient phonological representations or phonetic similarity would have to be precise in accounting for not only cases where the process is triggered but also cases where the process is *not* triggered; namely, it would have to explain why only the derived Tone 3 shows a variation in application rates and not the underlying Tone 3, and not the other way around. Furthermore, it would have to account for the fact that any other tones that are phonetically similar (along the relevant dimensions) do not trigger the process. While an evaluation of such an analysis is not possible without a concrete specification of the proposal, we suspect that, to explain the difference between derived Tone 3 and underlying Tone 3, one will have to make reference to performance factors anyway. Relatedly, we appeal to the need to prioritize relatively simple categorical phonological representations when they are sufficient to account for the observed patterns (Occam's razor/law of parsimony); in our case, the difference in application rates can be accounted for by independently needed performance factors, namely planning, and therefore we need not complicate our understanding of the relevant phonological (tonal) representations. For this reason, we see the feeding rule interaction as evidence of complete phonological neutralization of the derived Tone 3 from Tone 1 and Tone 4 sandhi processes. Furthermore, we use the processes to probe the phonetic (acoustic) consequences of the neutralizing processes in the case of the derived Tone 3 that in turn trigger Tone 3 sandhi.

(4) Application of Tone 3 sandhi before underlying Tone 3 in trisyllabic utterances

a. Tone 1 sandhi feeds Tone 3 sandhi

u	pɔ	tɕi
Mr. Wu	protect	car

'Mr. Wu protects cars.'

UR	T3 T3 T1
Tone 3 sandhi	**T2** T3 T1
SR	T2 T3 T1

b. Tone 4 sandhi feeds Tone 3 sandhi

u	tɛ	ɕiæ̃
Mr. Wu	catch	elephant

'Mr. Wu catches elephants.'

UR	T3 T3 T4
Tone 3 sandhi	**T2** T3 T4
SR	T2 T3 T4

To further ensure the phonological equality of derived Tone 3 and underlying Tone 3, we only analyze the derived Tone 3 tokens that actually trigger Tone 3 sandhi in this paper, which allows

Therefore, the observed difference in application rates between derived and underlying Tone 3s can also be cashed out in terms of planning difficulty related to prosodic phrases.

us to have perfect surface minimal pairs in each of our experiments. By doing so, we also exclude the possibility that any identified incomplete phonetic neutralization patterns arise as a result of averaging the outcomes of an optional phonological process, since we only look at the cases where we have reason to believe that the process applied. Despite the categorical phonological behavior of the derived Tone 3 in Huai’an, in the next two sections, we will show that there is substantial incomplete phonetic neutralization of derived Tone 3 and underlying Tone 3 for the feeding orders involving Tone 1 sandhi and Tone 4 sandhi.

4. Experiment 1: Tone 1 sandhi

4.1 Participants

We recruited 11 native speakers of Huai'an Mandarin via personal relationships in Huai’an City. The age range was from 37 to 55 years. Among them, 8 self-identified as female, and 3 as male. Due to the language standardization trend in mainland China (Ramsey, 1989), young speakers in Huai'an are generally bilingual and are native speakers of both Huai'an and Standard Mandarin. To minimize the influence of Standard Mandarin, we recruited older speakers who are only fluent in Huai'an in this study. All the participants were born and raised in Huai'an City. None of them had participated in any linguistic studies before or heard about the concept of incomplete neutralization.

4.2 Stimuli

The stimuli were composed of trisyllabic sentences with each syllable forming a separate word, therefore it is ensured that the tone sandhi processes are post-lexical and completely productive. Also, only right-branching utterances as in (3) are employed simply because not enough left-branching utterances could be constructed by us given the paradigm to be introduced immediately. The stimuli were divided into four sets as shown in (5). Furthermore, the third syllable was always Tone 1. The second syllable was one of the following possibilities: a) an underlying Tone 1 that optionally underwent Tone 1 sandhi to become Tone 3, b) an underlying Tone 3 that did not undergo any tone sandhi in this context. The first syllable was underlyingly a Tone 3 or a Tone 2. As a consequence of the possibilities in the second syllable, there were a few different possibilities for the first syllable, including: a) an underlying Tone 3 that could undergo Tone 3 sandhi to become Tone 2 with reference to the second syllable, b) an underlying Tone 2 that did not undergo any tone sandhi in this context. The four sets were only different in tonal patterns but not in segmental content. Furthermore, the crucial second syllable was always a voiceless unaspirated stop plus vowel sequence. Voiceless unaspirated stops were chosen to make sure that there is a consistent way to annotate the acoustic onset of the vowel by referring to the burst of the stop. The full stimulus list is summarized in Appendix 1. It is worth noting that one character [搭] may be pronounced with the only checked tone in Huai'an (Jiao, 2004; Y. Wang & Kang, 2012), which is an allophone of Tone 4 and appears only on monomoraic syllables ending with glottal stop. We excluded all checked tone productions when extracting f0 information.

(5) Four sets of stimuli in Experiment 1 [the syllables crucial for the current comparison are underlined and boldface]

a) underlying T3 following underlying T2: /T2 T3 T1/ → [T2 T3 T1]
b) underlying T3 following underlying T3: /T3 T3 T1/ → [T2 **<u>T3</u>** T1]
c) derived T3 following underlying T2: /T2 T1 T1/ → [T2 T3 T1] or [T2 T1 T1]
d) derived T3 following underlying T3: /T3 T1 T1/ → [T2 **<u>T3</u>** T1] or [T3 T1 T1]

Out of the above set of possibilities, the most crucial comparison is between two tones in the second syllable, namely, an underlying Tone 3 as in (b) and a derived Tone 3 as in the first possibility in (d). This particular comparison controls for the preceding surface context (derived Tone 2) and the following surface context (underlying Tone 1) and is therefore a perfect minimal pair. Furthermore, the two cases also show evidence that both tones are in fact categorically Tone 3 as they trigger Tone 3 sandhi on the preceding tone. Finally, as mentioned previously, the comparison allows us to exclude the possibility that any identified incomplete phonetic neutralization pattern arises as a result of averaging the outcomes of an optional phonological process. This is the crucial pair we will focus on in this experiment.

The set of possibilities also allows us to visually compare the derived Tone 3 against an underlying Tone 1 in the same surface context, as in the second possibility in (c) [albeit, the preceding syllable in this case is an underlying Tone 2 instead of a derived Tone 2].

Each participant produced 4 repetitions of 24 test and 27 filler sentences at a natural speech rate, which means each participant read a total of 204 sentences. All stimuli were randomized for each participant.

4.3 Procedure

The experiment was conducted entirely in Huai'an city. Each participant was recorded by a trained research assistant using Audacity (Audacity Team, 2019) and a Popu Line BK USB microphone on a Lenovo laptop in a quiet room that was either located in the participants' home or workplace. The participants were told that the purpose of the study was to collect some general information on Huai'an. None of the participants reported noticing the minimal pairs or the real purpose of the study being on tones in the post-experimental interview. The participants were instructed to read at a normal speech rate using their everyday voice, and the stimuli were presented in Chinese characters. The participants were also encouraged to read through the stimulus list to be familiar with the reading materials before producing them.

4.4 Measurement

Using Praat (Boersma & Weenink, 2021), the recordings were manually annotated by the first author, who is a native speaker of Huai'an. An example is shown in Figure *2*. Only the second syllable was marked and the annotation file had 6 tiers in total. The first tier marked the vowel of the second syllable for phonetic analysis. The first zero crossing at the beginning of the voicing of the target vowel and after the burst of the unaspirated stop was identified as the vowel onset, and the zero-crossing immediately following the vowel's final glottal pulse was identified as the vowel offset. All other tiers marked the whole second syllable to index phonological information and recording quality. The onset of the second syllable was marked just before the release burst of the initial stop and the offset of the second syllable corresponded with the offset of the nuclear vowel.

The second tier indicated the whole sentence in Pinyin, which is the official Romanization system for Chinese characters in China. The third tier was the tone sandhi condition where 'yes' meant the second syllable had undergone tone sandhi and 'no' meant it had not, the fourth and fifth tiers had the underlying tones and surface tones information respectively and the last tier had the quality of the recording. We only used productions that were marked 'good' in the analysis. The reasons that productions were not marked as 'good' included background noise, speech errors, any long delay while producing the utterance, and checked tone pronunciation. f0 was extracted only from the vowel at 5% steps with a script in Praat.

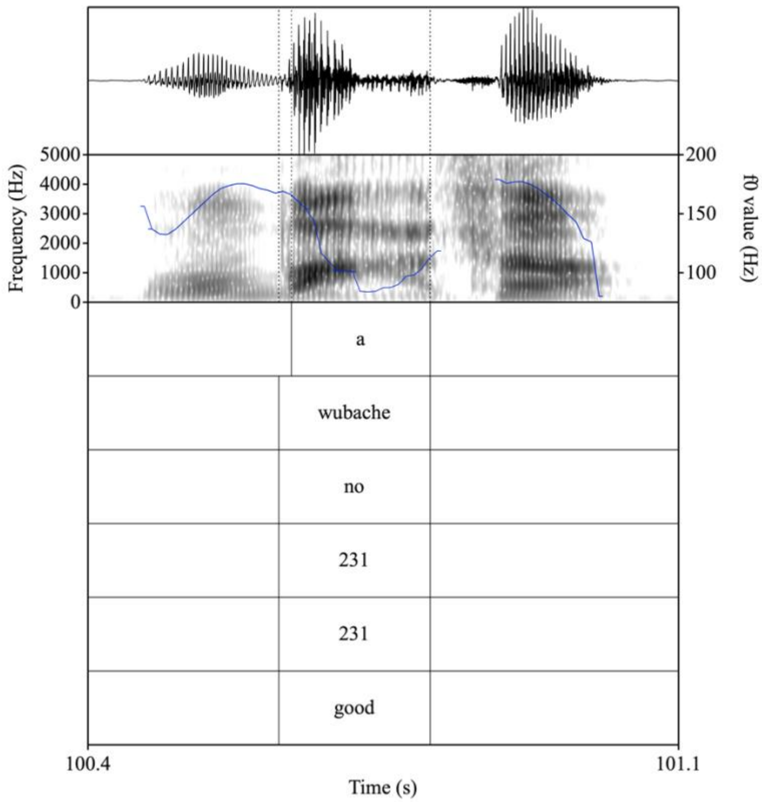


Figure 2: Annotation scheme of Experiment 1 (Tone 1)

To compare across different speakers and different vowels, z-score transformation was performed for each vowel of each speaker based on Hz scale (Laplace, 1820; Lobanov, 1971).

4.5 Results and statistical modelling

All data analyses in this paper were performed in R (R Core Team, 2021) using the **tidyverse** suite of packages (Wickham et al., 2019). The statistical modelling was done using the **lme4** package (Bates et al., 2021).[13]

The number of tokens for each possible combination of Underlying Representation and Surface Representation is summarized in Table *2*. The application rate of Tone 3 sandhi before underlying

[13] All the data presented in this article are available both in the form of csv files with the extracted measurements at the permanent link https://osf.io/9qcbr. The repository also includes the R script used to analyze and plot the data.

Tone 3 is 97.2% while the application rate before derived Tone 3 is 74.0%.[14] 71 tokens were not marked as 'good' and excluded, which accounts for 6.7% of all test stimuli.

UR	SR	Number of tokens
T2T3T1	T2T3T1	259
T3T3T1	T3T3T1	7
T3T3T1	T2T3T1	242
T2T1T1	T2T1T1	74
T2T1T1	T2T3T1	167
T3T1T1	T3T1T1	59
T3T1T1	T3T3T1	46
T3T1T1	T2T3T1	131

Table 2: Number of Tokens for UR and SR combination

The z-score transformed f0 contours on the crucial second syllable are shown in Figure *3*. As a reminder, the crucial comparison is between a derived Tone 3 and an underlying Tone 3 after derived Tone 2s in the same surface context --- the context establishes that both the Tone 3s are categorically Tone 3 as they trigger Tone 3 sandhi. We also present the tone contour for an underlying Tone 1 in the same surface context for visual comparison with the two crucial Tone 3s.

