SPEECH SEGMENTATION AND TONE SANDHI IN MANDARIN CHINESE

Sijia Zhang and Michael Wagner
University of British Columbia and McGill University

1. Introduction

When listening to speech, we effortlessly segment the incoming acoustic stream into words. How exactly we achieve this is still not fully understood. There are various acoustic cues that are important for segmentation, for example, word- or phrase-final lengthening. In addition, there are also phonological patterns that reveal the grouping of syllables into words and phrases. Many languages have so-called sandhi processes, phonological regularities that span word boundaries. Such sandhi processes often only apply between two words when they are 'close' to each other. For example, some sandhi processes seem to be sensitive to whether or not two words are within the same syntactic or prosodic domain. The application or non-application of a sandhi process can therefore provide effective perceptual cues for speech segmentation and sentence parsing, since they allow for an inference that two words are part of a prosodic or syntactic unit. This paper explores and compares the perception of both acoustic and phonological cues on speech segmentation in a tonal language, Mandarin.

1.1 Acoustic cues

A listener has to chunk the incoming signal of speech sounds into words. Duration is well known to be a cue to segmentation in English, since word-final syllables (Oller, 1973) and phrase-final syllables (Klatt, 1975) are lengthened. Less well known, and often not considered, is that intensity tends to decrease across the duration of a word, an effect that does not reduce to utterance-level intensity-downdrift. For example, Wagner (2022) found that intensity is higher in early syllables than in later syllables within nonce words, based on a production experiment in English. Thus, greater duration is a cue for finality within a word due to final lengthening, while higher intensity is a cue for initiality within a word. A similar cue distribution was found for the marking of grouping at the phrasal level in a production study reported in Wagner and McAuliffe (2017, 2019).

Apart from cueing grouping, intensity and duration are well known to play an important role in cueing prominence. Many studies have shown that focus prominence in English is cued both by an increase in duration and intensity (apart from other cues, such as pitch).

Given the dual uses of these cues, if a syllable is longer than expected given the speech rate and its segmental content, this can lead to one of the two inferences, namely, either that the syllable is prominent (*long* → *prominent*) or that the syllable is final (*long* →

Actes du congrès annuel de l'Association canadienne de linguistique 2022.
Proceedings of the 2022 annual conference of the Canadian Linguistic Association.

final). Similarly, an increase in intensity can be interpreted either as a cue to prominence (*loud* → *prominent*), or a cue to initiality within a word or a phrase (*loud* → *initial*).

These inferences based on duration and intensity can rationalize a generalization discovered in Bolton (1894), which today is sometimes called the 'iambic-trochaic law' (Hayes, 1995; Hay and Diehl, 2007), according to which listeners tend to hear iambs when long and short sounds alternate, but trochees when loud and soft sounds alternate. The inferences about grouping and prominence outlined above can explain this pattern, as argued in Wagner et al. (2021) and Wagner (2022): if in a sequence of sounds, every other syllable is longer than expected to an extent that cannot be plausibly attributed just to grouping or prominence, listeners will hear them as both final and prominent, and hence hear iambs. And if in a sequence of sounds, every other sound is louder than can be plausibly attributed just to grouping or prominence, listeners will hear them as both initial and prominent, and hence perceive trochees. This explanation makes sense only if listeners simultaneously parse the signal along two in principle orthogonal dimensions, prominence and grouping, and try to find the best explanation for the signal by positing a particular segmental content, prominence, and grouping structure.

This account was tested in Wagner et al. (2021) in six languages, including Mandarin, based on speech sequences of nonce syllables varying intensity and duration. Duration was observed as cross-linguistically robust cue for prominence (also a cue for grouping in English), and intensity was a cross-linguistically robust cue for grouping (also a cue for prominence in English). In Mandarin, duration was not reliable in cueing grouping. In other words, in Mandarin, longer syllables were not more likely to be perceived as word final. But louder syllables were more likely to be interpreted as initial. Essentially, the finding was that in Mandarin, of the four inferences observed in English, listeners only made three of them: *long* → *prominent*; *loud* → *prominent*; *long* → *initial*. [1]

1.2 Phonological cues

Apart from phonetic cues to phrasing and prominence, phonological patterns can also encode these dimensions. Our interest here is mainly Mandarin third tone (T3) sandhi.

Tones are used in Mandarin to make lexical contrasts, such as in the monosyllabic words *ma-55* (Tone 1) 'mom,' *ma-35* (Tone 2) 'hemp', *ma-214* (Tone 3) 'horse', *ma-51* (Tone 4) 'to scold'. The numbers 1-5 are adopted to denote the target level of tonal realizations as used in Chao (1968). When a word carrying the third tone (214) is followed by another word with the third tone, the first third tone can turn to Tone 2 (35). This process only applies when the two words are syntactically and prosodically close to each other (Shih, 1986; Chen, 2000; Zhang, 2022, and references therein).

