frontiers

Mandarin third tone sandhi may be incompletely neutralizing in perception as well as production

Stephen Politzer-Ahles[1*], Katrina Connell[2], Lei Pan[1], Yu-Yin Hsu[1]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong,
[2]Department of Spanish, Italian, and Portuguese; Pennsylvania State University, United States

Submitted to Journal:
Frontiers in Psychology

Specialty Section:
Language Sciences

Article type:
Original Research Article

Manuscript ID:
447184

Received on:
09 Jan 2019

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

SP and YH conceived the experiment. SP, KC, LP and YH designed the experiment. KC and LP created the stimuli. KC programmed the experiment and LP collected the data. SP and KC analyzed the data. All authors wrote and approved the manuscript.

Keywords

incomplete neutralization, Mandarin tone, third tone sandhi, visual world eye-tracking, Speech Perception

Abstract

Word count: 131

Mandarin third tone sandhi is traditionally assumed to be incompletely neutralizing in production but completely neutralizing in perception, based on metalinguistic judgment tasks in which participants cannot reliably identify the underlying tone of syllables neutralized by tone sandhi. We performed a visual world eye-tracking study to see if implicit sensitivity to the differences between the surface forms influences participants' eye movement patterns, even if they cannot consciously access this for identification tasks. We found a slight trend in this direction, with participants looking more towards orthographic representations that match the underlying form of the neutralized syllable they hear. The results are statistically inconclusive but suggest that this paradigm may be able to provide evidence that Mandarin neutralized tones are indeed incompletely neutralized, and that further research along these lines is warranted.

Funding statement

This research was supported by grant G-UACX from the Department of Chinese and Bilingual Studies to SP.

Ethics statements

(Authors are required to state the ethical considerations of their study in the manuscript, including for cases where the study was exempt from ethical approval procedures)

Does the study presented in the manuscript involve human or animal subjects: Yes

Please provide the complete ethics statement for your manuscript. Note that the statement will be directly added to the manuscript file for peer-review, and should include the following information:

- Full name of the ethics committee that approved the study
- Consent procedure used for human participants or for animal owners
- Any additional considerations of the study in cases where vulnerable populations were involved, for example minors, persons with disabilities or endangered animal species

As per the Frontiers authors guidelines, you are required to use the following format for statements involving human subjects:
This study was carried out in accordance with the recommendations of [name of guidelines], [name of committee]. The protocol was approved by the [name of committee]. All subjects gave written informed consent in accordance with the Declaration of Helsinki.
For statements involving animal subjects, please use:
This study was carried out in accordance with the recommendations of 'name of guidelines, name of committee'. The protocol was approved by the 'name of committee'.

If the study was exempt from one or more of the above requirements, please provide a statement with the reason for the exemption(s).
Ensure that your statement is phrased in a complete way, with clear and concise sentences.

This study was carried out in accordance with the recommendations of the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University.

Data availability statement

Generated Statement: All datasets generated for this study are included in the manuscript and the supplementary files.

Mandarin third tone sandhi may be incompletely neutralizing in perception as well as production

Stephen Politzer-Ahles[1*], Katrina Connell[2], Yu-Yin Hsu[1], Lei Pan[1]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, Hong Kong

[2]Department of Spanish, Italian & Portuguese; Pennsylvania State University, State College, Pennsylvania, United States

*** Correspondence:**
Stephen Politzer-Ahles
sjpolit@polyu.edu.hk

Keywords: incomplete neutralization, Mandarin tone sandhi, third tone sandhi, visual world eye-tracking, speech perception

Abstract

Mandarin third tone sandhi is traditionally assumed to be incompletely neutralizing in production but completely neutralizing in perception, based on metalinguistic judgment tasks in which participants cannot reliably identify the underlying tone of syllables neutralized by tone sandhi. We performed a visual world eye-tracking study to see if implicit sensitivity to the differences between the surface forms influences participants' eye movement patterns, even if they cannot consciously access this for identification tasks. We found a slight trend in this direction, with participants looking more towards orthographic representations that match the underlying form of the neutralized syllable they hear. The results are statistically inconclusive but suggest that this paradigm may be able to provide evidence that Mandarin neutralized tones are indeed incompletely neutralized, and that further research along these lines is warranted.

1 Introduction

A common phenomenon in language is *neutralization*, whereby two or more sounds that normally would be different, are instead changed to have the same surface pronunciation in a certain context. For example, standard Dutch has a distinction between phonologically voiced and voiceless stops like /d/ and /t/, but at the end of the word both are realized as [t] (the words *baat* "benefit" and *baad* "bathe" are both pronounced [baat]) (Warner, Jongman, Sereno, & Kemps, 2004). In many cases, however, these contrasts are not completely neutralized. For example, in the Dutch example above, there are still slight pronunciation differences between *baat* and *baad*, and these differences are perceptible to listeners (Warner et al., 2004). The same is true in English: for a minimal pair like *beat* and *bead*, while /t/ and /d/ may not differ in aspiration like they would at the onset of a word, there are still several subtle acoustic differences between them, as well as large acoustic differences in the vowels that precede them (see, e.g., Flege, Munro, & Skelton, 1992). Such a pattern is known as *incomplete neutralization* (for review see Nicenboim, Roettger, & Vasisth, 2018, among others).

Incomplete neutralization has important implications for psycholinguistic accounts of speech perception and production, which must account for the consequences of incompletely neutralized representations in lexical storage and in speech comprehension. It also has implications for phonological theory—for example, struggles to theoretically account for phonological opacity (situations where a surface form exists in a context that looks like it should have triggered a change to a different surface form, but somehow does not) may be obviated if one assumes that an incompletely neutralized surface form does not actually trigger the relevant phonological alternation and thus is not actually opaque (see, e.g., Zhang, Lai, & Sailor, 2011). Finally, incomplete neutralization is pervasive; it may even be the norm, as many putatively neutralized contrasts have been found to have some remaining acoustic correlates that distinguish them, and patterns where two neutralized forms are actually indistinguishable in both acoustic measurements and perception seem to be rare (Kim & Jongman, 1996).[1]

Some neutralization patterns are argued to be incomplete in production but complete in perception: that is to say, two putatively neutralized sounds have reliable acoustic differences which can be detected with computer-assisted measurements and statistical methods, but human listeners cannot reliably detect them when tested. A classic example of this case is third tone sandhi in Mandarin. Standard Mandarin has lexical tones, such that segmentally identical syllables with different tones may correspond to different morphemes (e.g., [kwa˥][2], with a High tone, is the pronunciation of the morpheme written瓜, "melon", among others; whereas [kwa˥˨], with a Falling tone, is the pronunciation of the morpheme written 挂, "to hang"). Crucially, two tones, Rising and Low, are putatively neutralized in a certain context. When a Low tone (also known as a third tone or "tone three", hence the name "third tone sandhi") is preceded by another within the same intonational domain, it is instead pronounced as a Rising tone (see Kuo, Xu, and Yip, 2007, and Zhang & Lai, 2010, among others, for further details regarding other phonological and morphosyntactic constraints on this pattern). For instance, the morpheme written 雨 is normally pronounced [ɥy˨˩], with Low tone; but when it appears before another Low tone, as in the compound word 雨伞 [ɥy˧˥ san˨˩] "umbrella", it is instead produced with a Rising tone, homophonic with the morpheme written 鱼 [ɥy˧˥], "fish". This alternation causes the distinction between Low and Rising tones to be neutralized in pre-Low positions that license third tone sandhi. Many acoustic studies, though, have found that the distinction is not completely neutralized: a "Rising" tone derived via tone sandhi from an underlying Low tone tends to be slightly lower and have a slightly later turning point in its tonal contour (Chen, Wiltshire, & Li, 2017; Cheng, Chen, & Gubian, 2013; Liu, 2013, Peng, 2000; Yuan & Chen, 2014; Zhang & Lai, 2010; Zhang & Peng, 2013; among others). On the other hand, several studies have also shown that speakers asked to identify the underlying tone (i.e., if they are presented with an ambiguous Rising-Low disyllable and asked to identify it as an underlyingly Low-Low disyllable or an

[1] Alternatively, though, it is possible that the abundance of attested incomplete neutralization patterns could also be due to methodological practices. In short, if one contrast has many possible acoustic correlates that could be tested and could be quantified in different ways, and these are all checked for evidence of incomplete neutralization, this greatly increases the chance of finding spurious differences (see Nicenboim et al., 2018, for discussion; and Roettger, under review, for a general introduction to this statistical issue). Meta-analytic methods, as well as frequent replication, provide evidence that some classic incomplete neutralization patterns like Dutch and German final devoicing are indeed reliable (Nicenboim et al, 2018); it would be valuable to extend these methods, as well as other meta-analytic methods like the p-curve (Simonsohn, Nelson, & Simmons, 2014) to other less studied cases of putatively incomplete neutralization.