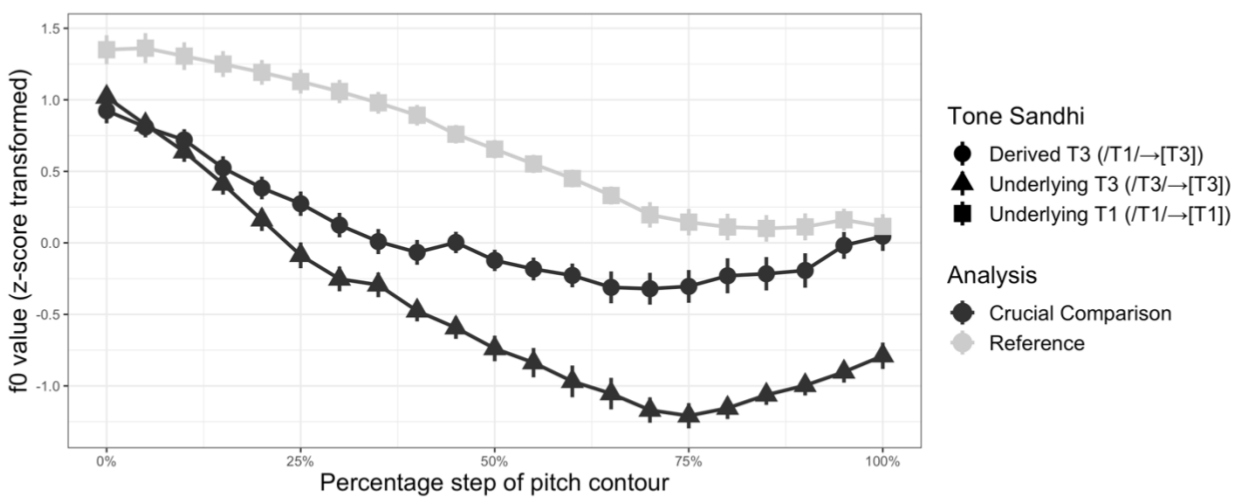


Figure 3: Contours comparison of the second syllable in Experiment 1 (Tone 1)
(Error bars indicate standard error)

Based on the visual inspection of the data, the derived Tone 3 seems to start as an underlying Tone 3 and ends as an underlying Tone 1. And the contour shape of the derived Tone 3 is close to that of an underlying Tone 3. Furthermore, the comparison between underlying Tone 3 and derived Tone 3 clearly shows that the neutralization is incomplete.[15]

---

[14]As mentioned before, the difference in application rates does not necessarily inform us of any differences in phonological representations since the difference is accountable through independently needed mechanisms, namely the difficulty in planning longer utterances.

[15] To further address the concern that incomplete neutralization patterns identified in this study may arise as a result of averaging the outcomes of an optional phonological process, the distributions of underlying Tone 1, derived Tone

For the purposes of statistical modelling, we used just the two-group factor (underlying Tone 3 vs. derived Tone 3), and ignored underlying Tone 1, in order to simplify the modelling and address only the crucial question of whether or not the underlying and derived Tone 3s have incompletely neutralized. The results turn out to support the observation that the neutralization is indeed incomplete phonetically.

In dealing with time course data, traditional techniques like t-tests and ANOVA have to divide continuous time into multiple time bins and therefore have to make multiple comparisons. This method has been argued by Mirman (2017) to be problematic for increasing the risk of 'false positives'. Since each time bin incurs the nominal 5% false positive rate implied by '$p < 0.05$', overall, the false positive rate with multiple time bins and multiple comparisons will be much higher than a single comparison.

To solve this problem, multiple analysis methods have been developed, which includes Smooth Spline Analysis of Variance (SS-ANOVA) (Y. Wang, 1998), Generalized Additive Model (GAM) (Hastie & Tibshirani, 1990) and Growth Curve Analysis (GCA) (Mirman et al., 2008a; Mirman, 2017). In this paper, we follow S. Chen et al. (2017), in modelling f0 contours using Growth Curve Analysis. Growth Curve Analysis uses multilevel linear regression to avoid multiple comparisons and has been argued to be a useful modelling technique in different fields (Baldwin & Hoffmann, 2002; McArdle & Nesselroade, 2003; inter alia). To apply Growth Curve Analysis in Huai'an tones, we started with a simple model as in (6) (Mirman et al., 2008).

(6) $$Y_{ij} = (\gamma_{00} + \zeta_{0i}) + (\gamma_{10} + \zeta_{1i}) * Time_{ij} + \varepsilon_{ij}$$

Here i is the i^{th} f0 (z-score transformed) contour and j is the j^{th} time point, and Y_{ij} is the f0 (z-score transformed) value for i^{th} contour at j^{th} time point. γ_{00} is the population average value for the intercept, ζ_{0i} is individual variation on the intercept, γ_{10} is the population average value for the fixed effect of time, ζ_{1i} is individual variation on the fixed effect of time and ε_{ij} is the error term.[16] To optimize the model for the data, we employed higher-order polynomial functions, and allowed individuals to vary on each term only when those terms reached significance according to chi-square likelihood ratio tests (S. Chen et al., 2017; S. Chen & Li, 2021; inter alia). In Mandarin languages, a Tone Bearing Unit (TBU), which is assumed to be the syllable or the rhyme or the nucleus of the rhyme, has been widely argued to be associated with at most three tonal targets (Bao, 1990, 1992; Duanmu, 1994; inter alia). Therefore, the most complex tones can only have one change of direction, which will produce U-shaped contours, such as high-low-high and low-high-low. To conform to this observation, we only considered up to second-order functions to ensure that the final model is not more complex than a U-shape contour. Also, orthogonal polynomials were used to make sure that the linear and quadratic terms were not correlated (Mirman, 2017). After optimizing the model by including all significant terms, we first treated underlying Tone 3 and derived Tone 3 as the same and modelled them as one single contour to get Model 1. Then we built models that treat them as different, namely, models that include a tone sandhi condition (underlying Tone 3 vs. derived Tone 3) to do model comparison. Based on Model 1, tone sandhi condition is first allowed to affect only intercept to get Model 2. Then tone sandhi condition is

3, and underlying Tone 3 are shown for each time step in the supplementary materials. Crucially, the derived Tone 3 distribution is generally uni-modal, and distinct from the other two distributions, across the time-steps. Thus, there is no evidence of an averaging artifact over optional surface representations for the derived Tone 3 case.

[16] Note, the individual variation terms, $\zeta 0i$ and $\zeta 1i$, are akin to the random intercept by participant, and random slope of time by participant.

allowed to affect both intercept and linear term to get Model 3. Finally, tone sandhi condition is allowed to affect all fixed effects, which include intercept, linear term and quadratic term, and the outcome is Model 4. Chi-square likelihood ratio test was used to determine whether two minimally different models differ significantly.

The result shows that the difference between underlying Tone 3 and derived Tone 3 is in fact supported by model comparisons. The addition of a tone sandhi condition improves the model on the intercept as shown by comparing Model 1 and Model 2 ($\chi^2(1)$=331.81, p<0.01), on the linear term as shown by comparing Model 2 and Model 3 ($\chi^2(1)$= 118.34, p<0.01) and on the quadratic term as shown by comparing Model 3 and Model 4 ($\chi^2(1)$= 14.99, p<0.01). Figure *4* shows how the full model (Model 4) fits the observed data. The parameter estimates for the full model are summarized in Table *3*.

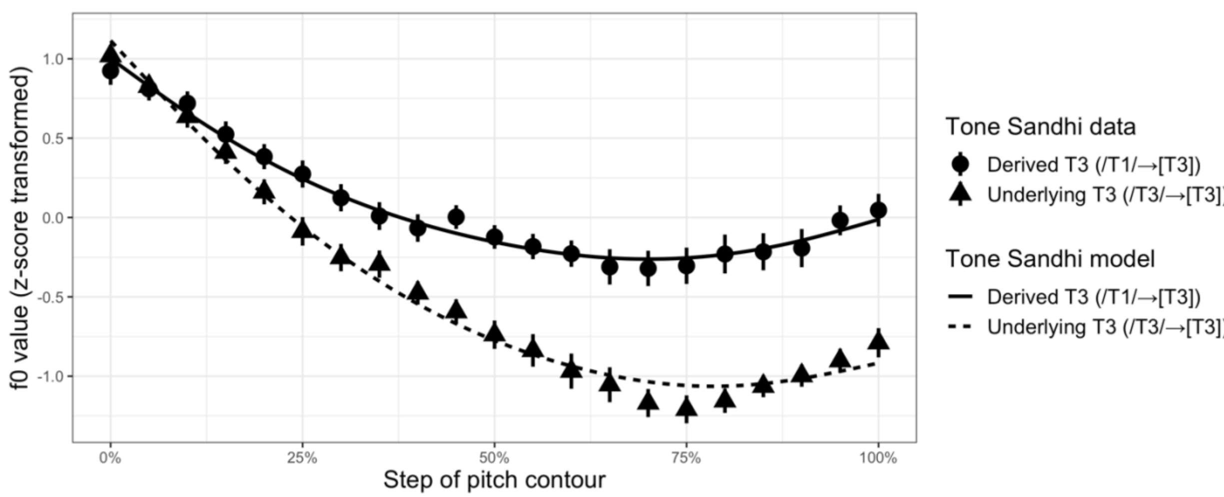


Figure 4: Observed data and Growth Curve Model fits for derived and underlying Tone 3 (Error bars indicate standard error)

	Estimate	Std. Error	t	p
Intercept	0.07	0.06	1.13	0.28
Linear	-16.01	2.14	-7.47	<0.01
Quadratic	9.57	1.59	6.04	<0.01
Tone Sandhi: Intercept	-0.50	0.03	-19.68	<0.01
Tone Sandhi: Linear	-13.64	1.24	-10.99	<0.01
Tone Sandhi: Quadratic	4.80	1.23	3.90	<0.01

Table 3: Parameter estimates of the full model with the assumption of tone sandhi affecting every fixed effect (baseline: derived Tone 3)

Moreover, the effect size of incomplete neutralization is large in Tone 1 sandhi. The mean difference in f0 between underlying Tone 3 and derived Tone 3 across all steps is 18 Hz, which is more than 2 times the Just Noticeable Difference of f0 value (7 Hz) for Mandarin speakers (Jongman et al., 2017). Furthermore, across the last 10 steps (step 11 to step 20), the f0 difference is over 22 Hz, which is more than 3 times the Just Noticeable Difference. The f0 difference (f0 of derived Tone 3 - f0 of underlying Tone 3) of each step is summarized in Table *4*. Recall that the underlying premise from those who criticize the small effect size of incomplete neutralization is that only if the differences were robust and large in size, the existence of such an effect should be accepted as functionally relevant.[17] According to that standard, the case of Huai'an Tone 1 sandhi is clearly a case of phonetically incomplete neutralization.

Step	f0 difference (Hz)	Step	f0 difference (Hz)
0	-5	11	22
1	-2	12	25
2	2	13	26
3	4	14	29
4	7	15	30
5	12	16	32
6	12	17	30
7	11	18	28
8	15	19	29
9	22	20	25
10	21		

Table 4: f0 difference of each step in Experiment 1 (Tone 1)

To show that the pattern is not unique to Tone 1 sandhi and to extend the scope of the current study, we ran a second experiment on Tone 4 sandhi process in Huai'an.

5. Experiment 2: Tone 4 sandhi

5.1 Participants

We recruited 20 native speakers of Huai'an Mandarin also via personal relationships in Huai'an City. The age range was from 33 to 57 years old. Again, to minimize the influence of Standard Mandarin, we avoided younger speakers in this study. Among them, 16 self-identified as female, and 4 as male. 5 speakers had also participated in Experiment 1. The interval between the two experiments was about 7 months; the 5 participants from Experiment 1 failed to guess and were not told the purpose of Experiment 2. Like Experiment 1, all the participants were born and raised in Huai'an City. Other speakers have not participated in any linguistic studies before or heard about the concept of incomplete neutralization.

5.2 Stimuli

The stimuli were organized in the same way as in Experiment 1. The four sets of trisyllabic sentences are shown in (7), and the full stimulus list is summarized in Appendix 2.

[17] We acknowledge that it is not entirely clear to us what is intended by the use of phrases such as "functional relevance", since many aspects of linguistic behavior might be important to the speaker/listener while not stemming from the grammar per se; however, we retain the phrase here to reflect the terminology in the sub-field.

(7) Four sets of stimuli in Experiment 2 [the syllables crucial for the current comparison are underlined and boldface]

a) underlying T3 following underlying T2: /T2 T3 T4/ → [T2 T3 T4]
b) underlying T3 following underlying T3: /T3 T3 T4/ → [T2 **<u>T3</u>** T4]
c) derived T3 following underlying T2: /T2 T4 T4/ → [T2 T3 T4] or [T2 T4 T4]
d) derived T3 following underlying T3: /T3 T4 T4/ → [T2 **<u>T3</u>** T4] or [T3 T4 T4]

As with Experiment 1, the crucial comparison is between two tones in the second syllable, namely the underlying Tone 3 in (b) and the derived Tone 3 as in the first possibility in (d). This comparison allows us to control for the surface context, while also establishing that the two tones are indeed categorical Tone 3s since they trigger Tone 3 sandhi on the preceding tone. Furthermore, as mentioned previously, the comparison allows us to exclude the possibility that any identified incomplete phonetic neutralization pattern arises as a result of averaging the outcomes of an optional phonological process.

The set of possibilities also allows us to look at an underlying Tone 4 in roughly the same surface context, as in the second possibility in (c), for visual comparison.

Each participant produced 4 repetitions of 20 test sentences at a natural speech rate with 20 fillers, which means that each participant read a total of 160 sentences.

5.3 Procedure

The procedure was identical to that of Experiment 1.