There are multiple factors that affect the likelihood of the application of T3 sandhi between two words. T3 sandhi is often reported to be obligatory or near obligatory if two monosyllabic words form a binary constituent. An exception are cases where the two words

[1] Other studies have found in contrast that while duration is a phonetic correlate of Mandarin stress, intensity is not (Lin et al., 1984; Qu, 2013).

are separated by an intonational phrase boundary, for example when the first word is a topic constituent (Liu and Chen, 2020). Tone 3 sandhi is less likely to apply if a T3 is followed by a T3 word that already forms a constituent with a following word, even less likely if the two words carrying T3 are each part of a larger constituent with the word that precedes and follows them, respectively. Another important factor is speech rate (Shih, 1986; Chen, 2000): T3 is more likely to apply, especially over domains of more than two words, with faster speech rates (Shih, 1986; Chen, 2000, and references therein). In very careful speech, T3 can even fail to apply in constituents of two syllables. Perhaps relatedly, speakers do not always apply T3 when producing two novel words with T3 (Zhang and Lai, 2010).

Compositionality may also play a role, such that T3 is more likely to apply if two words have an idiomatic meaning than if their meaning is compositional. In fact, if a two-word sequence does not have a compositional meaning, speakers might just learn the word with the surface T2 tone, instead of positing an underlying T3 and deriving it by tone sandhi (Fu, 2022). This type of "under-learning" problem mostly occurs in compositionally opaque words, such as *ma-214.ji-214* 'ant' where the initial syllable does not contribute to any meaning. However, in compositionally transparent words, such as *ɥy-214.san-214* 'umbrella', underlearning is not observed.

Shih (1986) argues that the generalization about T3 sandhi is best captured by a prosodic approach based on foot formation, which can explain why T3 sandhi is more likely to apply across weaker prosodic boundaries (e.g., across feet or words), than larger prosodic boundaries (e.g., phrase boundaries), although it could be that prosody itself is just another factor affecting whether two T3 are 'close' in the relevant sense, but not the only one. For example, Shih (1986) observes certain cases where T3 sandhi applies across prosodic boundaries, even stronger ones, which seems to go against a purely prosodic account.[2]

The convergence of a variety of different factors including speech rate, syntax, semantic transparency, and prosody is typical of certain types of sandhi phenomena, such as tapping in English. One way to think about what unites these factors is that they all influence the span of production planning. The locality of planning can then potentially explain their effect: if the phonology of two words is planned together, only then, does flapping in English or T3 sandhi in Mandarin get a chance to apply. Under this perspective, the difference in the locality conditions between flapping and T3 sandhi is that Tone sandhi phenomena are only sensitive to the presence of another upcoming word, with no references to its phonological properties, like Xiamen tone sandhi, are not sensitive to prosody or speech rate in the same way (Wagner, 2012). (Shih, 1986; Chen, 2000, and references therein) See Tanner et al. (2017); Kilbourn-Ceron (2017); Kilbourn-Ceron et al. (2020) for experiments exploring the idea that sandhi phenomena are constrained by the locality of production planning.

Whatever the precise account of the locality effect observed in T3 sandhi, the interaction of tone sandhi with constituency, whether mediated by prosody and/or planning

[2] Although Shih (1986) analyzes these cases as effects of cyclic rules application.

locality, makes tone sandhi a potential cue for parsing. Lai and Li (2022), for example, show that T3 sandhi can disambiguate structural syntactic ambiguities.

Some previous studies have also reported a relation between T3 sandhi and prominence. This is in principle not unexpected, given tone sandhi patterns in other languages, for example a production study by Yiu (2019) on Southern Min found higher prominence cues in the citation syllable (the syllable without tone changing) than the sandhi syllable (the syllable that changes its tone). But whether prominence plays a role in T3 sandhi in Mandarin remains controversial. Some authors have presented analyses that assume it does (see, e.g., Qu, 2013). However, Shih (1986) argues at least word stress is irrelevant for T3 sandhi. Shih (1997) discusses some interesting interactions with focus prominence, though. If T3 is indeed influenced by prominence, then manipulating whether or not T3 sandhi has applied might influence a listener's prominence judgments.