[2] Transcriptions in brackets are representations in the International Phonetic Alphabet. We use the following symbols to indicate tones: ˥ for High tone, ˧˥ for Rising tone, ˨˩ for Low tone, and ˥˨ for Falling tone.

underlyingly Rising-Low one) cannot reliably identify the correct tone (Liu, 2013; Peng, 2000; Wang & Li, 1967; Zhang & Peng, 2013). Thus, it is generally believed that while Mandarin third tone sandhi is not completely neutralizing in production (i.e., a Rising tone resulting from tone sandhi is acoustically different from an underlyingly Rising tone), it is completely neutralizing in perception (i.e., native listeners cannot reliably hear the contrast).

Nonetheless, some results in the extant literature do suggest that the subtle contrast between underlying and sandhi-derived Rising tones may influence listeners' perception. These come from a priming experiment, a visual world eye-tracking experiment, a discrimination task, and an identification task, described below.

Zhou and Marslen-Wilson (1997) report two auditory-auditory priming experiments in which participants heard primes and targets that were each disyllabic words, and in which the first syllable was sometimes an underlyingly or a sandhi-derived Rising tone. In their first experiment, the critical targets began with a sandhi-derived Rising tone (e.g., [t͡sʰaiˊˀ t͡ɕʰyˇ˩] 采取 "to carry out", the first syllable of which has Low tone in its citation form but has Rising tone in this word because it is followed by another Low tone), and reaction time in performing lexical decision to these items was sped up (relative to a control condition with unrelated primes) when they were immediately preceded by primes with a Low tone in the first syllable, like [t͡sʰaiˇ˩ xʊŋˊˀ] 彩虹 "rainbow". On the other hand, in their second experiment (and in a direct replication of their second experiment), the critical targets began with an underlyingly Rising tone (e.g., [t͡sʰaiˊˀ panˋ˥] 裁判 "referee", the first syllable of which is pronounced with Rising tone even in its citation form), and reaction time to these targets was slowed down (relative to a control condition with unrelated primes) when they were preceded by primes with a Low tone in the first syllable, like [t͡sʰaiˇ˩ naˋ˥] 采纳 "to adopt [a measure]". These patterns appear to be opposite: a sandhi target is facilitated by a Low prime whereas a Rising target is inhibited by a Low prime. This might be taken as evidence that sandhi-derived and underlying Rising tones were treated differently by the parser. However, the experiment was not designed to test this issue, and this result could be accounted for without recourse to incomplete neutralization. For example, the critical syllables were embedded in meaningful two-syllable words, which are sufficient to allow participants to identify the underlying form even if they can't hear the difference between the incompletely neutralized tones. E.g., the form [t͡sʰaiˊˀ t͡ɕʰyˇ˩] corresponds to an existing Low-Low disyllable word 采取 and does not correspond to any existing Rising-Low disyllable word. Thus, a participant does not need to be able to hear the difference between sandhi-derived and underlying tones to be able to recognize that the first syllable in this word has undergone tone sandhi; the participant only needs to recognize the following syllable and to recognize that this combination only corresponds to one existing word, one where the first syllable is underlyingly Low.

Speer and Xu (2008) report two priming experiments and a visual world eye-tracking experiment, specifically designed to investigate the comprehension of ambiguous neutralized forms. Participants heard Mandarin sentences in which the critical Rising-tone word was in a sandhi-licensing context, and thus could be interpreted as either underlyingly Rising or underlyingly Low. For example, one sentence (今天海边__很多) meant something like "Today by the sea there was/were[3] a lot of [qʏˊˀ]." In the Mandarin sentence (which has a different word order than English), the critical syllable [qʏˊˀ] is

[3] Mandarin verbs are not morphologically marked for number, so the verb was the same in either sentence and thus could not be a cue to the identity of the critical ambiguous word.

in a context that licenses tone sandhi. Thus, this word could either be interpreted as meaning "fish" (鱼, the citation form of which is [qy˧˥] with Rising tone), or as meaning "rain" (雨, the citation form of which is [qy˨˩] with low tone, but which would change to [qy˧˥] with Rising tone in this sandhi-licensing context). The sentence stimuli were cross-spliced such that sometimes the critical ambiguous syllable was one produced as a natural Rising tone, with its concomitant acoustic cues, and sometimes the critical ambiguous syllable was one produced with a sandhi-derived Rising tone, with its concomitant acoustic cues. In the visual world eye-tracking experiment, participants heard these sentences while looking at a screen with both potential characters (the one corresponding to the interpretation with Rising tone and the one corresponding to the interpretation with Low tone) shown on the screen, along with distractor characters. An eye tracker monitored where the participants' eyes were looking as they heard the critical ambiguous syllable. Counterintuitively, when participants heard the sentence with a sandhi-derived Rising tone in it, they initially looked at the character corresponding to the Rising tone more than the one corresponding to the Low tone; also counterintuitively, when they heard the sentence with an underlyingly Rising tone in it, they initially looked at the Low-tone character more than the Rising tone character. These results are opposite what one would initially predict (it makes the most sense to assume that if participants can recognize the difference, they would look at the appropriate character more than the inappropriate character). Nonetheless, they are potentially consistent with the notion that participants are subtly sensitive to the difference between the incompletely neutralized tones (although they are also consistent with other possibilities, of course; e.g., they may be a Type 1 error or they may be due to some other unexpected factor).

Liu (2013) tested six listeners' explicit metalinguistic identification and discrimination of underlying and sandhi-derived Rising tones. In the identification task they heard ambiguous disyllables like [t͡ɕʰi˧˥ ma˨˩], in which the first syllable was either underlyingly Rising or was changed to Rising via tone sandhi, and had to select the appropriate orthographic representation (for this example, either 起码 /t͡ɕʰi˨˩ ma˨˩/ "at least" or 骑马 /t͡ɕʰi˧˥ ma˨˩ / "ride a horse"). In the discrimination they heard disyllable pairs, where either both disyllables were the same, or one disyllable was a production with an underlying Rising and one a production with a sandhi-derived Rising tone. They then judged whether they were the same or different. The participants were not significantly more accurate than chance level in identifying the tones, but they were significantly above chance in discriminating them. Above-chance discrimination does not indicate that the tones were incompletely neutralized, though, because there are many other ways that different acoustic cues can allow participants to perform well on within-category discrimination (not just in tone perception and not just in incomplete neutralization contexts). Evidence that the tones are incompletely neutralized should take the form of either above-chance ability to identify the appropriate tones, or better discrimination across tone categories (underlying and sandhi-derived Rising tones) than within either category; above-chance discrimination in of itself is not necessarily indicative of sensitivity to a categorical difference between these tones, as opposed to sensitivity to any other differences between any pair of tokens. (We could make an analogy to dishes from different culinary traditions. For example, if a person eats dishes at a Hunan-style restaurant and a Sichuan-style restaurant, that person might be able to notice the difference between a dish from one restaurant and a dish from the other restaurant, as well as the difference between two dishes from the same restaurant, without necessarily being sensitive to the difference between Hunan cuisine and Sichuan cuisine. Likewise, being able to tell the difference between two stimuli—which may differ in many other respects other than their tones—is not necessarily evidence that participants are sensitive to the general difference between derived and underlying Rising tones.)