5.4 Measurement

The recordings were manually annotated by the first author but with a somewhat different scheme. For this experiment, both the first and second syllables were marked. The first syllable was marked to confirm that derived Tone 3 can in fact trigger Tone 3 sandhi on this syllable. An example is shown in Figure *5*. The annotation file had five tiers in total. The criteria for marking vowels and syllables remained the same. The first tier marked the vowel of the syllable. All other tiers marked the whole second syllable to index phonological information and recording quality. The second tier indicated the position of the syllable inside the sentences where a first syllable was marked '1' and a second syllable was marked '2', the third tier contained the Pinyin of the whole sentence followed by the underlying tone of the syllable. The fourth tier marked whether the syllable underwent tone sandhi. And the last tier indicated the quality of the recording. Similar to the previous experiment, we only used productions of recordings that were marked 'good'. The f0 extraction, normalization and visualization processes are identical to those in the previous experiment.

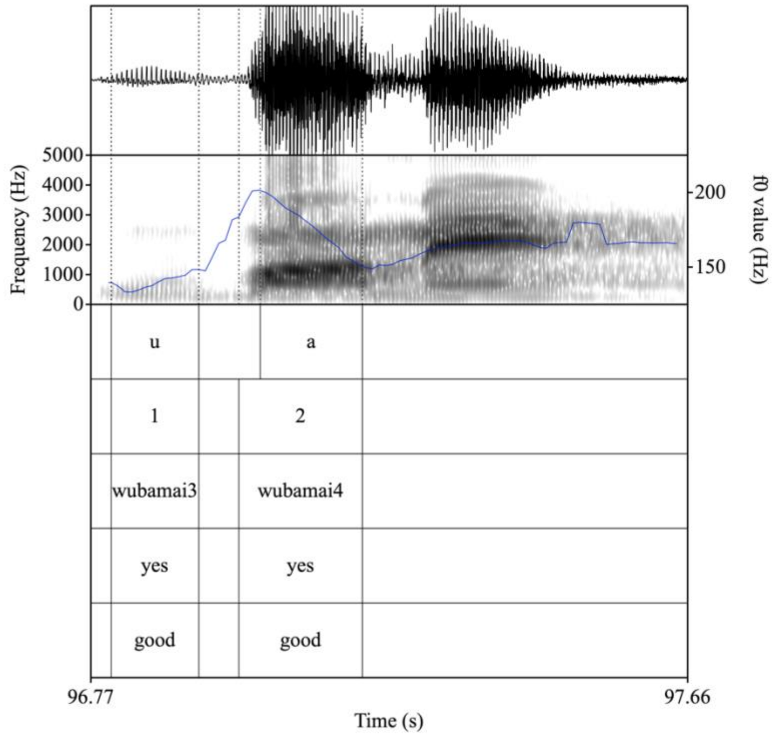


Figure 5: Annotation scheme of Experiment 2 (Tone 4)

5.5 Results and statistical modelling

The number of tokens for each possible combination of Underlying Representation and Surface Representation is summarized in Table *5*. The application rate of Tone 3 sandhi before underlying Tone 3 is 94.8% while the application rate before derived Tone 3 is 24.2%.[18] 79 tokens were not marked as 'good' and excluded, which accounts for 5.0% of all test stimuli.

[18] Again, we point out that the difference in application rates does not necessarily inform us of any differences in phonological representations. The difference can be accounted for by the difficulty in planning longer utterances. We also recognize that the application rate of derived Tone 3 from Tone 4 here is different from derived Tone 3 from Tone 1 in Experiment 1 (74.0%), and as one anonymous reviewer pointed out, such a difference may be used to argue for a phonological difference between the Tone 3s from different underlying sources. However, such a difference may simply be due to different groups of speakers in two independent experiments or even more simply that the observed difference in effect sizes is simply random variation, as would be expected between any two experiments measuring the same phenomenon. Future study is needed to compare Huai'an Tone 1 sandhi and Tone 4 sandhi on the same group of speakers in one single experiment. If the application rates are indeed replicable, an intriguing possibility that we note for the readers is that the phonetic difference between derived Tone 3 from Tone 1 and derived Tone 3 from Tone 4 may itself serve as a performance factor (related to the differential difficulty in implementing different tones that in turn effects planning) that can account for the difference of application rate outside phonology. At the moment, the explanations based on planning difficulty are simply speculative since it is not clear how phonological planning can affect tone sandhi application rate, and future study is needed to quantify the size of variation caused by planning difficulty.

UR	SR	Number of tokens
T2T3T4	T2T3T4	386
T3T3T4	T3T3T4	20
T3T3T4	T2T3T4	368
T2T4T4	T2T4T4	156
T2T4T4	T2T3T4	212
T3T4T4	T3T4T4	98
T3T4T4	T3T3T4	213
T3T4T4	T2T3T4	68

Table 5: Number of Tokens for UR and SR combination

The z-score transformed f0 contours on the crucial second syllable are shown in Figure *6*. Again, the crucial comparison is between a derived Tone 3 and an underlying Tone 3 after a derived Tone 2 in the same surface context. We also present the tone contour for an underlying Tone 4 in the same surface context for visual comparison with the two crucial Tone 3s.

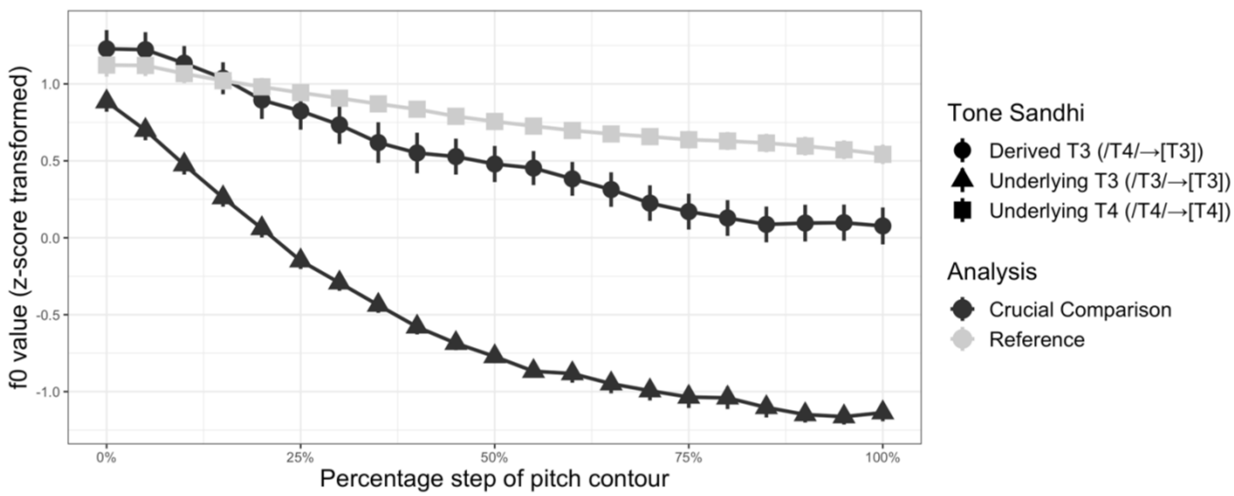


Figure 6: Contours comparison of the second syllable in Experiment 2 (Tone 4)
(Error bars indicate standard error)

Based on visual inspection of the data, the pattern seems to be different from the case of Tone 1 sandhi. The derived Tone 3 seems to start as an underlying Tone 4,[19] instead of as an underlying Tone 3 as in Experiment 1. Furthermore, the derived Tone 3 gradually deviates from underlying Tone 4 through the whole contour; note, this is in contrast to Experiment 1, where the derived Tone 3 ended up at a value almost identical to the underlying Tone 1. However, the contour shape of the derived Tone 3 is again close to that of an underlying Tone 3 as in Experiment 1. Despite

[19] As observed by one anonymous reviewer, more accurately, derived Tone 3 from Tone 4 actually starts higher than underlying Tone 4 as shown in Figure *6*, although the difference in raw pitch between derived Tone 3 and underlying Tone 4 is very small (less than 5 Hz for the first 4 steps). As suggested by this anonymous reviewer, this pattern may be due to some effort to maintain contrast between derived Tone 3 and underlying Tone 4 since the remainder of derived Tone 3 is relatively high and close to underlying Tone 4.

the difference, incomplete phonetic neutralization is again clearly observed in the comparison between underlying Tone 3 and derived Tone 3.[20,21]

The modelling method remains the same as in Experiment 1, and four models are generated. The observation of incomplete phonetic neutralization is again supported by model comparisons. The addition of a tone sandhi condition improves the model on the intercept as shown by comparing Model 1 and Model 2 ($x^2(1)$=1429.23, p<0.01), the linear term as shown by comparing Model 2 and Model 3 ($x^2(1)$= 66.22, p<0.01) and the quadratic term as shown by comparing Model 3 and Model 4 ($x^2(1)$= 32.67, p<0.01). Figure *7* shows how the full model with the assumption of tone sandhi affecting every fixed effect fits the observed data. And the parameter estimates for full model are summarized in Table *6*.

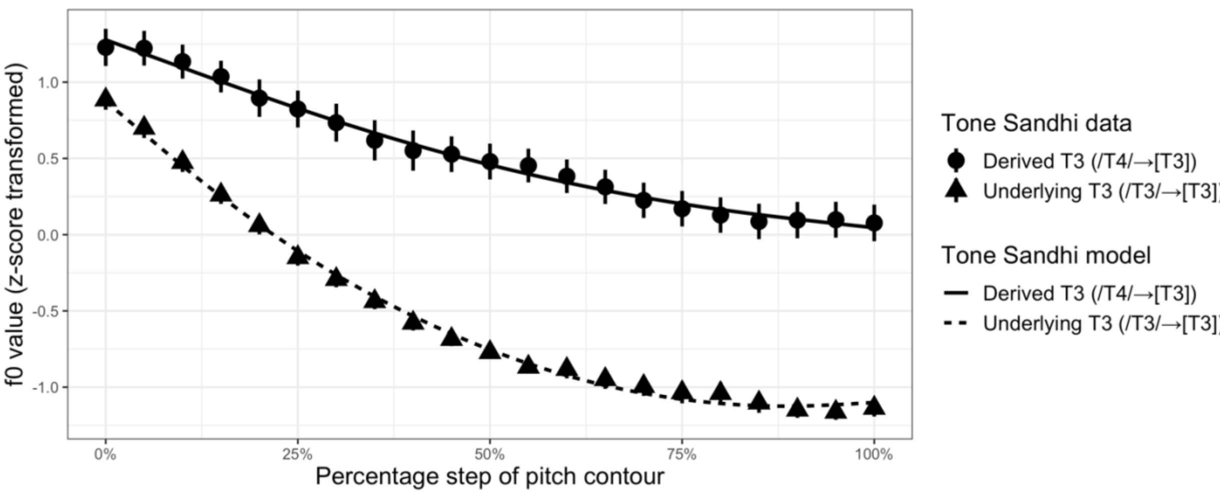


Figure 7: Observed data and Growth Curve Model fits for derived and underlying Tone 3 (Error bars indicate standard error)

	Estimate	Std. Error	t	p
Intercept	0.50	0.06	8.26	<0.01
Linear	-22.09	2.23	-9.89	<0.01
Quadratic	4.38	1.33	3.29	<0.01
Tone Sandhi: Intercept	-1.01	0.02	-43.58	<0.01
Tone Sandhi: Linear	-10.43	1.26	-8.30	<0.01
Tone Sandhi: Quadratic	7.17	1.25	5.75	<0.01

Table 6: Parameter estimates of the full model with the assumption of tone sandhi affecting every fixed effect (baseline: derived Tone 3)

[20] We recognize that the underlying Tone 3 in Experiment 2 appears to be slightly phonetically different from its counterpart in Experiment 1, especially at the tonal offset position. Tone 3, which is usually used to represent low tone in Mandarin languages, generally involves creakiness. The small difference between underlying Tone 3s in Experiments 1 and 2 may be caused by inconsistency related to the Praat f0 estimation algorithms for creaky sounds. Whether it is due to this issue or due to simple random variation is something that we leave for further inquiry.

[21] Again, to further address the concern that incomplete neutralization patterns identified in this study may arise as a result of averaging the outcomes of an optional phonological process, the distributions of underlying Tone 4, derived Tone 3, and derived Tone 3 are shown for each time step in the supplementary materials. Again, crucially, the derived Tone 3 distribution is generally uni-modal, and distinct from the other two distributions, across the time-steps. Thus, there is no evidence of an averaging artifact over optional surface representations for the derived Tone 3 case.

Again, the effect size of incomplete neutralization is also large in Tone 4 sandhi. The mean difference in f0 between underlying Tone 3 and derived Tone 3 across all steps is 17 Hz, which is more than 2 times the Just Noticeable Difference of f0 value (7 Hz) for Mandarin speakers (Jongman et al., 2017). Also, across the last 11 steps (step 9 to step 20), the f0 difference is over 21 Hz, which is more than 3 times the Just Noticeable Difference. The f0 difference (f0 of derived Tone 3 - f0 of underlying Tone 3) of each step is summarized in Table *7*. Therefore, the case of Huai'an Tone 4 sandhi should also be safely defined as phonetically incomplete neutralization, and not susceptible to the criticism of a small effect size.