1.3 This study

In this study, we look at the perception of sequences of alternating syllables, in order to compare the effect of phonetic and phonological cues on a listener's percept. Our sound sequences consist of repetitions of 'reversible' words, that is sequences of two syllables that have a meaning for both the parse where one is considered the first in a bisyllabic word or last. For example, the word *k^hɤ-214 (Tone 3).k^how-214 (Tone 3)* means 'delicious', but it is reversible in the sense that swapping the syllables also results in a word: *k^how-214.k^hɤ-214*, meaning 'thirsty'. When creating a sequence of alternating syllables based on these two words, and if we obscure how it starts by adding a ramp and noise, the sequence will be ambiguous between being a sequence of repetitions of the word *k^hɤ-214.k^how-214* or of *k^how-214.k^hɤ-214*.

The crucial property of our reversible words is that each word part underlying has T3. While the sequence can in principle be heard as repetitions of one disyllabic compound word or the other, tone sandhi can then help to disambiguate the sequence. If tone sandhi applies to all instances of *k^hɤ-214*, the sequence should tend to be perceived as *k^hɤ-35.k^how-214* 'delicious'. Otherwise tone sandhi would have to apply across the boundary of the bisyllabic word, but not within the word, which should be hard if not impossible. If, however, tone sandhi applies to *k^how-214*, the sequence should tend to be perceived as *k^how-35. k^hɤ-214* 'thirsty'. In fact, the pronunciation of these words with the first syllable undergoing tone sandhi is preferred, even if it is not obligatory.

Tone sandhi, therefore, provides a potentially powerful cue to speech segmentation in these examples, i.e., what we have called here 'grouping'. By looking at sound sequences, we can compare the effect of such a phonological to the effect of acoustic cues, which are prototypically studied in this paradigm.

If the cues in Mandarin work as in English in Wagner (2022), then if one of the syllables (say *k^hɤ-214*) is relatively loud, listeners may interpret this as a cue for prominence, or as a cue to initiality within a word or phrase. If they take it as a sign of initiality, then they will hear the sequence as repetitions of the word *k^hɤ-214.k^how-214*.

If that syllable is relatively long, listeners may interpret this as a cue for prominence or finality within a word or phrase. If they take duration to cue finality, they will hear the sequence as repetitions of the word *k^how-214.k^hɤ-214*. If, by contrast, they attribute loudness or length to prominence for a given syllable, then there will be greater variability in the grouping decisions, that is, in the answer to the question of which word listeners will hear repeated in the sequence.

In our experiment, we will ask listeners both which word they heard and which syllable in the word was prominent. Listeners try to explain the signal by trying to explain its causes. Grouping and prominence are two dimensions of organization of the sounds that each can cause aspects of the stimulus. Decisions about grouping and prominence are mutually informative and constraining, similar to decisions about the size and distance in the visual domain. The prediction is then that the two decisions mutually affect each other, since variability explained by grouping will not have to be explained by prominence, and vice versa.

The decisions about grouping and prominence are also predicted to be mutually informative, as variation in the data already 'explained' by prominence percept will no longer inform grouping decisions and vice versa. This predicts that listeners will take one decision into account when making the other decision. The hypothesis that grouping and prominence perception is mutually informative was tested and confirmed for English in Wagner (2022) and Wagner et al. (2021). The present study explores parsing along two dimensions in more detail in Mandarin.

The first goal of our study is to test whether the cues to prominence and segmentation in sound sequences will be similarly interpreted by Mandarin listeners when the stimuli are actual words rather than nonce syllables. This is particularly important since our prior study (Wagner et al., 2021) used stimuli originally designed for English listeners to study segmentation in Mandarin. The present study will help establish whether these results generalize intuitions of Mandarin speakers based on Mandarin stimuli.

The second goal is to compare the effect of acoustic cues to the effect of phonological cues. The expectation given the phonological pattern as described in the literature is that T3 sandhi will be an effective cue for grouping, but not for prominence.

2. Methodology

2.1 Stimuli

We selected bisyllabic Mandarin words with underlying Tone 3 on each syllable. We chose words that are reversible, in that they would still form an existing word, albeit with a different meaning, when the order of the syllable is swapped. We used four such reversible words, trying to use pairs that have roughly the same lexical frequency, as shown in 1. To avoid the under-learning problem reported in Fu (2022) (see discussion above), the four disyllabic word pairs used in the current design are arguably compositionally transparent.

The first author, who is a native speaker of Mandarin, recorded each of the 8 words

Table 1. The four reversible word pairs used in the experiment. Each syllable in each word carries underlying T3 (214).