Finally, in a recent perception study, Lin and Hsu (2018) presented natural tokens with underlying or sandhi-derived Rising tones to four listeners who were instructed to indicate whether the tone they heard was Rising, Low, or something in between. When hearing a sandhi-derived Rising tone, participants identified it as Rising 53% of the time and as Low 29% of the time. On the other hand, when hearing an underlyingly Rising tone, they identified it as Rising 79% of the time, and as Low 8% of the time. While the reliability of these observations is limited by the small scale of the study, this is another piece of evidence that listeners may perceive derived and underlying Rising tones slightly differently, even if they are not completely accurate in identifying them.

Overall, there is some evidence that listeners may be sensitive to the difference between incompletely neutralized Mandarin tones, but the evidence is weak and spotty, with some of it coming from studies not designed to test this issue and some coming from counterintuitive results that are difficult to explain. Something that is noteworthy is that, to our knowledge, all the studies suggesting that listeners are not sensitive to this difference are based on explicit metalinguistic judgment tasks. On the other hand, some of the potential evidence that listeners are sensitive to the difference comes from online measures like eye movements and reaction times, which participants do not have direct control over and which may reflect processes that participants are not consciously aware of. Because of these conflicting sets of results, we hypothesized that listeners may be able to hear the difference between sandhi-derived and underlying Rising tones at the unconscious, automatic level, but not able to consciously access that for a metalinguistic judgment.

We test this with a visual world eye-tracking experiment, using a design that is essentially a simplification of that used by Speer and Xu (2008). Participants heard ambiguous disyllabic words, without a sentence context, like [duˊ pən˩˧], which might correspond to the word 读本 "reading book", where the citation form of the first syllable has Rising tone, or to the word 赌本 "bookie", where the citation form of the first syllable has Low tone. Like in the design of Speer and Xu (2008), we predict that, if participants are somewhat sensitive to the difference between the incompletely neutralized tones, they should look more at the word with an underlyingly Low first syllable when they hear a token that was spoken as a production of the word with underlying Low tone, compared to when they hear a token that was spoken as a production of the word with underlying Rising tone. By using single words rather syllables embedded in sentences, we aimed to minimize potential complications related to sentence processing and the plausibility or semantic fit between critical words and the rest of the sentence—although, of course, there are still other differences between the critical targets, like their frequencies, stroke counts, and other lexical properties. Crucially, however, in both conditions we are comparing (auditory stimulation with the underlyingly Rising-tone word, and stimulation with the sandhi-derived Rising-tone word), the Low and Rising target words on the screen are the same, so any lexical differences between them would not cause a difference between conditions. If the Low target is looked at more when hearing the sandhi-derived Rising form than the underlyingly Rising form, as far as we know this can only be due to listeners' sensitivity to the phonetic difference between these forms.

2 Methods

All experiment materials, data, de-identified participant demographic information, and analysis scripts are available at https://osf.io/ursh9/. Experiment methods were pre-registered at https://osf.io/35ang/register/5771ca429ad5a1020de2872e.

2.1 Participants

60 native speakers of Mandarin (mean age 23.55, age range 18-34, 48 women and 12 men) participated in the study. Participants were required to filled in the Language Background Questionnaire before the experiment. According to the questionnaire, all participants came from homes where Mandarin was the primary language in use between birth and 5 years old and they listed Mandarin as the most dominant language used in their daily life. They had normal or corrected-to-normal vision and hearing. They provided informed consent to participate and were compensated in cash. All experiment procedures were approved by the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University (project reference # HSEARS20171012002). Two additional volunteers participated in the experiment but their data were not included in the analysis due to technical issues relating to the eye tracker.

2.2 Materials

The experiment consisted of two auditory conditions: a sandhi-derived Rising tone and an underlyingly Rising tone. The critical stimuli consisted of 14 pairs of disyllabic words that were identical in their segmental structure and differed in the underlying tone of the initial syllable. In the sandhi-derived Rising tone condition, these words had an initial Low tone and a final Low tone, which causes the initial low tone to be phonetically realized as a rising tone (i.e.,土改 /tʰuꜿ kaiꜿ/ [tʰuꜿ kaiꜿ]). In the Underlying Rise condition, the initial tone was a Rising tone and the final tone was a Low tone (i.e.,涂改 /tʰuꜿ kaiꜿ/ [tʰuꜿ kaiꜿ]). The words across conditions were maximally similar in terms of frequency, homophone density, neighborhood density, and neighborhood frequency for the word as a unit, as well as for each syllable individually, summarized in Table 1.

A female native speaker of Mandarin Chinese recorded the words in isolation in a sound attenuated room with a Telefunken M-80 dynamic microphone and a Focusrite Scarlett 2i2 sound interface. Three repetitions were recorded at slow, normal, and fast speeds. The words were cut from the main audio recording and no further manipulation was done. Tokens were selected from the normal-speed repetition only.

In the eye-tracking experiment, a critical display presented 4 words in characters. Two of the words formed the target pair, as described above. The other two words served as distracters. The words in the target pairs matched in their segments and differed only in the initial tone (i.e., 涂改 /tʰuꜿ kaiꜿ/ versus 土改 /tʰuꜿ kaiꜿ/). The words in the distracter pairs also matched in their segments and differed only in the initial tone (i.e., 安保 /anꜿ pauꜿ/ - 暗堡 /anꜿ pauꜿ/), but the target and distracter pairs differed from each other in all of their segments.

Given the salience of the sandhi contrast, there was a concern that participants would realize that the focus of the experiment was on sandhi, which may then impact their eye-movements during the tasks. To avoid this issue, several steps were taken to distract away from the prominence of the critical sandhi trials. First, the structure of a single display was carefully controlled. All words in the experiment had a low tone as the second syllable. In every display, each of the four tones was present as an initial syllable once. As such, a single display would have four words present: one High+Low, one Rising+Low, one Low+Low, and one Falling+Low. In this way, it stands to reason that there would occasionally be sandhi items that may be difficult to tell apart, but since every item has a low tone second syllable and each tone appears as the initial tone once in a display, it was hoped that the sandhi items would be considered an coincidence, as opposed to the focus of the experiment. Second, forty-two filler trials were included. The most crucial of these fillers was a set of 14 fillers that were identical to the critical trials in all respects, except that which pairs served as targets and distractors

was reversed, with the sandhi pair now serving as the distractor pair and one of the other words (with High or Falling tone on the first syllable) serving as the target. This filler set is crucial, in that it establishes that a sandhi pair can be present in the display but not targeted. This prevents participants from being able to know *a priori* if the sandhi pair will be targeted without hearing any acoustic input and prevents baseline effects. The listener must wait for segmental information to determine which of the two pairs is being targeted, and then for the tone to determine which item in the pair is being targeted.

The remaining fillers were identical to the critical trials and crucial filler set except that what tone patterns that constituted the pairs of words was switched. While the Rising+Low and Low+Low items matched in their segments in the critical and first filler sets, the Rising+Low and Falling+Low items matched in their segments and the Low+Low and High+Low items matched in their segments.

All pair types served as the targeted pair an equal number of times throughout the experiment. Each of the four tones served as the initial syllable for a target word an equal number of times through the experiment. Each of the four tones were present on the screen as the initial syllable an equal number of times through the experiment.

Additionally, the characters of the second syllables of the targeted pairs were controlled so that for half of the trials the second characters matched (i.e., 夹板 - 甲板) and for half they mismatched even though their pronunciations were the same (i.e., 宣纸- 选址). All critical trials had matching second characters.