Step	f0 difference (Hz)	Step	f0 difference (Hz)
0	-4	11	24
1	-1	12	23
2	1	13	24
3	3	14	23
4	5	15	23
5	11	16	23
6	14	17	24
7	16	18	26
8	19	19	27
9	22	20	27
10	22		

Table 7: f0 Difference of each step in Experiment 2 (Tone 4)

The coding in Experiment 2 also allowed us to answer another question that we did not answer for Experiment 1. In Experiment 1, we impressionistically coded whether or not the first syllable was in fact subject to Tone 3 sandhi. One could have argued that this impressionistic coding could have been inaccurate, and was based on a perceptual bias of the annotator (first author). To address this concern, it would have been optimal if we could have shown through phonological behavior that the derived Tone 2 is indeed phonologically identical to underlying Tone 2. Although historically Tone 2 sandhi (Tone 2 + Tone 2 → Tone 3 + Tone 2) existed in Huai'an (Y. Wang & Kang, 2012), this tone sandhi rule was not observed in our fieldwork in early 2020 probably due to the influence of the standard language, as is generally observed in other languages (Labov, 1963; Milroy, 2001; inter alia). And no researchers before have tested if derived Tone 2 can trigger another tone sandhi process. Therefore, we cannot verify if the derived Tone 2 can trigger Tone 2 sandhi like an underlying Tone 2. Furthermore, we are not aware of any other phonological processes in the language that are triggered by Tone 2. As a result, it is not possible to establish Tone 2 category by phonological behavior in Huai'an and we turn to provide phonetic evidence for the Tone 2 identity of the derived rising tone.

To make some inroads into the question of the phonological nature of the (putatively) derived Tone 2 in initial position, in Experiment 2, we also annotated the first syllable, and are therefore able to observe the f0 contours for derived Tone 2 (from underlying Tone 3) and compare it to an underlying Tone 2 to see if the impressionistic coding was appropriate. The tone contours of the z-score transformed f0 for the relevant first syllables are shown in Figure *8*. For the benefit of the reader, we also present the tone contour for an underlying Tone 3 on the first syllable that comes from a derived Tone 3 failing to trigger Tone 3 sandhi on the preceding syllable. By doing so, a three-way visual comparison is possible at the position of the first syllable under the same phonological environment, i.e. before derived Tone 3.

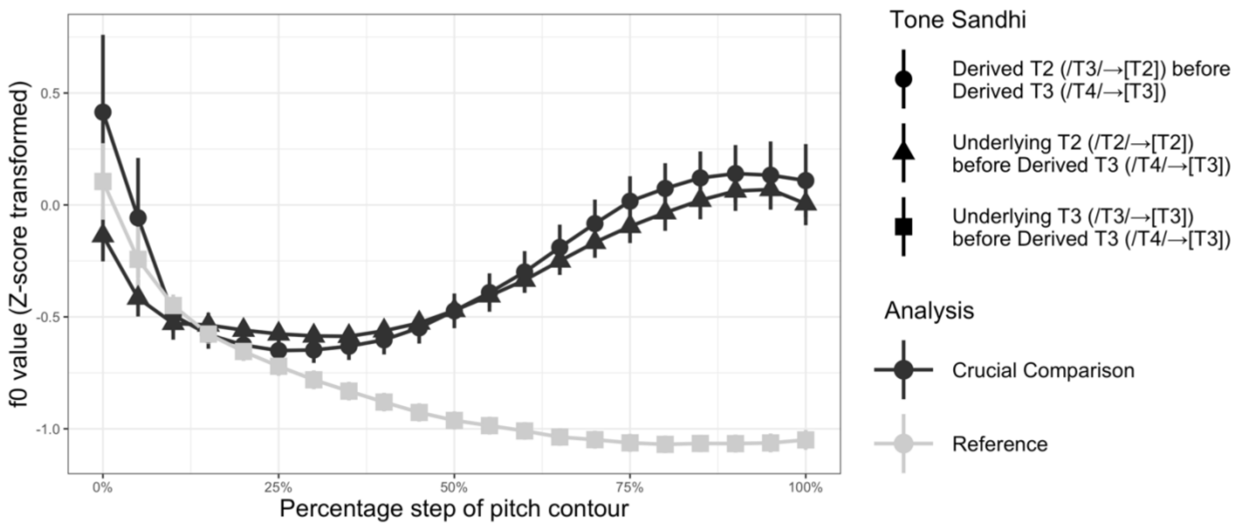


Figure 8: Contours comparison of the first syllable in Experiment 2 (Tone 4)
(Error bars indicate standard error)

Based on the visual inspection of the data, the derived Tone 2 that undergoes Tone 3 sandhi with reference to the following derived Tone 3 is phonetically highly similar to an underlying Tone 2 with regard to the f0 contour. Both derived Tone 2 and underlying Tone 2 f0 contours are phonetically very different from underlying Tone 3. Furthermore, as with the other tone sandhi processes discussed in this paper, there is incomplete phonetic neutralization of the derived Tone 2 (from an underlying Tone 3) and the underlying Tone 2 in the first syllable. With the modelling method introduced in Section 4.5, the addition of a Tone Sandhi condition improves the model on the quadratic term as shown by comparing Model 3 and Model 4 ($x^2(1)$= 4.96, p=0.03), but not on the intercept as shown by comparing Model 1 and Model 2 ($x^2(1)$=2.16, p=0.14) or the linear term as shown by comparing Model 2 and Model 3 ($x^2(1)$= 1.10, p=0.29). Figure *9* shows how the full model (Model 4) with the assumption of tone sandhi affecting every fixed effect fits the observed data. And the parameter estimates for the full model are summarized in Table *8*.

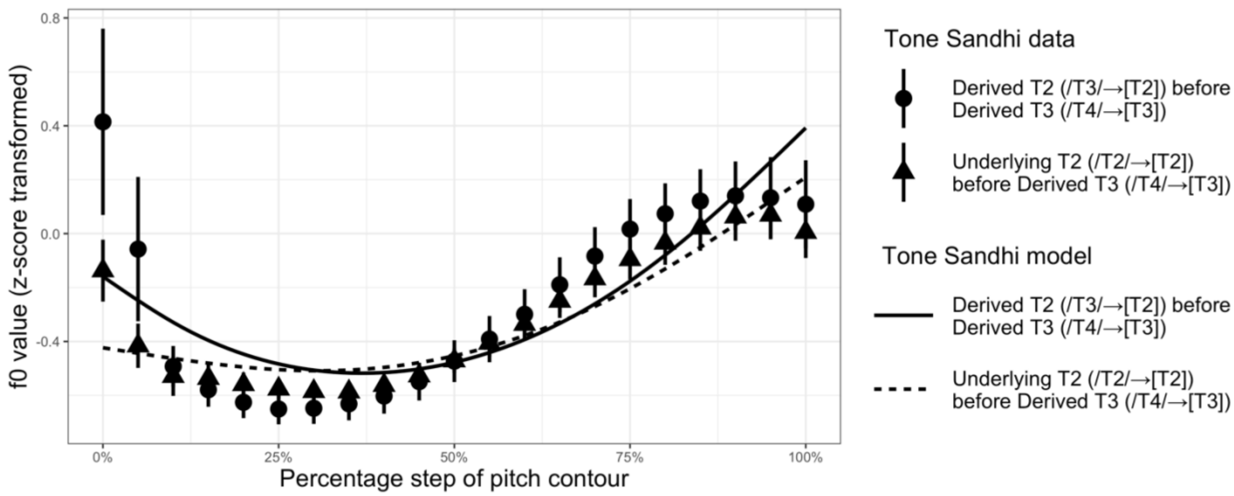


Figure 9: Observed data and Growth Curve Model fits for derived and underlying Tone 2 before derived Tone 3 (Error bars indicate standard error)

	Estimate	Std. Error	t	p
Intercept	-0.26	0.06	-4.07	<0.01
Linear	10.00	2.69	3.72	<0.01
Quadratic	9.63	1.89	5.11	<0.01
Tone Sandhi: Intercept	-0.04	0.03	-1.51	0.13
Tone Sandhi: Linear	-1.32	1.30	-1.02	0.31
Tone Sandhi: Quadratic	-2.88	1.29	-2.23	0.02

Table 8: Parameter estimates of the full model with the assumption of tone sandhi affecting every fixed effect (baseline: derived Tone 2)

However, consistent with our larger claim, this should not be interpreted as incomplete phonological neutralization. The mean difference in f0 between underlying Tone 2 and derived Tone 2 across all steps is only 1 Hz, which is *much* lower than the Just Noticeable Difference of f0 value (7 Hz) for Mandarin speakers (Jongman et al., 2017). The f0 difference (f0 of underlying Tone 2 - f0 of derived Tone 2) of each step is summarized in Table *9*. This indicates that native speakers of Huai'an may not be able to distinguish underlying vs derived Tone 2s and therefore are likely to analyze them to be in the same category in phonology. It is worth noting that an assumption has been made here that a phonetic difference that is *much* smaller than or around Just Noticeable Difference means phonologically complete neutralization, and a phonetic difference that is *much* bigger than the Just Noticeable Difference is compatible with both phonologically complete neutralization (as in Huai'an Tone 1 and Tone 4 sandhis) and phonologically incomplete neutralization. We acknowledge that some previous studies on incomplete neutralization have shown that phonetic differences that are smaller than the relevant Just Noticeable Difference are still perceptually distinguishable (Port & O'Dell, 1985; Warner et al., 2004; inter alia). However, the substantial phonetic difference between derived Tone 2 and underlying Tone 3 and the phonetic similarity between derived Tone 2 and underlying Tone 2 are difficult to account for by any mechanism known to us other than Tone 3 sandhi – it cannot simply be random variation or a co-articulatory change. Therefore, the impressionistic coding was in our opinion appropriate.

Step	f0 difference (Hz)	Step	f0 difference (Hz)
0	-8	11	2
1	-8	12	2
2	0	13	1
3	2	14	1
4	2	15	1
5	3	16	1
6	3	17	1
7	3	18	2
8	3	19	3
9	2	20	3
10	2		

Table 9: f0 Difference of each step for first syllable in Experiment 2 (Tone 2)

To summarize the results of Experiment 2, we showed using the feeding interaction between Tone 4 sandhi and Tone 3 sandhi, that the Tone 4 sandhi results in a phonological completely *derived* Tone 3. Despite this phonologically complete neutralization, we observed (a rather large) incomplete neutralization between the derived Tone 3 and underlying Tone 3 in the same surface tonal context. The experiment therefore replicates the results of Experiment 1.

6. Discussion

This paper offers two clear cases of incomplete neutralization based on data from Huai'an high-register tone sandhi processes. We observed robust phonetic differences (with large effect sizes) between a *derived* Tone 3 and an *underlying* Tone 3 in two independent experiments. This indicates that the observed effect is not likely to be a 'false positive' or functionally unimportant. Moreover, the cases of Huai'an circumvent the potential interference of orthography by presenting stimuli in Chinese characters. Therefore, some previous criticisms related to experimental design and the interpretation of data do not apply to the current Huai'an evidence.

A crucial aspect of the paper is that we first established that the relevant tone sandhi processes are in fact phonological processes. To establish this fact, we look at the phonological behavior of the derived tones, which to us is the best way of establishing phonological representations. More specifically, we looked at cases of tone sandhi that had feeding interactions, namely high-register tone sandhis including Tone 1 sandhi (Experiment 1) and Tone 4 sandhi (Experiment 2) feed Tone 3 sandhi in Huai'an Mandarin. This establishes the fact that the Tone 1 and Tone 4 sandhi processes are indeed cases of phonological neutralization. Despite this, we observed incomplete phonetic neutralization between underlying Tone 3 and derived Tone 3s stemming from the two tone sandhi processes. Consequently, our results establish the fact that phonologically complete neutralization can still be phonetically incomplete.

6.1 Phonological representation of Mandarin tone

It is worth noting that the interpretation of Huai'an tone sandhi cases as incomplete neutralization relies on the general consensus that a Mandarin tone is a single phonological unit despite appearing as a phonetic tonal contour (Bao, 1990, 1992; Yip, 1989; inter alia). Based on this view, phonologically, it is not possible for part of a Mandarin contour tone to neutralize while part of the tone remains unchanged. Perhaps, the most convincing evidence for this single-phonological-unit representation in Mandarin languages comes from Contour Tone spreading. And the most discussed case is undoubtedly Danyang (Chan, 1991; M. Y. Chen, 1991; Yip, 1989; data from Lü, 1980). The pattern of interest is given in (8):

(8) a) 2-syllable: hl. lh
b) 3-syllable: hl. hl. lh
c) 4-syllable: hl. hl. hl. lh

Note: 'h' means high tone, 'l' means low tone, and '.' indicates syllable boundary.

According to the analysis by Yip (1989), in these cases, a falling tone is associated with the first syllable and a rising tone is associated with the last syllable. Then the falling tone spreads rightwards over the domain as one single unit. If the falling tone is not a unit in phonology, one would expect only the low tone but not the whole contour to spread. A similar phenomenon of tone spreading is also found in Changzhi (Hou, 1983). It is worth noting that Duanmu (1994) challenges the above evidence by pointing out that Contour Tone spreading examples are only found in two languages and restricted to certain morpho-syntactic structures. However, since Changzhi City and Danyang City are geographically far away from each other (roughly 456 miles or 734 kilometers apart), tone spreading may be discovered in more languages and potentially more morpho-syntactic structures. To summarize, despite dispute, the tone spreading pattern itself offers strong support for phonological contour tone. It is also worth pointing out that despite disagreement with the single-unit analysis of contour tones, Duanmu (1994) claims that tone sandhis results in a categorical change, which is argued in this paper to support the interpretation of incomplete neutralization in Huai'an.

With the above phonological viewpoint of tonal representations as backdrop, in Huai'an, the fact that both *derived* Tone 3 and *underlying* Tone 3 can trigger Tone 3 sandhi suggests that a derived Tone 3 is phonologically identical to underlying tone 3. In fact, to the best of our knowledge, we are not aware of any Mandarin languages where only underlying Tone 3 triggers Tone 3 sandhi, and not derived Tone 3 - this would be a correlation that is accounted for by analyzing it as phonological neutralization. However, a phonetic difference on any part of the contour between a derived contour tone and its underlying counterpart indicates phonetic incomplete neutralization of the whole contour tone unit. In the case of Tone 1 and Tone 4 sandhis in Huai'an, there is a clear phonetic difference at the tonal offset position as shown in Experiment 1 and 2.

Based on the above, we would like to explicitly acknowledge that our claims in the paper about incomplete phonetic neutralization in the face of complete phonological neutralization are contingent on the phonological representations we have assumed. As we see it, it cannot be any other way. Note, the argument for incomplete neutralization in any language depends on a certain set of assumed phonological representations. For example, in German, the interpretation of incomplete neutralization depends on the devoicing rule actually resulting in a [-voice] feature (or equivalent). If the devoicing process results in some other phonological representation with similar phonetics, then the whole issue of incomplete neutralization vanishes, and there is no need to entertain any more gradience in the phonological system to explain the observed phonetic patterns. In fact, a version of such a featural account is implied by Hale et al. (2007), who argue that language-specific phonetics can in fact be accounted for by different phonological feature combinations. Similarly, in Huai'an, it is possible to explain what is observed in the phonetics by changing or adding new phonological representations, but then of course independent evidence of the same representations in the language or in other related languages generally needs to be provided, otherwise it becomes an ad hoc, and therefore unjustified, claim. More generally, any set of representations or computations cannot simply be post-hoc accounts of the data/patterns but need to be independently justified claims.

6.2 Desiderata for any explanation for incomplete neutralization

With the two clear cases of incomplete neutralization, the next step is naturally the explanation for incomplete neutralization. Due to the limitation of the current study, the exact source of incomplete neutralization cannot be pinpointed. However, we would like to lay out the desiderata that we think any explanation of incomplete neutralization must achieve and illustrate the problems with previous explanations alongside.

(9) Desiderata for a theory of incomplete neutralization

a) The simplest explanation of why incomplete neutralization exists as a phenomenon.
b) An explanation for the actual distribution of effect sizes among different phonological processes and within a single phonological process.
c) An explanation of why 'over-neutralization' is never observed.
d) An explanation of how a feeding interaction is possible where the derived representation still incompletely neutralizes with the element that triggers the process.
e) Related to (d)), an explanation of why incompletely neutralized segments can trigger the process, but other phonetically similar segments do not.

First, to ensure the priority of a relatively simple theoretical model, explanations that can solve the problem while retaining a relatively simple phonological model should be considered first (Occam's razor/law of parsimony). Consequently, if independently needed performance mechanisms have the potential to account for the observation of incomplete phonetic neutralization, they should be prioritized. Consistent with this principle, in the current study, the difference in Tone 3 sandhi application rates is assigned to independently needed performance factors of phonological planning, and therefore there is no need to complicate our understanding of the relevant phonological (tonal) representations. For the explanation of incomplete neutralization, beyond previously identified factors such as orthography and task effects, the best performance factors in our opinion that need to be explored further are again phonological planning (Kilbourn-Ceron & Goldrick, 2021; Tanner et al., 2017; Wagner, 2012) or cascaded activation of morphemes during production (Goldrick & Blumstein, 2006). If they are able to account for the patterns, we would be able to maintain a much simpler and, consequently, more predictive phonological theory.

The second challenge facing theories of incomplete neutralization is the systematic disparity in effect sizes (9b)). Any proposed theory should explain among the observed cases why effect sizes of incomplete neutralization are rather small in devoicing processes (as in German, Dutch, Russian …), but can be quite large as in Huai'an tone sandhis or Japanese vowel lengthening. Moreover, the proposed explanation should also account for the newly found disparity in effect sizes within a single phonological process as in two Huai'an tone sandhis. In Experiment 1, the effect size is very small at the tonal onset position as shown in Table *4*, and the effect size become quite large as the contour progresses. A similar pattern is also found in Experiment 2 as shown in

Table *7*. A model that can simply account for a variety of effect sizes misses the systematic nature among different neutralization processes and within a single time-varying neutralization process.

The third challenge is that the proposed explanation should not only predict cases of 'incomplete neutralization' where the derived category is phonetically close to an underlying category (and in fact, between the phonetic manifestation of two underlying categories – its own UR and the phonological representation it is putatively changing to), but also avoid predicting cases of 'over-neutralization' where the degree of application is beyond the phonetic distribution of the underlying category it is neutralizing to (9c)). Back to the case of German devoicing, under the scenario of 'incomplete neutralization', the phonetic cues of *derived* voiceless stops fall between *underlying* voiceless stops and *underlying* voiced stops. While under the scenario of 'over-neutralization', the phonetic cues of *underlying* voiceless stops fall between *derived* voiceless stops and *underlying* voiced stops. However, only 'incomplete neutralization' has been observed in examined languages including Huai'an. This observation would be particularly problematic for purely exemplar representations (Brown & McNeill, 1966; Bybee, 1994; Goldinger, 1996, 1997; Port & Leary, 2005; Roettger et al., 2014; inter alia). Many previous theories account for the absence of over-neutralization by proposing some mechanism where phonetically incomplete neutralization is simply intermediate between two representations as it results from a blend of all phonetic cues of two distinct representations (Braver, 2019; Gafos & Benus, 2006; Smolensky et al., 2014; Van Oostendorp, 2008).[22] Such theories are either not specific enough, or other independently needed mechanisms need to incorporate to capture the systematic disparity in effect sizes illustrated in the previous challenge (9b)).

The fourth challenge that any theory of incomplete neutralization faces is to explain how a feeding interaction is possible where the derived representation still incompletely neutralizes with the element that triggers the process (9d)). In the case of Huai'an, the Tone 3 output of the high-register tone sandhi processes can feed the low-register Tone 3 sandhi process as in (3) despite incompletely neutralizing with underlying Tone 3 in the phonetics. Any categorical theory of phonological representations naturally accounts for this as is observed rule process/rule interactions. Of course, it is possible for a theory of gradient phonological representations to do so too; however, to assess the effectiveness of such a theory, one needs to grapple with the specifics of the representations and computations proposed. Back to the issue of Tone 3 sandhi application rate difference, if one were to propose that the differential application rates are a consequence of gradient phonological representations, where phonetic proximity triggers application of a process, then one has to address two things: first, why do we see the gradience in application rates with the derived category but not with the underlying category, though both vary in terms of phonetic manifestations?; second, we need to ensure that other phonetically similar sounds do not trigger the process too (9e)). For example, in German, though both voiced obstruents and sonorants are phonetically voiced, only voiced obstruent devoice at the end of a prosodic word. One may grant that the distinction between obstruents and sonorants is a difference in phonological representations, however, by making use of such distinction, a view of category is implicitly implemented.

We raise these challenges here to move the goal-post in a constructive direction on the debate about incomplete neutralization. Given the above desiderata, we believe previous explanations are not perfectly satisfying, and therefore the phenomenon of incomplete neutralization remains an open problem.

[22] Note, typically, incomplete neutralization is argued to be a blend of the surface representation and the underlying representation, or the surface representation and a base representation, or two co-activated surface representations.

7. Conclusion

The primary goal of this paper is to offer two clear cases of incomplete neutralization using Huai'an data. Our results suggest that incomplete phonetic neutralization can in fact have a large effect size, and more importantly that the phenomenon does not automatically inform us of (gradient) phonological representations. Furthermore, echoing the general advice in Roettger et al. (2014), we'd like to encourage more work on the topic and on our particular claim, since the acceptance of any phenomenon should not be based on a single study or a single language, and only by accumulating converging evidence from different methodologies can we be more certain of it.

Finally, the phenomenon of incomplete neutralization highlights a discrepancy between the *Standard generative view of Phonology* (Kenstowicz, 1994; Pierrehumbert, 2002), wherein the output of phonological computation (the surface phonological representation) uniquely feeds into a phonetics module, and *Classic generative view of Phonology*, where phonology is seen as *knowledge* (Chomsky, 1965; Chomsky & Halle, 1965, 1968; inter alia). Note, both views represent feed-forward models, where phonological computation feeds into phonetic manifestations, but phonetic manifestations cannot feed into phonological computation. However, as per the latter view, linguistic performance is a multi-factorial problem, and linguistic knowledge (i.e., competence) is only one of the many factors involved (Chomsky, 1964, 1965; Schütze, 1996; Valian, 1982; Warner et al., 2004; inter alia).[23]

Our results from Huai'an tone neutralizations are problematic for the *Standard generative view of Phonology* - if phonetic manifestations depend solely on the output of phonology and nothing else, then it is of course the case that such a view cannot account for cases where phonological neutralization can still result in distinctness in the phonetics. However, our results are not in conflict with the *Classic generative view of Phonology*. Phonology, as per this latter view, is conceived of as grammatical knowledge that is used by a speaker to map a string of lexical items in a specific syntactic structure to articulations, and the use of this knowledge is affected by multiple other performance factors. Consequently, gradience in performance, and more specifically differences in speech production between two identical surface phonological representations, are not surprising, i.e., there is no tension between incomplete phonetic neutralization and categorical phonological neutralization for the *Classic generative view of Phonology*; instead, the actual mystery as per this view has always been with any observed cases of *complete* phonetic neutralization stemming from a process of phonological neutralization.

Competing interests

The authors have no competing interests to declare.

References

Audacity Team. (2019). Audacity [Computer Program]. Version 2.3.2, retrieved October 3, 2019 from: https://www.audacityteam.org/.

[23] We are not aware of any explicit argumentation that has ever been put forward in support of the **Standard generative view** over the **Classic generative view**. So, we are at a loss as to precisely when and, more importantly, *why* this change in viewpoints occurred. Here, we simply note the discrepancy.

Badia Margarit, Antonio M. (1962). *Gramática catalana*, 1. Madrid: Editorial Gredos.
Baldwin, Scott A. & Hoffmann, John P. (2002). The dynamics of self-esteem: A growth-curve analysis. *Journal of youth and adolescence*, *31*(2), 101-113.
Bao, Zhiming. (1990). *On the nature of tone* [PhD Thesis]. Massachusetts Institute of Technology.
Bao, Zhiming. (1992). Toward a typology of tone sandhi. *Annual Meeting of the Berkeley Linguistics Society*, *18*(2), 1-12.
Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steven. (2021). *Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4.
Boersma, Paul & Weenink, David. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.41, retrieved 25 March 2021 from http://www.praat.org/.
Braver, Aaron. (2019). Modelling incomplete neutralisation with weighted phonetic constraints. *Phonology*, *36*(1), 1-36.
Braver, Aaron & Kawahara, Shigeto. (2016). Incomplete neutralization in Japanese monomoraic lengthening. In *Proceedings of the Annual Meetings on Phonology* (Vol. 2).
Brown, Roger & McNeill, David. (1966). The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, *5*(4), 325-337.
Bybee, Joan L. (1994). A view of phonology from a cognitive and functional perspective. *Cognitive Linguistics*, *5*(4), 285-305.
Chan, Marjorie K. (1991). Contour-tone spreading and tone sandhi in Danyang Chinese. *Phonology*, 237-259.
Chao, Yuen-Ren. (1930). A System of Tone-Letters. *Le Maître Phonétique*, *30*: 24-27.
Charles-Luce, Jan & Dinnsen, Daniel A. (1987). A reanalysis of Catalan devoicing. *Journal of Phonetics*, *15*(2), 187-190.
Chen, Matthew Y. (1991). An overview of tone sandhi phenomena across Chinese dialects. *Journal of Chinese Linguistics Monograph Series*, *3*, 111-156.
Chen, Matthew Y. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press.
Chen, Si & Zhang, Caicai & McCollum, Adam G. & Wayland, Ratree. (2017). Statistical modelling of phonetic and phonologised perturbation effects in tonal and non-tonal languages. *Speech Communication*, *88*, 17-38.
Chen, Si & Li, Bin. (2021). Statistical modeling of application completeness of two tone sandhi rules. *Journal of Chinese Linguistics*, *49*(1), 106-141.
Chinese Government. (1958). *Hanyu pinyin fang'an [Scheme of the Chinese Phonetic Alphabet]*.
Chomsky, Noam. (1964). The development of grammar in child language: Formal discussion. *Monographs of the Society for Research in Child Development*, *29*, 35-39.
Chomsky, Noam. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.
Chomsky, Noam & Halle, Morris. (1965). Some controversial questions in phonological theory. *Journal of linguistics*, *1*(2), 97-138.
Chomsky, Noam & Halle, Morris. (1968). *The sound pattern of English*. Harper and Row, New York.
Cohn, Abigail C. (1993). Nasalisation in English: Phonology or phonetics. *Phonology*, *10*(1), 43-81.
Dinnsen, Daniel A. & Charles-Luce, Jan. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics*, *12*(1), 49-60.