Critical items were not repeated across audio conditions. The audio conditions for each item were separated out onto 2 lists in a Latin square design, and participants were randomly assigned to a list. In List 1, participants heard a token that was produced as an underlyingly Rising tone for half of the critical trials and a token that was produced as a sandhi-derived Rising tone for the other half of the critical trials. This pattern was reversed for List 2. The displays were identical across the lists and conditions, only the audio was manipulated.

2.3 Procedure

The experiment was compiled using Experiment Builder software (SR Research). Participants' eye-movements were recorded with a desktop EyeLink 1000 Eye Tracker recording at 1000 Hz (1 gaze position sample recorded every millisecond). The experiment began with a calibration of the participants' pupil and corneal reflection. This calibration was followed by the practice session of 4 trials. After any questions were answered, the experiment began. A trial began with four words appearing on the screen in Times New Roman size 80 font in white on a black background in a non-displayed 2x2 grid. The words remained on the screen for 3,000ms (preview time). This time allowed participants to pre-activate the pronunciations of each of the words and to familiarize themselves with their locations. No auditory stimulus was heard during this presentation. After the 3,000ms preview, the images disappeared, and a fixation cross appeared in the middle of the screen for 500ms to return the participant's gaze to a neutral starting point. As the fixation cross disappeared, the words reappeared on the screen in the same locations as during the preview, and an auditory stimulus was heard though headphones. This auditory stimulus was the target word for that trial, heard in isolation. Participants were instructed to click on the word spoken as quickly as possible. Once the participant clicked, a blank screen appeared for 700ms, after which the next trial began. Both eye-movements (recorded from the target-word onset in the auditory stimulus) and selection accuracy were recorded.

2.4 Data analysis

The data were exported using SR Research DataViewer software in a samples report with an interest period from the word onset until the participant clicked on an item. The samples report provides information on the individual samples of gaze position and does not pre-determine groupings of samples into fixations. The resulting file was pre-processed using a Python script written by SR Research that primarily bins time into 8ms bins, calculates the proportion and count data for each of the four interest areas for each bin, excludes blinks and saccades, and removes several unnecessary columns. This script also excludes fixations that fell outside of the interest areas. The interest areas were set to a 300x200 pixel rectangle around the approximately 150x75 pixel words (The width of this region varied slightly depending on the characters of the word). This allows for slight human or tracker error in the accuracy of the tracking. The interest areas did not overlap and were separated by at least 200 pixels of blank space. The size and arrangement of the interest areas was pre-programmed into the experiment before data collection began.

After this preprocessing, visualization and statistical analysis were carried out using R (R Core Team, 2016). Trials in which the participant did not click on either the Low- or the Rising-tone visual target word were excluded from further analysis. The hypothesis that there were more looks to the Low target when hearing the sandhi-derived than the underlying Rising stimulus was evaluated with a cluster-based permutation test (Maris & Oostenveld, 2007). This test, originally developed for analyzing event-related potential data, controls for multiple comparisons and removes some researcher degrees of freedom (albeit creating others) by eliminating the need for the researcher to choose a particular time window for analysis. Instead, in this test, an uncorrected comparison is done at every sample, clusters of adjacent samples that pass an *a priori* threshold are formed (to address the fact that consecutive samples of data are not statistically independent), and a Monte Carlo permutation test is performed to evaluate the statistical significance of the difference. For our test, we used a paired-sample t-test at each time bin to test the hypothesis that the looks to the Low target were greater when hearing the sandhi-derived Rising stimulus than when hearing the underlying Rising stimulus.[4] Any series of one or more adjacent time bins with uncorrected $p<.1$ in this test was marked as a temporal cluster (we used a relatively liberal clustering threshold based on our *a priori* interest in effects that may be weak but long-lasting, which are best detected with a loose threshold, per Maris & Oostenveld, 2007). The t-statistics in each cluster were summed, and the greatest t-statistic sum across all clusters was saved as the test statistic. Next, 1000 Monte Carlo permutation statistics were drawn. For each permutation, the condition labels were randomly permuted in each participant (since there were only two conditions, this essentially means that for each participant we flipped a coin to randomly determine whether to keep the original pairing between data and condition label or to switch them), then a paired t-test was conducted on each time bin, the t-statistics were summed within each cluster (where a "cluster" is the same bin or series of bins identified as a cluster in the original analysis of the observed data), and the largest of the t-statistic sums was taken as the permutation statistic. The distribution of these 1000 permutation statistics formed the permutation distribution, and the final corrected p-value for the cluster-based permutation test was the proportion

[4] In our pre-registered plan we said that we would logit-transform the data first, but we ended up not doing this because the individual participant data included many zeros, the logit of which is undefined. While our data do not meet the distributional assumptions required for getting a p-value from a t-test, it is important to note that neither these t-tests nor the p-values derived from them were used to draw statistical inferences or conclusions in this study. Rather, they are only used to put time bins into clusters for the purpose of creating a test statistic. The final p-value, and the inferences and conclusions drawn with its help, are based on a non-parametric permutation test, as described by Maris & Oostenveld (2007). Thus, this non-parametric test is still valid, regardless of how the initial sample-wise tests for clustering are performed.

of permutation statistics that were greater than or equal to the original test statistic. This test does not license inferences about exactly when a difference between conditions starts and ends; rather, it licenses an inference about whether there is a difference between conditions overall, anywhere in the curve, after controlling for multiple comparisons.

3 Results

3.1 Behavioural

Overall, participants clicked on the character corresponding to the Low-tone target on 56% of trials when they heard the stimulus with the sandhi-derived Rising tone, and on 52.6% of trials when they heard the stimulus with the underlyingly Rising tone. This is in the direction one might expect if listeners are sensitive to the difference, but this difference was not significant in a logistic mixed-effects model with maximal random effects for participants and items (b=0.28, z=1.15, p=.249). As shown in Figure 1, while some items tended to be preferentially identified with the Low or Rising word target (likely due to differences between the two targets' frequencies or other lexical characteristics), the choice of which word to click was not strongly or reliably influenced by the auditory stimulus.

---Figure 1 about here---

3.2 Eye-tracking

The left side of Figure 2 shows the proportion of looks to the Low target over time, as a function of which auditory stimulus was heard. The right side shows the proportion of looks to the Rising target; while we did not perform statistical analyses on these, since we only pre-registered analysis for the looks to the Low target, we show the data here for completeness. For each condition, the proportion of fixations peaks around 50% rather than reaching 100%; this is not surprising, since the stimuli are ambiguous. Similar to what Speer and Xu (2008) observed, the early portion of the time windows shows a counterintuitive pattern, with participants looking more at the putatively inappropriate target—i.e., looking at the Low tone more when hearing an underlying Rising tone word than when hearing a sandhi-derived Rising tone, and looking at the Rising tone word more when hearing a sandhi-derived Rising tone than when hearing an underlying Rising tone. Crucially, in the latter portion of the window, participants appear to look more at the putatively appropriate target: i.e., for targets corresponding to a word whose first syllable is Low tone, participants look more at these targets when hearing a Rising tone that was derived from a Low tone via tone sandhi, compared to when hearing an underlyingly Rising tone. This trend was not significant, however, in our pre-registered statistical analysis (p=.351).

---Figure 2 about here---

Figure 3 shows the pattern across participants and items. While the effect was statistically significant in the direction we predicted, it is also clear that it varies substantially across participants and also varies somewhat across items. The pattern is much clearer for items, given that each item had observations from as many as 30 participants per condition, whereas each participant had observations from only 7 or fewer items per condition. We performed an additional exploratory cluster-based permutation test, relaxing the cluster threshold to p<.3 (as looser thresholds are more sensitive for detecting weak but long-lasting effects [Maris & Oostenveld, 2007], like the one observed here) and calculating the fixation proportions by item rather than by participant, since the

effects for items were more stable. This yielded a *p*=.098 effect, although *p*-values are not interpretable for this test since it is not confirmatory.