Dmitrieva, Olga. (2005). *Incomplete neutralization in Russian final devoicing: Acoustic evidence from native speakers and second language learners* [PhD Thesis]. University of Kansas.

Duanmu, San. (1994). Against contour tone units. *Linguistic Inquiry*, *25*(4), 555-608.

Duanmu, San. (2007). *The phonology of standard Chinese*. OUP Oxford.

Du, Naiyan & Lin, Yen-Hwei. (2021). Post-Lexical tone 3 sandhi domain-building in Huai'an Mandarin: multiple domain types and free application. *University of Pennsylvania Working Papers in Linguistics*, *27*(1), 6.

Dunbar, Ewan. (2013). *Statistical Knowledge and Learning in Phonology* [PhD Thesis]. University of Maryland, College Park.

Ernestus, Mirjam & Baayen, R. Harald. (2006). The functionality of incomplete neutralization in Dutch: The case of past tense formation. In Goldstein, Louis M., Whalen, Douglas H. & Best, Catherine T. (eds.), *Papers in Laboratory Phonology VIII*. Mouton de Gruyter, 27-49.

Ernestus, Mirjam & Lahey, Mybeth & Verhees, Femke & Baayen, R. Harald. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America*, *120*, 1040-1051.

Ferreira, Fernanda & Swets, Benjamin. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57-84.

Fougeron, Cécile & Keating, Patricia A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728-3740.

Fourakis, Marios & Iverson, Gregory K. (1984). On the 'incomplete neutralization' of German final obstruents. *Phonetica*, *41*(3), 140-149.

Gafos, Adamantios I. & Benus, Stefan. (2006). Dynamics of phonological cognition. *Cognitive Science*, *30*(5), 905-943.

Goldinger, Stephen D. (1996). Words and voices: Episodic traces in spoken word identification and recognition in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166-1183.

Goldinger, Stephen D. (1997). Words and voices: Perception and production in an episodic lexicon. In Johnson, Keith & Mullennix, John W. (eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, 33-65.

Goldrick, Matthew & Blumstein, Sheila E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649-683.

Hale, Mark & Kissock, Madelyn & Reiss, Charles. (2007). Microvariation, variation, and the features of universal grammar. *Lingua*, *117*(4), 645-665.

Hastie, Trevor & Tibshirani, Robert. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hou, Jingyi. (1983). Changzhi fangyan jilüe [Notes on the Changzhi dialect]. *Fangyan [Dialect]*, *1983*(4), 260-274.

Huang, Borong & Liao, Xudong. (2017). *Xiandai hanyu [Contemporary Chinese]*. Higher Education Press.

Itô, Junko. (1990). Prosodic minimality in Japanese. *CLS*, *26*(2), 213-239.

Itô, Junko & Armin Mester. (1999). The phonological lexicon. In Tsujimura, Natsuko. (eds.), *The handbook of Japanese linguistics*. Maiden, Mass. & Oxford: Blackwell, 62-100.

Itô, Junko & Mester, Armin. (2003). Weak layering and word binarity. In Honma, Takeru & Okazaki, Masao & Tabata, Toshiyuki & Tanaka, Shin'ichi (eds.), *A new century of*

phonology and phonological theory: a Festschrift for Professor Shosuke Haraguchi on the occasion of his sixtieth birthday. Tokyo: Kaitakusha, 26-65.
Itô, Junko & Mester, Armin. (2013). Prosodic subcategories in Japanese. *Lingua*, *124*, 20-40.
Japan Broadcasting Corporation. (1998). *The Japanese Language Pronunciation and Accent Dictionary*. Tokyo: NHK.
Jassem, Wiktor & Richter, Lutosława. (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics*, *17*(4), 317-325.
Jiao, Lidong. (2004). *Huai'an fangyan de shengdiao fenxi [An analysis of tones in Huai'an dialect]* [Master Thesis]. Tianjin Normal University.
Jongman, Allard & Qin, Zhen & Zhang, Jie & Sereno, Joan A. (2017). Just noticeable differences for pitch direction, height, and slope for Mandarin and English listeners. *The Journal of the Acoustical Society of America*, *142*(2), EL163-EL169.
Jongman, Allard & Wang, Yue & Moore, Corinne B. & Sereno, Joan A. (2006). Perception and production of Mandarin Chinese tones. In Li, Ping & Tan, Lihai & Bates, Elizabeth & Tzeng, Ovid J. L. (eds.), *Handbook of East Asian psycholinguistics*. Cambridge, UK: Cambridge University Press, 209-217.
Kenstowicz, Michael J. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell.
Kharlamov, Viktor. (2012). *Incomplete neutralization and task effects in experimentally-elicited speech: Evidence from the production and perception of word-final devoicing in Russian* [PhD Thesis]. Université d'Ottawa/University of Ottawa.
Kilbourn-Ceron, Oriana & Goldrick, Matthew. (2021). Variable pronunciations reveal dynamic intra-speaker variation in speech planning. *Psychonomic Bulletin & Review*, *28*, 1365-1380.
Labov, William. (1963). The social motivation of a sound change. *Word*, *19*(3), 273-309.
Laplace, Pierre S. (1820). *Théorie analytique des probabilités*. Courcier.
Leben, William R. (1973). *Suprasegmental phonolog*y [PhD Thesis]. Massachusetts Institute of Technology.
Li, Rong. (1989). Hanyu fangyan de fenqu [The geographic division of Chinese dialects]. *Fangyan [Dialect]*, *4*, 19.
Lionnet, Florian. (2017). A theory of subfeatural representations: The case of rounding harmony in Laal. *Phonology*, *34*(3), 523-564.
Lobanov, Boris M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606-608.
Lü, Shuxiang. (1980). Danyang fangyan de shengdiao xitong [The tonal system of the Danyang dialect]. *Fangyan [Dialect]*, *1980*(2), 85-122.
McCarthy, John J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207-263.
Manaster Ramer, Alexis. (1996). A letter from an incompletely neutral phonologist. *Journal of Phonetics*, *24*(4), 477-489.
Mascaró, Joan. (1987). Underlying voicing recoverability of finally devoiced obstruents in Catalan. *Journal of Phonetics*, *15*(2), 183-186.
Matsui, Mayuki. (2015). *Roshiago ni okeru yuuseisei no tairitsu to tairitsu no jakka: Onkyo to chikaku [Voicing contrast and contrast reduction in Russian: Acoustics and perception]* [PhD Thesis]. Hiroshima University.
McArdle, John J. & Nesselroade, John R. (2003). Growth curve analysis in contemporary psychological research. In Velicer, Wayne F. & Schinka, John A. (eds.), *Handbook of psychology: Research methods in psychology*, *2*. New York: Wiley, 447-480.

McCollum, Adam. (2019). Gradient morphophonology: Evidence from Uyghur vowel harmony. In *Proceedings of the Annual Meetings on Phonology* (Vol. 7).
Mester, Armin. (1990). Patterns of truncation. *Linguistic Inquiry*, *21*(3), 478-485.
Milroy, James. (2001). Language ideologies and the consequences of standardization. *Journal of sociolinguistics*, *5*(4), 530-555.
Mirman, Daniel. (2017). *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.
Mirman, Daniel & Dixon, James A. & Magnuson, James S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of memory and language*, *59*(4), 475-494.
Mori, Yoko. (2002). Lengthening of Japanese monomoraic nouns. *Journal of Phonetics*, *30*(4), 689-708.
Pierrehumbert, Janet B. (2002). Word-specific phonetics. In Gussenhoven, Carlos & Warner, Natasha. (eds.), *Papers in Laboratory Phonology VII*. Mouton de Gruyter, 101-139.
Pierrehumbert, Janet B., Beckman, Mary E. & Ladd, D. Robert. (2000). Conceptual foundations of phonology as a laboratory science. *Phonological Knowledge: Conceptual and Empirical Issues*, 273-304.
Piroth, Hans G. & Janker, Peter M. (2004). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics, 32*(1), 81-109.
Port, Robert F. & Leary, Adam P. (2005). Against formal phonology. *Language*, *81*(4), 927-964.
Port, Robert F. *&* O'Dell, Michael L. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*(4), 455-471.
Poser, William J. (1990). Evidence for foot structure in Japanese. *Language*, *66*(1), 78-105.
Prince, Alan & Paul Smolensky. (1993). *Optimality Theory*, Rutgers University Center for Cognitive Science Technical Report 2, Rutgers University, New Brunswick, New Jersey.
R Core Team. (2021). R: A language and environment for statistical computing. [Computer program] Version 1.4.1106, retrieved 25 March, 2021 from https://www.rstudio.com.
Ramsey, S. Robert. (1989). *The languages of China*. Princeton University Press.
Roettger, Timo B. & Winter, Bodo & Grawunder, Sven & Kirby, James & Grice, Martine. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *43*, 11-25.
Schütze, Carson T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
Shen, Xiaonan S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, *18*(2), 281-295.
Silverman, Daniel. (2006). *A critical introduction to phonology: Of sound, mind, and body*. A&C Black.
Slowiaczek, Louisa M. & Dinnsen, Daniel A. (1985). On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics*, *13*(3), 325-341.
Slowiaczek, Louisa. M., & Szymanska, Helena J. (1989). Perception of word-final devoicing in Polish. *Journal of Phonetics, 17*(3), 205-212.
Smolensky, Paul & Goldrick, Matthew & Mathis, Donald. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102-1138.

Tanner, James & Sonderegger, Morgan & Wagner, Michael. (2017). Production planning and coronal stop deletion in spontaneous speech. Laboratory Phonology: *Journal of the Association for Laboratory Phonology*, *8*(1), p.15.

Valian, Virginia. (1982). Psycholinguistic experiment and linguistic intuition, in Simon, Thomas W. & Scholes, Robert J. (eds.), *Language, Mind, and Brain, Hillsdale*. NJ: Lawrence Erlbaum, 179-188.

Van Oostendorp, Marc. (2008). Incomplete devoicing in formal phonology. *Lingua*, *118*(9), 1362-1374.

Wagner, Michael. (2002). The role of prosody in laryngeal neutralization. *MIT Working Papers in Linguistics*, *42*, 373-392.

Wagner, Valentin & Jescheniak, Jörg D. & Schriefers, Herbert. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 423.

Wagner, Michael. (2012). Locality in phonology and production planning. *McGill working papers in linguistics*, *22*(1), 1-18.

Wang, Chiung-Yao & Lin, Yen-Hwei. (2011). Variation in Tone 3 Sandhi: The case of prepositions and pronouns. In *Proceedings of the 23rd North American Conference on Chinese Linguistics*, 138-155.

Wang, Yuedong. (1998). Mixed-Effects Smoothing Spline ANOVA. *Journal of the Royal Statistical Society, Series. B*, *60*(1), 159-174.

Wang, Yifeng & Kang, Jian. (2012). Huai'an nanpian fangyan liangzizu lianxu biaodiao fenxi [An analysis of disyllabic tone sandhi in southern Huai'an dialect]. *Journal of Shenyang Institute of Engineering (Social Sciences)*, *8*(3), 357-359.

Warner, Natasha & Jongman, Allard & Sereno, Joan & Kemps, Rachèl. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*(2), 251-276.

Whalen, Douglas H. (1991). Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America*, *90*(4), 2311.

Whalen, Douglas H. (1992). Further results on the duration of infrequent and frequent words. *Journal of the Acoustical Society of America*, *91*(4), 2339-2340.

Wickham, Hadley & Averick, Mara & Bryan, Jennifer & Chang, Winston & McGowan, Lucy D. A. & François, Romain ... & Yutani, Hiroaki. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686.

Wright, Richard. (2004). Factors of lexical competition in vowel articulation. In Local, John & Ogden, Richard & Temple, Rosalind. (eds.), *Papers in laboratory phonology VI*, 75-87. Cambridge: Cambridge University Press.

Xu, Yi. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, *95*(4), 2240-2253.

Xu, Yi. (1997). Contextual tonal variations in Mandarin. *Journal of phonetics*, *25*(1), 61-83.

Yang, Jilin. (1995). *Zhongguo zhongxiaoxue baikequanshu* [*Encyclopedia for elementary school and middle school students*]. Harbin Publishing House.

Yip, Moira J. (1980). *The tonal phonology of Chinese* [PhD Thesis]. Massachusetts Institute of Technology.

Yip, Moira J. (1989). Contour tones. *Phonology*, *6*(1), 149-174.

Yip, Moira J. (2002). *Tone*. Cambridge University Press.

Zhang, Ning. (1997). The avoidance of the third tone sandhi in Mandarin Chinese. *Journal of East Asian Linguistics*, *6*(4), 293-338.