---Figure 3 about here---

Overall, the results are inconclusive, consistent with both the presence and the absence of a difference between conditions. On the one hand, the pre-registered analysis did not yield a statistically significant effect, so we are not able to reject the possibility that the tones are completely neutralized in perception. On the other hand, visual inspection of the data suggest that it may not be reasonable to conclude that they are completely neutralized either (keeping in mind that failure to reject the null hypothesis is not the same as acceptance of the null hypothesis), and exploratory analysis suggests that our pre-registered analysis plan may not have been as sensitive as it could have; i.e., if we replicate the study with more items, and perform confirmatory analysis on a more stable proportion measure and use a more appropriate clustering threshold, then it may be possible to detect a significant effect. None of these facts justify concluding that there is a reliable difference in *this* experiment, as these are all post-hoc decisions about researcher degrees of freedom (e.g., Roettger, ms.; Simmons et al., 2011) and could just reflect capitalizing on noise patterns. At the very least, however, they suggest that this paradigm *may* be capable of providing evidence for incomplete neutralization in the perception of Mandarin Low and Rising tones, and they demonstrate that this issue is worth further investigation and replication.

4 Discussion

We used visual world eye-tracking to examine whether participants are somewhat sensitive to the underlying form of underlying vs. sandhi-derived Rising tones in Mandarin. Using eye-tracking, we examined whether participants were more likely to look at an orthographic form representing an underlying Low tone (e.g., 赌本 "bookie", /du˨ pən˨/) when they heard a token of [duˀˊ pən˨] that was produced as a reading of this underlying form (with the tone of its first syllable changed to Rising tone because of tone sandhi) and less likely to look at this orthographic form when they heard a token of [duˀˊ pən˨] that was produced as a reading a form with Low tone underlyingly (such as 读本 "reading book", /duˀˊ pən˨/). The results of our eye-tracking experiment were inconclusive: there was a numerical trend in the direction we predicted, which was not statistically significant but also not consistent with there being no difference. Statistically speaking, with a result like this we do not have evidence to conclude either that there is a difference or that there is no difference; rather, we can only conclude that the result is indeterminate between these possibilities and that further research with higher power (especially with more items, and perhaps also with more participants) is needed to resolve the issue. At least, however, the results suggest that the present paradigm may be a way to observe partially intact discrimination in the perception of Mandarin tone sandhi, challenging the previously accepted view that this alternation is completely neutralizing in the perception domain.

If perception of Mandarin Low and Rising tones is indeed not completely neutralized by tone sandhi (a speculation which needs further research to confirm), it is worth noting that this clearly is a weak trend. Even if participants are found to reliably identify sandhi-derived Rising tones as underlyingly Low slightly more often than they identify underlyingly Rising tones as Low, this trend is far from 100%; if anything, we expect that further research will show participants to be either at chance or just slightly above chance. This is very different, then, than usual phonological contrasts that speakers can very reliably distinguish. This would put Mandarin third tone sandhi on a similar level as other classical incomplete neutralization cases, such as Dutch final devoicing, where participants also show a slight (better than chance but still small and very imperfect) sensitivity to incompletely neutralized

sounds (e.g., Warner et al., 2004). Thus, this perceptual difference may be of little practical use to interlocutors during real-time language comprehension.

To further examine why the experiment may have failed to observe a significant sensitivity to incompletely neutralized Mandarin tone sandhi, we checked the F0 tracks of the stimuli to confirm whether the speaker actually did incompletely neutralize their tones as is typically reported. Figure 4 shows the pitch tracks for each item. While some items, such as 1, 7, and 12, do indeed show the typical pattern, with underlying Rising tone (solid red line) being higher overall than sandhi-derived Rising tones (dashed blue line), some show little difference (e.g., items 5 and 11) or a difference in the opposite direction (e.g., items 3 and 8). Thus, the incomplete neutralization of Rising and Low tones in production cannot be said to be reliable in this speaker.

---Figure 4 about here---

Did listeners more accurately distinguish Rising from Low tones in those items there the production distinguished them better? This does not appear to be the case, as shown by comparing Figures 4 and 5. The items which showed the largest or most reliable difference in eye-tracking are not necessarily the ones which showed the largest difference in production.

---Figure 5 about here---

Overall, then, we can see that the incomplete neutralization of Low and Rising tones in production is real (i.e., statistical significant and widely replicated) but is not entirely reliable across or within speakers (i.e., it does not happen all the time for every item produced by every speaker), and that perception of these neutralized tones is also not very reliable: either it is completely neutralized, as previous studies have argued, or even if it is not completely neutralized, it is still far less than perfect. In fact, the unreliability of the contrast in production may play a role in why the distinction is not reliably used in perception. We can think of three possible mechanisms for why perception of incompletely neutralized tones may be poor. First, it may be that the acoustic difference between underlying and sandhi-derived Rising tones is imperceptible or difficult to perceive, even when it is present. If that is the case, we expect that perception of these tones will be inaccurate on all items, in experiments such as the present one. Secondly, it is possible that the acoustic difference between the surface forms is indeed perceptible when the surface forms are not completely neutralized, but that this incomplete neutralization does not always occur (as seen in our own stimuli, shown in Figure A). In that case, we would expect that participants are fairly reliable at identifying the tones, but only on the items that show good incomplete neutralization in production; this does not seem very consistent with our findings. Finally, a last possibility is that the difference between incompletely neutralized tones is perceptible when it is present, but that participants know it is not a reliable cue they do not regularly use it to drive their perception. If that is the case, we would expect perception to be inaccurate on all items, the same as in the first possibility outlined above. A potential piece of evidence for this last possibility would be if participants show above-chance accuracy at perceiving these tones in an on-line, implicit measure like eye movements, but not in an explicit metalinguistic measure like end-state judgments; that would be consistent with the notion that participants can perceive the difference somewhat but end up not using it to inform their eventual categorization of the sound. This is indeed a pattern suggested by the present results (in our case both the eye movements and the clicks do show non-significant trends in the direction consistent with incomplete neutralization, but the trend for eye movements seems stronger than that for clicks; although, given that these are very different measures with different properties, it is difficult to directly compare the

sizes of these measures, so future research will be valuable to confirm whether perception of tones really is less neutralized in eye movements than in end-state responses).

A surprising aspect of our results is that, as shown in Figure 2, there was a period in the early time window at the beginning of the syllable where participants showed a trend in the opposite of the expected direction: they looked at the Low target slightly more when hearing underlying Rising tones compared to when hearing sandhi-derived (underlyingly Low) tones. We had not predicted this, and it was not significant in our analysis (in fact it could not have been, as our analysis used one-sided tests). This pattern was also observed, however, by Speer and Xu (2008). Thus far we have no explanation for it. Given that it has shown up in two experiments by now, however, it is highly subjective, and in future research it will be important to examine whether this pattern is replicated, now that there is substantial *a priori* reason to expect it. If future research finds this pattern to be reliable, an explanation will be needed for why this counterintuitive eye movement pattern emerges in tone perception.

One limitation of the study is that, while we used a large number of listeners and items, the experiment only used one speaker. Thus, we cannot be sure that the results will generalize to perception of tokens from other speakers; this is an open question for future research. We do note that using one speaker is currently standard for studies in this area (Peng [2000], Liu [2013], and Zhang & Peng [2013] each used one speaker in their perceptual tasks, and Wang and Li [1967] used two). Nonetheless, acoustic differences between underlying and sandhi-derived Rising tones may not be constant. For instance, the acoustic difference between sandhi-derived and underlyingly Rising tones in real words reported in several studies (Peng, 2000; Liu, 2013; Zhang & Lai, 2010) was not significant in the study by Zhang and Peng (2013); and even when a difference is significant, it is not necessarily present for all speakers (see, e.g., Liu, 2013, and Zhang and Peng, 2013). Therefore, examining the extent to which incompletely neutralized perception generalizes across speakers (if it occurs at all) is a valuable question for future study.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author Contributions

SP and YH conceived the experiment. SP, KC, LP and YH designed the experiment. KC and LP created the stimuli. KC programmed the experiment and LP collected the data. SP and KC analyzed the data. All authors wrote and approved the manuscript.

7 Funding

This research was supported by grant G-UACX from the Department of Chinese and Bilingual Studies to SP.

8 Ethics statement

This study was carried out in accordance with the recommendations of the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

The protocol was approved by the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University.

9 Acknowledgments

The authors would like to thank Jueyao Lin for assistance with stimulus creation.

10 References

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*, e10729.

Chen, S., Wiltshire, C., & Li, B. (2017). Statistical modeling of Mandarin tone sandhi: neutralization of underlying pitch targets. *Proceedings of the World Academy of Science, Engineering, and Technology, 11*.

Cheng, C., Chen, J., & Gubian, M. (2013). Are Mandarin sandhi tone 3 and tone 2 the same or different? The results of functional data analysis. *The 27th Pacific Asia Conference on Language, Information, and Computation*.

Cousineau, D. (2005). Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*, 42-45.

Flege, J., Munro, M., & Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin, and Spanish. *Journal of the Acoustical Society of America, 92*, 128-143.

Kim, H., & Jongman, A. (1996). Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean. *Journal of Phonetics, 24*, 295-312.

Kuo, Y., Xu, Y., & Yip, M. (2007). The phonetics and phonology of apparent cases of iterative tone change in Standard Chinese. In *Phonology and Phonetics: Tones and Tunes, volume 2*, 212-237. Berlin: Mouton de Gruyter.

Lin, Y., & Hsu, Y. (2018). Whether and how do Mandarin sandhied tone 3 and underlying tone 2 differ? *The 32nd Pacific Asia Conference on Language, Information and Computation*.

Liu, X. (2013). 上声变调的声学与感知实验研究 [Acoustic and perceptual research on third tone sandhi]. 文教资料 [*Culture and Education Data*], *29*, 139-142. (In Chinese)

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*, 177-190.

Morey, R. (2008). Confidence intervals from normalized data: a correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64.

Neergard, K., Xu, H., & Huang, C. (2016). Database of Mandarin neighborhood statistics. *Proceedings of the Conference on Language Resources and Education*.

Nicenboim, B., Roettger, T., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics, 70*, 39-55.

Peng, S. (2000). Lexical versus phonological representations of Mandarin sandhi tones. In *Language acquisition and the lexicon: Papers in laboratory phonology V*, Michael Broe and Janet Pierrehumbert [Eds.]. pp. 152-167. Cambridge, UK: Cambridge University Press.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roettger, T. (ms.). Researcher degrees of freedom in phonetic research. https://psyarxiv.com/fp4jr

Simonsohn, U., Nelson, L., & Simmons, J. (2014). P-curve: a key to the file drawer. *Journal of Experimental Psychology: General*, 143, 534-547.

Speer, S., & Xu, L. (2008). Processing lexical tone in third-tone sandhi. Talk presented at *Laboratory Phonology* 11.

Wang, W., & Li, K. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research, 10*, 629-636.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics, 32*, 251-276.

Yuan, J., & Chen, Y. (2014), 3rd tone sandhi in standard Chinese: a corpus approach. *Journal of Chinese Linguistics, 42*, 218-237.

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology, 27*, 153-201.

Zhang, J., Lai, Y., & Sailor, C. (2011). Modeling Taiwanese speakers' knowledge of tone sandhi in reduplication. *Lingua, 121*, 181-206.

Zhang, C., & Peng, G. (2013). Productivity of Mandarin Third Tone sandhi: A wug test. In Peng G. and Shi F. (Eds.) *Eastward Flows the Great River: Festschrift in Honor of Prof. William S-Y. Wang on his 80thBirthday*, pp. 256-282. Hong Kong: City University of Hong Kong Press.

Zhou, X., & Marslen-Wilson, W. (1997). The abstractness of phonological representation in the Chinese mental lexicon. In *Cognitive Processing of Chinese and Related Asian Languages*, Hsuan-Chih Chen [Ed.], 32-27. Hong Kong, Hong Kong: The Chinese University Press.

1 Data availability statement

The datasets generated for this study can be found in the Open Science Foundation repository [https://osf.io/ursh9/].

Tables

Table 1: Summary statistics about lexical properties of the materials. Frequency is listed in natural log of words/characters per million; for words that did not appear as the corpus, we coded them as having 1 occurrence (0.03 per million) to avoid undefined log values. For each cell, the first row shows the measure for the whole word, and the second row shows the measure for the first and second syllable, in parentheses. Frequency, measures are from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010) and homophone/neighbourhood density measures are from the Database of Mandarin Neighborhood Statistics (Neergard, Xu, & Huang, 2016).

	Frequency	Homophone Density	Neighbourhood Density	Neighbourhood Frequency	Stroke Count
Underlyingly Rising	-0.95 (4.49, 6.14)	0.86 (8.43, 3.57)	0.97 (13.93, 13.43)	-0.56 (2.94, 3.32)	16.43 (8.43, 8.07)
Sandhi- derived Rising	-0.17 (5.09, 6.14)	1.00 (5.50, 3.57)	0.93 (15.50, 13.43)	-0.16 (3.96, 3.93)	15.50 (7.50, 8.07)

Figure captions

Figure 1. Proportion of clicks on the visual character target corresponding to the word with Low tone in its first syllable, depending on what audio stimulus was heard. The x-axis shows the proportion of clicks to this target when participants heard an auditory stimulus in which the first syllable's Rising tone was derived through tone sandhi, and the y-axis shows the proportion when the auditory stimulus' first syllable's Rising tone was underlyingly Rising. A diagonal line from the lower left to the upper right corner indicates where the two proportions are equal. The rest of the plot shows a cloud of points, either purple squares or gray circles; each purple square indicates these proportions for one item, and each gray circle indicates these proportions for one participants. Because some participants' points are directly on top of one another, the gray circles are opaque; darker gray circles indicate more participants' data appearing at this same point, and lighter gray circles indicate fewer participants. If the Low target is reliably clicked more after hearing a sandhi-derived Rising stimulus than after hearing an underlyingly Rising stimulus, then most of the observations should be below the diagonal. If most observations are on the diagonal or symmetrically distributed around the diagonal, this indicates that there is little effect of the auditory stimulus.

Figure 2. Proportion of looks to each word target over time. The figure consists of two panels, one representing proportion of looks to the word associated with an underlying Low tone (left) and one representing proportion of looks to the word associated with an underlying Rising tone (right). In each panel, the x-axis represents time since the onset of the auditory stimulus, and the y-axis represents proportion of looks to the given target. A solid red line indicates the proportion of looks to that target when the participant heard a token pronounced with an underlyingly Rising tone, and a dashed blue line indicates the proportion of looks when the participant heard a token pronounced with a Rising tone derived from tone sandhi. Each line is surrounded by an opaque ribbon representing a difference-adjusted by-participant Cousineau-Morey interval (Cousineau, 2005; Morey, 2008), such that where one condition's interval does not include the other condition's mean and vice versa, the conditions might be significantly different; note that these intervals are only an aid for visualization/exploration and will not necessarily correspond to the statistical results reported in the prose, as they are two-sided and do not take multiple comparisons into account. In the left panel (looks to Low tone), from about 250 to 500 ms there appear to be more looks to the Low target when

hearing underlying Rising tones, and from about 750 ms to the end of the window (1500 ms) there are more looks when hearing sandhi-derived Rising tones. In the right panel (looks to Rising target) the reverse of this pattern is seen: initially more looks to the Rising target when hearing a sandhi-derived Rising tone, and then more when hearing an Underlying rising tone.