Appendix 1: Stimuli for Experiment 1 on Tone 1

Sentence	IPA	Pinyin	Word-by-word gloss	Translation of the whole sentence	UR	SR
吴把车	u pa tɕi	wu ba che	'Mr. Wu' 'play' 'car'	'Mr. Wu plays with cars.'	T2T3T1	T2T3T1
吴鼓分	u ku fən	wu gu fen	'Mr. Wu' 'encourage' 'points'	'Mr. Wu tries to increase points.'	T2T3T1	T2T3T1
吴打车	u ta tɕi	wu da che	'Mr. Wu' 'call' 'car'	'Mr. Wu calls for a taxi.'	T2T3T1	T2T3T1
吴把虾	u pa xa	wu ba xia	'Mr. Wu' 'play' 'shrimp'	'Mr. Wu plays with shrimp.'	T2T3T1	T2T3T1
吴摆虾	u pɛ xa	wu bai xia	'Mr. Wu' 'place' 'shrimp'	'Mr. Wu places shrimp (in a plate).'	T2T3T1	T2T3T1
吴保车	u pɔ tɕi	wu bao che	'Mr. Wu' 'protect' 'car'	'Mr. Wu protects cars.'	T2T3T1	T2T3T1
吴扒车	u pa tɕi	wu ba che	'Mr. Wu' 'grasp' 'car'	'Mr. Wu catches cars.'	T2T1T1	T2T1T1/ T2T3T1
吴估分	u ku fən	wu gu fen	'Mr. Wu' 'estimate' 'scores'	'Mr. Wu estimates scores.'	T2T1T1	T2T1T1/ T2T3T1
吴搭车	u ta tɕi	wu da che	'Mr. Wu' 'take' 'cars'	'Mr. Wu gets a ride.'	T2T1T1	T2T1T1/ T2T3T1
吴扒虾	u pa xa	wu ba xia	'Mr. Wu' 'smash' 'shrimp'	'Mr. Wu smashes shrimp (to eat).'	T2T1T1	T2T1T1/ T2T3T1
吴掰虾	u pɛ xa	wu bai xia	'Mr. Wu' 'break off' 'shrimp'	'Mr. Wu breaks off shrimp (to eat).'	T2T1T1	T2T1T1/ T2T3T1
吴包车	u pɔ tɕi	wu bao che	'Mr. Wu' 'rent' 'car'	'Mr. Wu rents cars.'	T2T1T1	T2T1T1/ T2T3T1
武把车	u pa tɕi	wu ba che	'Mr. Wu' 'play' 'car'	'Mr. Wu plays with cars.'	T3T3T1	T2T3T1

武鼓分	u ku fən	wu gu fen	'Mr. Wu' 'encourage' 'points'	'Mr. Wu tries to increase points.'	T3T3T1	T2T3T1
武打车	u ta tɕi	wu da che	'Mr. Wu' 'call' 'car'	'Mr. Wu calls for a taxi.'	T3T3T1	T2T3T1
武把虾	u pa xa	wu ba xia	'Mr. Wu' 'play' 'shrimp'	'Mr. Wu plays with shrimp.'	T3T3T1	T2T3T1
武摆虾	u pɛ xa	wu bai xia	'Mr. Wu' 'place' 'shrimp'	'Mr. Wu places shrimp (in a plate).'	T3T3T1	T2T3T1
武保车	u pɔ tɕi	wu bao che	'Mr. Wu' 'protect' 'car'	'Mr. Wu protects cars.'	T3T3T1	T2T3T1
武扒车	u pa tɕi	wu ba che	'Mr. Wu' 'grasp' 'car'	'Mr. Wu catches cars.'	T3T1T1	T3T1T1/ T2T3T1
武估分	u ku fən	wu gu fen	'Mr. Wu' 'estimate' 'scores'	'Mr. Wu estimates scores.'	T3T1T1	T3T1T1/ T2T3T1
武搭车	u ta tɕi	wu da che	'Mr. Wu' 'take' 'cars'	'Mr. Wu gets a ride.'	T3T1T1	T3T1T1/ T2T3T1
武扒虾	u pa xa	wu ba xia	'Mr. Wu' 'smash' 'shrimp'	'Mr. Wu smashes shrimp (to eat).'	T3T1T1	T3T1T1/ T2T3T1
武掰虾	u pɛ xa	wu bai xia	'Mr. Wu' 'break off' 'shrimp'	'Mr. Wu breaks off shrimp (to eat).'	T3T1T1	T3T1T1/ T2T3T1
武包车	u pɔ tɕi	wu bao che	'Mr. Wu' 'rent' 'car'	'Mr. Wu rents cars.'	T3T1T1	T3T1T1/ T2T3T1

Appendix 2: Stimuli for Experiment 2 on Tone 4

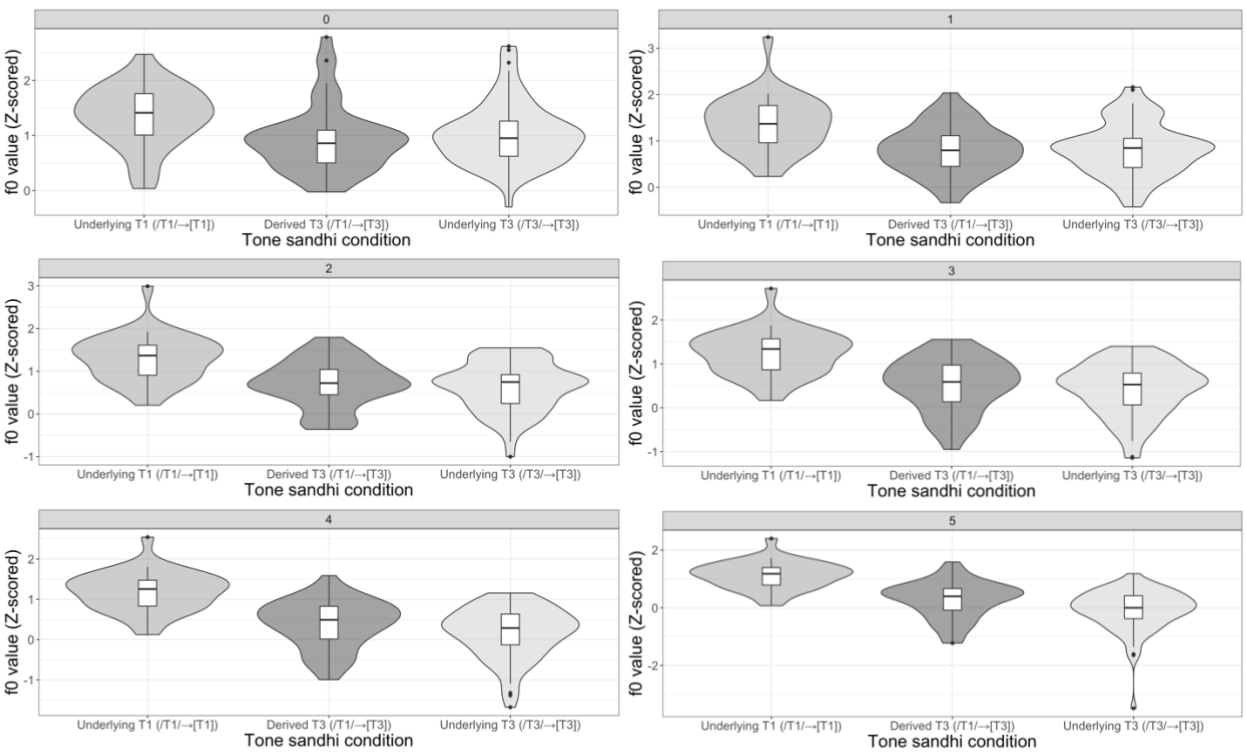
Sentence	IPA	Pinyin	Word-by-word gloss	Translation of the whole sentence	UR	SR
吴保税	u pɔ suɛi	wu bao shui	'Mr. Wu' 'protect' 'tax'	‘Mr. Wu is under bond.’	T2T3T4	T2T3T4
吴躲肉	u to ʐəɯ	wu duo rou	'Mr. Wu' 'avoid' 'meat'	'Mr. Wu avoids eating meat.'	T2T3T4	T2T3T4

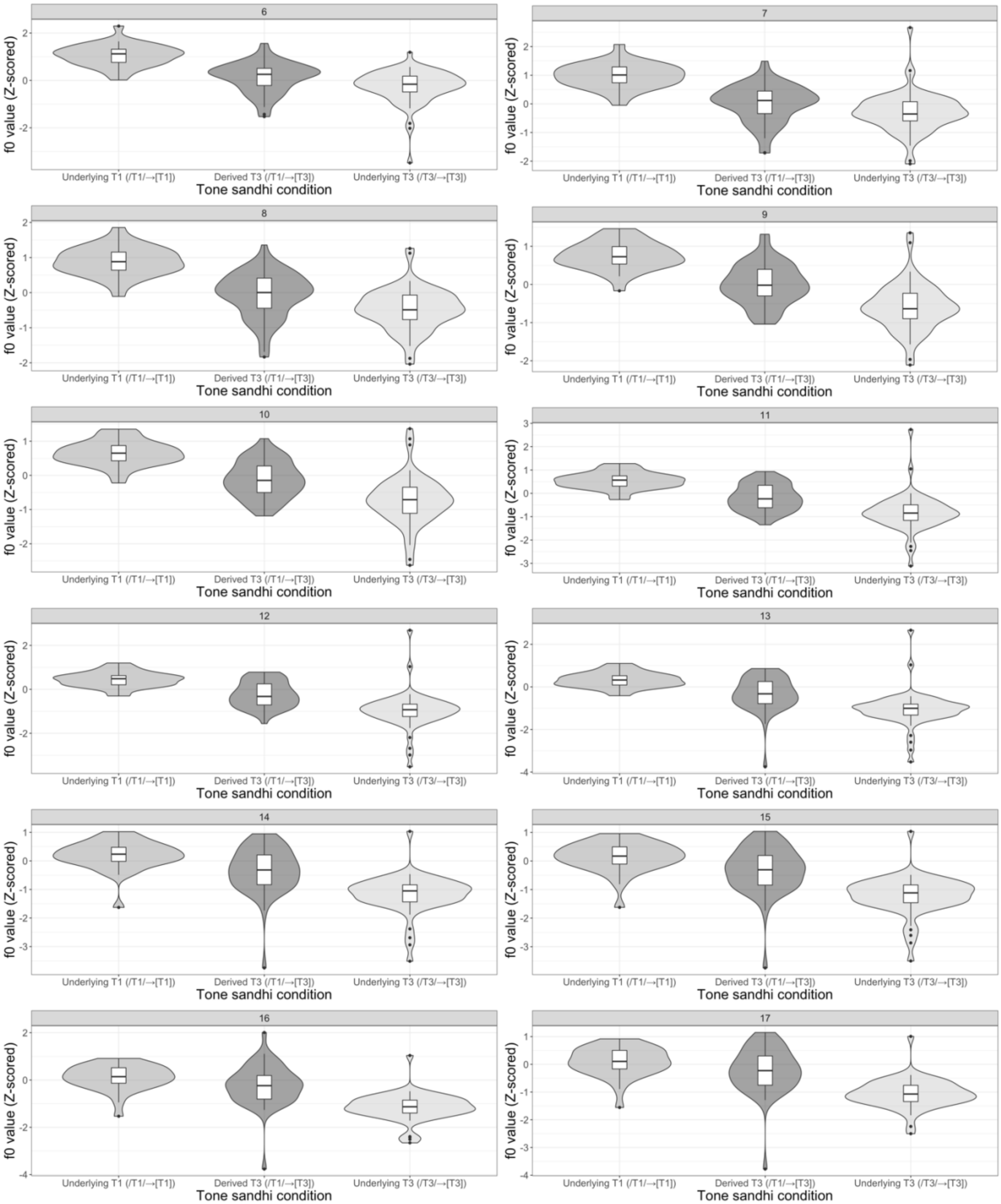
吴把脉	u pa mɛ	wu ba mai	'Mr. Wu' 'touch' 'blood vessel'	'Mr. Wu diagnoses by touching blood vessels.'	T2T3T4	T2T3T4
吴逮象	u tɛ ɕiæ̃	wu dai xiang	'Mr. Wu' 'catch' 'elephant'	'Mr. Wu catches elephants.'	T2T3T4	T2T3T4
吴补炮	u pu pʰɔ	wu bu pao	'Mr. Wu' 'replenish' 'cannons'	'Mr. Wu replenishes the stock of cannons.'	T2T3T4	T2T3T4
吴报税	u pɔ suɛi	wu bao shui	'Mr. Wu' 'declare' 'tax'	'Mr. Wu does taxes'	T2T4T4	T2T4T4/ T2T3T4
吴剁肉	u to zəɯ	wu duo rou	'Mr. Wu' 'chop' 'meat'	'Mr. Wu chops meat.'	T2T4T4	T2T4T4/ T2T3T4
吴罢卖	u pa mɛ	wu ba mai	'Mr. Wu' 'stops' 'sell'	'Mr. Wu stops selling (to protest).'	T2T4T4	T2T4T4/ T2T3T4
吴带象	u tɛ ɕiæ̃	wu dai xiang	'Mr. Wu' 'take along' 'elephant'	'Mr. Wu takes along elephants.'	T2T4T4	T2T4T4/ T2T3T4
吴布炮	u pu pʰɔ	wu bu pao	'Mr. Wu' 'deploy' 'cannons'	'Mr. Wu deploys cannons.'	T2T4T4	T2T4T4/ T2T3T4
武保税	u pɔ suɛi	wu bao shui	'Mr. Wu' 'protect' 'tax'	‘Mr. Wu is under bond.’	T3T3T4	T2T3T4
武躲肉	u to zəɯ	wu duo rou	'Mr. Wu' 'avoid' 'meat'	'Mr. Wu avoids eating meat.'	T3T3T4	T2T3T4
武把脉	u pa mɛ	wu ba mai	'Mr. Wu' 'touch' 'blood vessel'	'Mr. Wu diagnoses by touching blood vessels.'	T3T3T4	T2T3T4
武逮象	u tɛ ɕiæ̃	wu dai xiang	'Mr. Wu' 'catch' 'elephant'	'Mr. Wu catches elephants.'	T3T3T4	T2T3T4
武补炮	u pu pʰɔ	wu bu pao	'Mr. Wu' 'replenish' 'cannons'	'Mr. Wu replenishes the stock of cannons.'	T3T3T4	T2T3T4
武报税	u pɔ suɛi	wu bao shui	'Mr. Wu' 'declare' 'tax'	'Mr. Wu does taxes'	T3T4T4	T3T4T4/ T2T3T4