Figure 3. Incomplete neutralization effect (difference between looks to Low tone when hearing a sandhi-derived Rising tone vs. when hearing an underlying Rising tone) by item (left panel) and by participant (right panel). Each panel shows a red line indicating the mean difference over time, and multiple gray lines indicating the difference for a given participant or item over time. When a line is above zero, this indicates that there were more looks to the Low target when hearing sandhi-derived Rising tone than when hearing underlying Rising tone, at this time. The red line is surrounded by a shaded red ribbon indicating a two-tailed 95% confidence interval based on the t-statistic; this is just a visualization aid and does not necessarily correspond to the statistical results in the prose, as it represents two-sided tests and does not take multiple comparisons into account. In each graph, there is a region around 1000ms where the mean difference is pretty reliably above zero (consistent with this same effect shown in Figure Y and through the statistical analysis). At this time period, 11 out of 14 items are above zero as well and only three are below zero; whereas for participants, a substantial number of participants are below zero.

Figure 4. Pitch tracks for the first syllable of each item, as a function of whether it was pronounced with an underlying Rising tone (solid red lines) or sandhi-derived Rising tone (dashed blue lines). The figure includes fifteen subplots arranged in a 3x5 grid, where the first fourteen subplots each represent the first syllable of one item, and the last shows the legend and axis labels. In each subplot, the horizontal axis represents time in a syllable, and the vertical axis represents fundamental frequency. Each subplot shows two upward-curving lines, representing the fundamental frequencies of underlying and sandhi-derived Rising tones over the course of the syllable. Seven of the fourteen subplots show the red line (representing underlying Rising tone) mostly higher than the dashed blue line (representing sandhi-derived Rising tone), consistent with typically observed incomplete neutralization in third tone sandhi. Three plots show the red line representing underlying Rising tone mostly below the blue line representing sandhi-derived Rising tone, two show the two lines pretty much on top of one another, and two show the two lines crossing each other more or less evenly.

Figure 5. Proportion of looks to the Low target over time, for each item. The plot consists of 15 subplots in a 3x5 grid, as in the previous figure (Figure 4). As in Figure 2, in each subplot shows the x-axis represents time since the onset of the auditory stimulus, and the y-axis represents proportion of looks to the given target. Solid red lines indicate the proportion of looks to that target when the participant heard a token pronounced with an underlyingly Rising tone, and dashed blue lines indicate the proportion of looks when the participant heard a token pronounced with a Rising tone derived from tone sandhi.

Figure 1.TIF

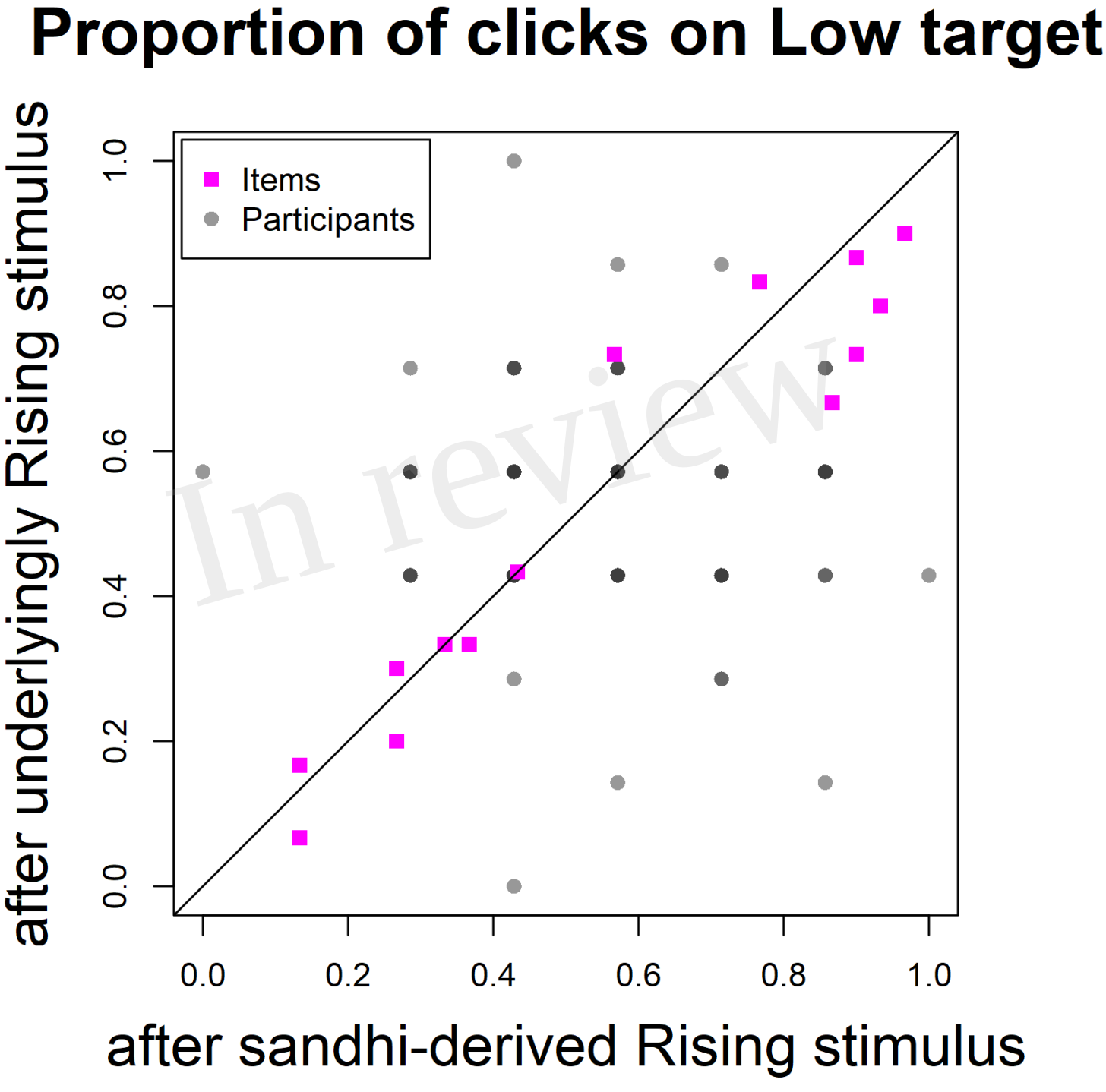
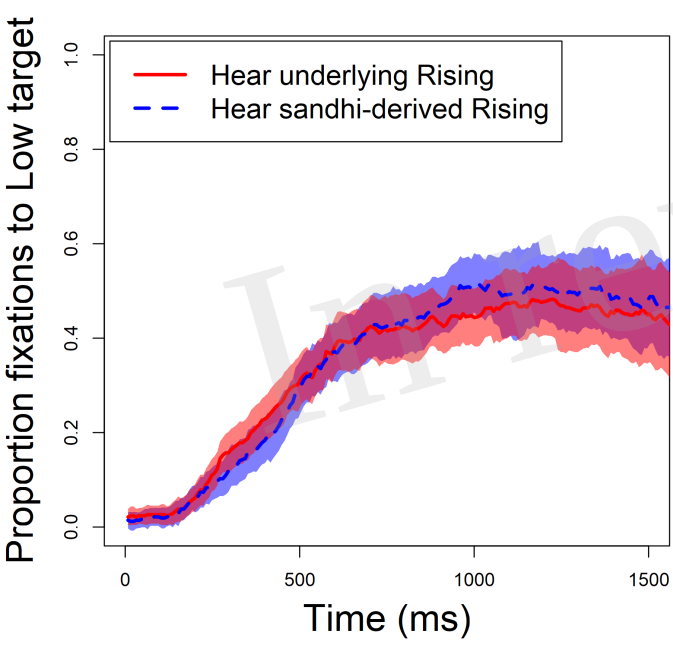