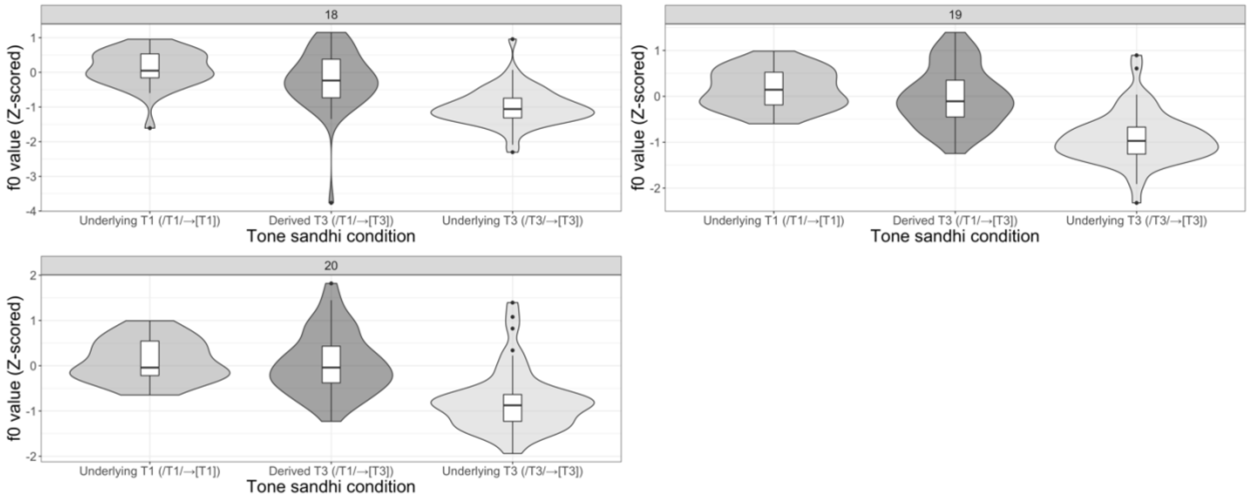
武剁肉	u to z̥əɯ	wu duo rou	'Mr. Wu' 'chop' 'meat'	'Mr. Wu chops meat.'	T3T4T4	T3T4T4/ T2T3T4
武罢卖	u pa mɛ	wu ba mai	'Mr. Wu' 'stops' 'sell'	'Mr. Wu stops selling (to protest).'	T3T4T4	T3T4T4/ T2T3T4
武带象	u tɛ ɕiæ̃	wu dai xiang	'Mr. Wu' 'take along' 'elephant'	'Mr. Wu takes along elephants.'	T3T4T4	T3T4T4/ T2T3T4
武布炮	u pu pʰɔ	wu bu pao	'Mr. Wu' 'deploy' 'cannons'	'Mr. Wu deploys cannons.'	T3T4T4	T3T4T4/ T2T3T4

Supplementary Materials

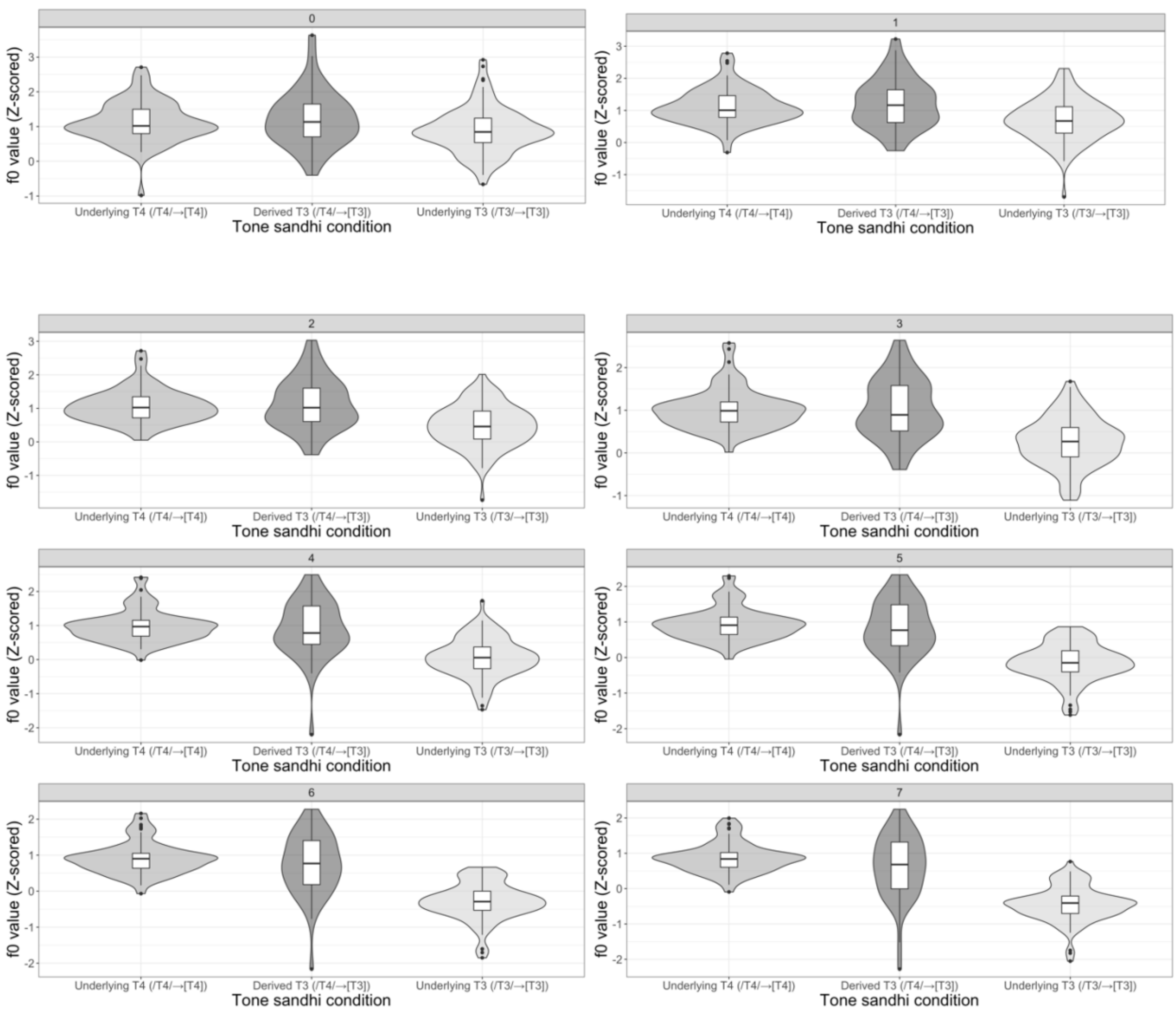
Distribution of underlying Tone 1, derived Tone 3 and underlying Tone 3 in each step in Experiment 1.

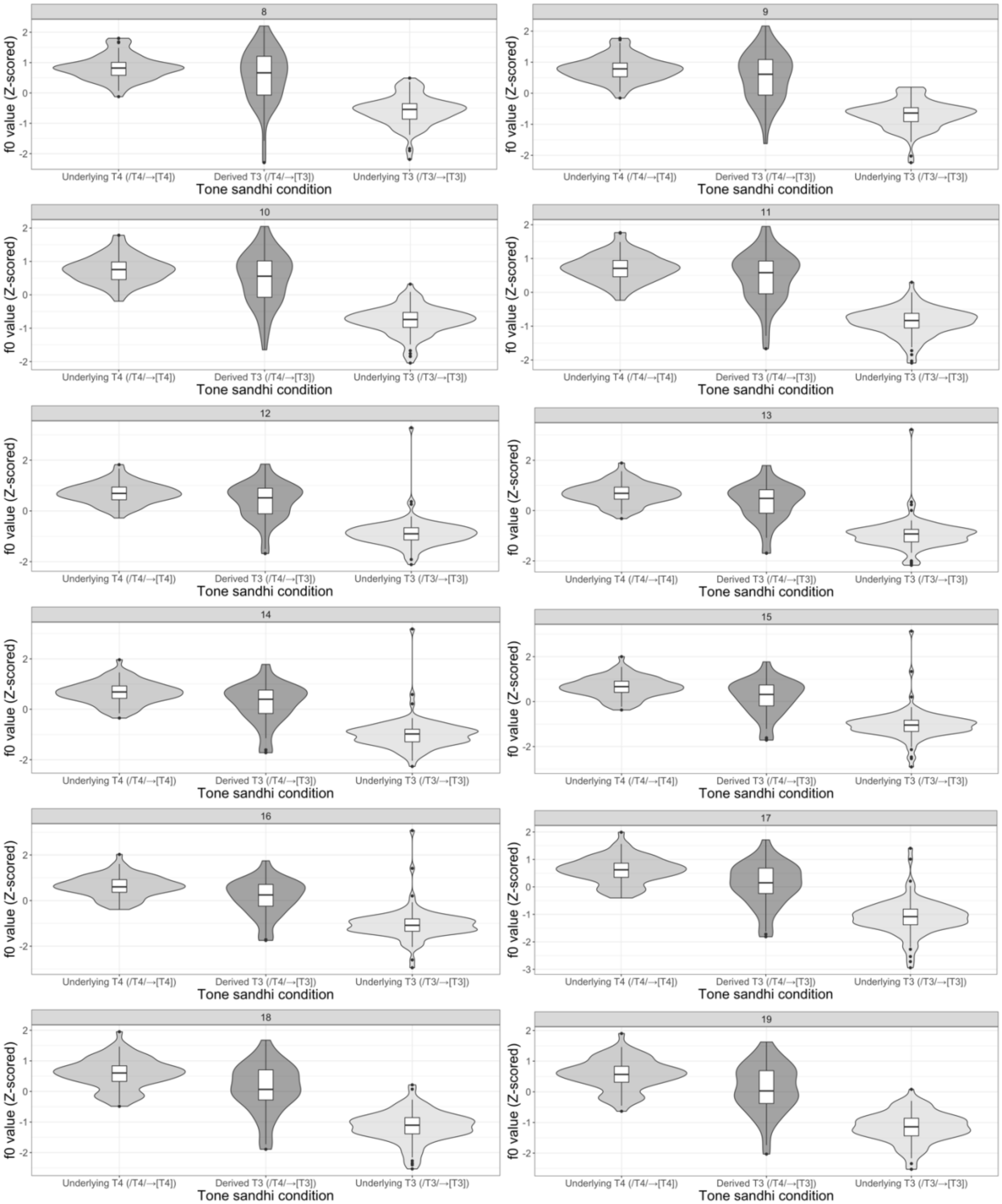


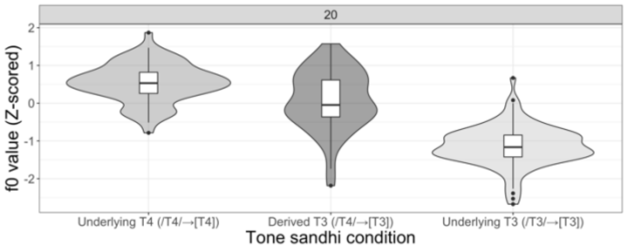

6
7
8
9
10
11
12
13
14
15
16
17
f0 value (Z-scored)
Underlying T1 (/T1/→[T1])
Derived T3 (/T1/→[T3])
Underlying T3 (/T3/→[T3])
Tone sandhi condition



Distribution of underlying Tone 4, derived Tone 3 and underlying Tone 3 in each step in Experiment 2.




8
9
10
11
12
13
14
15
16
17
18
19
f0 value (Z-scored)
Underlying T4 (/T4/→[T4])
Derived T3 (/T4/→[T3])
Underlying T3 (/T3/→[T3])
Tone sandhi condition


20
f0 value (Z-scored)
2
1
0
-1
-2
Underlying T4 (/T4/→[T4])
Derived T3 (/T4/→[T3])
Underlying T3 (/T3/→[T3])
Tone sandhi condition