Figure 2.TIF



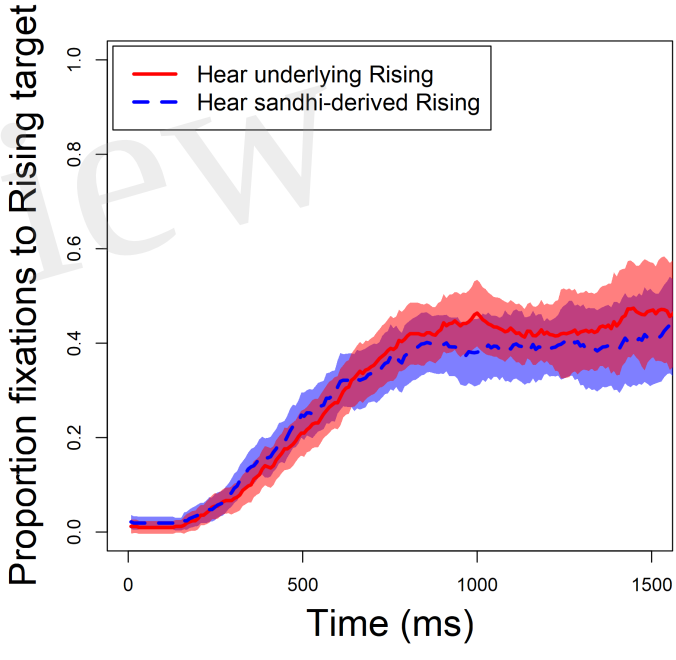


Figure 3.TIF

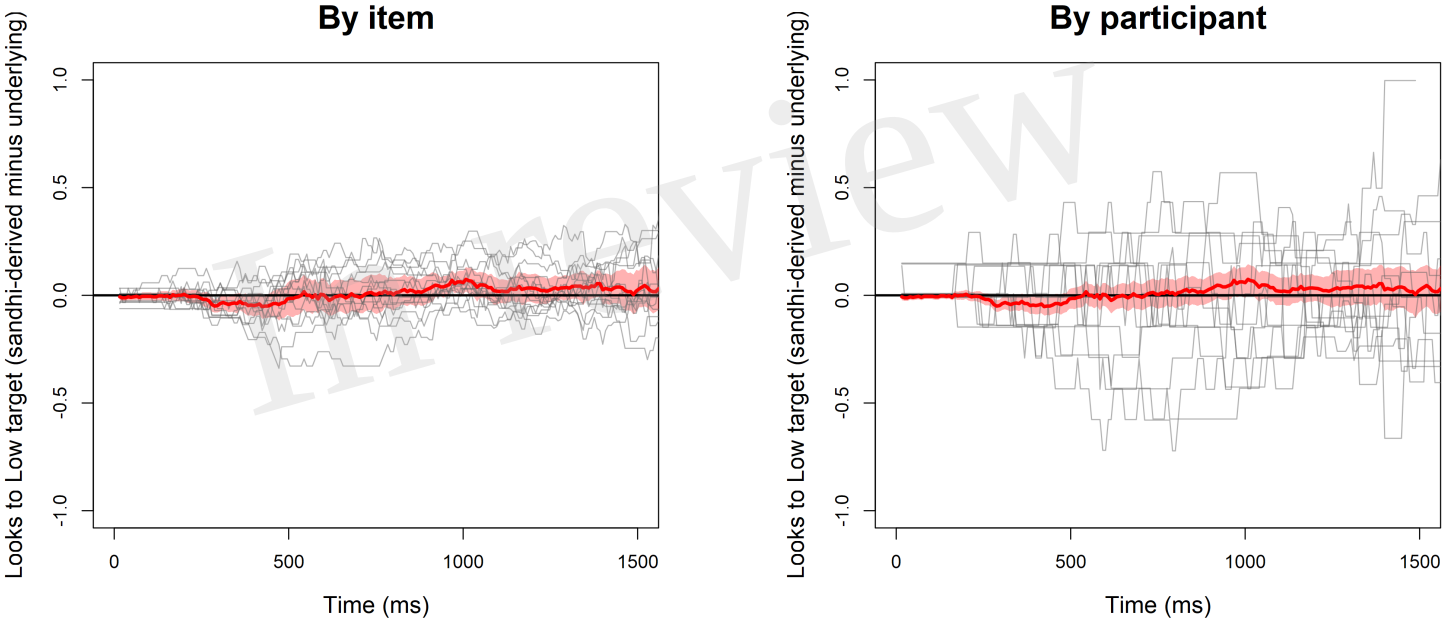


Figure 4.TIF

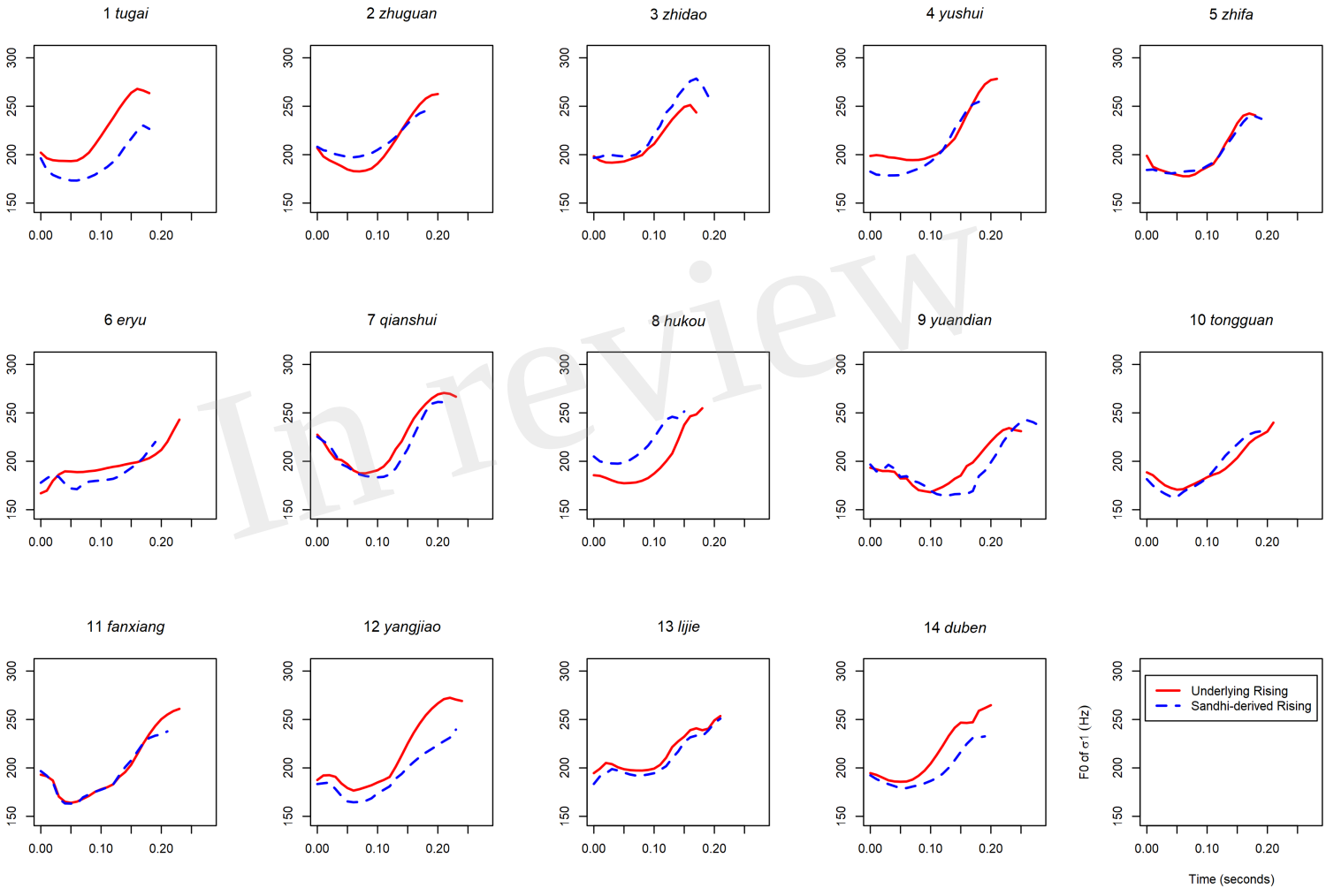


Figure 5.TIF