Word 1		**Word 2**	
k^{h}ɤ.k^{h}ow	‘delicious’	k^{h}ow.k^{h}ɤ	‘thirsty’
ny.tsɨ	‘woman’	tsɨ.ny	‘offspring’
fa.ɥy	‘French’	ɥy.fa	‘grammar’
jow.xɑw	‘friendly’	xɑw.jow	‘good friend’

(the 2 halves of each reversible pair in Table 1) in two different tonal combinations: T3T3 (i.e. in citation form), T2T3 (i.e. the surface form after sandhi applies). Words in T3T2 (i.e. the surface form corresponding to words with reversed syllable orders) were generated by cutting and rearranging the syllables of the recorded words in T2T3 with reversed syllable orders. This leads to a total of 24 sound files. The third tone was consistently recorded as a falling-rising tone for all the tonal combinations.

We used a script in the speech analysis software Praat (Boersma and Weenink, 1996) to manipulate each of the recorded words using resynthesis in 6 ways, adjusting the intensity and duration of the two syllables. We created a baseline stimulus that scaled each syllable to a mean intensity of 70 dB and scaled each syllable to a duration of 400 msec (referred to as level short/short for duration and soft/soft for intensity).

For duration-varied sequences, we kept the mean intensity constant at 70 dB, and increased the duration of either syllable to 550 msec (yielding the conditions long-short, short-long). We also created a long-long sequence with each syllable at 550 msec.

For intensity-varied sequences, we kept the duration constant at 400 msec, and increased the intensity of one of the syllables by 7 dB (yielding conditions loud-soft, soft-loud).

We then created ambiguous sequences of repetitions of each word by alternating the two syllables, varying whether they were pronounced with sandhi (e.g. *k^{h}ɤ-35.k^{h}ow-214.k^{h}ɤ-35.k^{h}ow-214...*, or *k^{h}ɤ-214.k^{h}ow-35.k^{h}ɤ-214.k^{h}ow-35...*; the latter sequence corresponds to sandhi applied in the word with a reversed syllable order), or without (e.g. *k^{h}ɤ-214.k^{h}ow-214.k^{h}ɤ-214.k^{h}ow-214...*). This led to 144 different sound sequences, one based on each of the 8 (original words) * 3 (tone realizations) * 6 (acoustic manipulation conditions) = 144 words, by repeating a given word 12 times. The syllables in the sequence were equally spaced without pauses. Each recording lasted 12s. The beginning of the sequence was faded in by filtering with the first half of a cosine function over a window of 5s, and over the same time window white noise was faded out. This manipulation was intended to decrease the effects of the underlying syllable order.

2.2 Participants

Twenty native speakers of Mandarin took part in the experiment remotely using their own laptops or mobile devices. They were either students at McGill University in Montreal, Canada, or university students in China. Before the experiment, participants filled out a language questionnaire. The experiment was followed by a music questionnaire and a post-experiment questionnaire. We do not report on the details of these questionnaires here for reasons of space, but all of the participants were L2 learners of English, and some had knowledge of other languages, such as French.

2.3 Procedure

We used the Prosodylab Experimenter (Wagner, 2021), a set of javascripts making use of the jsPsych toolbox (De Leeuw, 2015) to generate online experiments. The Prosodylab Experimenter generates an experiment based on a tab-separated spreadsheet, along with some others files such as the instructions to the participant. Participants were tested remotely and used their own desktop, laptop or mobile devices and internet to access the experiment webpage at their homes. Each participant saw 72 out of the 144 stimuli. We selected stimuli for a given participant in a Latin Square design, such that for a given word pair, each participant encountered an equal number of sequences generated from the 2 syllable orders. Trials were presented in random order. Participants were instructed that one or the other syllable of a word may be 'emphasized', for example in a corrective realization. All instructions and questionnaires were presented in Mandarin Chinese. In order to ensure that participants were using headphones, as instructed, we conducted a headphone screener test (Woods et al., 2017).

On a given trial, participants were asked to listen carefully to the recordings of Mandarin word sequences. They were first asked a question about grouping, a binary forced-choice question about which word of the two possibilities they heard repeated in the sequence. The choices were presented with two clickable buttons labelled with the words in Chinese characters. They were then asked to rate the sequence as to how natural the sequence sounded, on a sliding scale. Finally, they were asked which of the two syllables they heard as emphasized, the two options were presented in the pinyin form of the two syllables. To avoid confusion, the options of the binary forced-choice questions were randomized between participants, but kept constant within participants.

While the participants chose between the actual words for the grouping question, and chose a particular syllable (e.g. *k^hɤ*) for the prominence task, our research questions are not about specific syllables, but about whether a syllable that is loud, long and/or has undergone tone sandhi is heard as prominent and group-initial. We therefore coded the syllables as [syllable1 syllable2] depending on which came linearly first in the sequence. So if a sequence was based on repetitions *k^hɤ-214.k^how-214*, the syllable *k^hɤ* would be coded as syllable1. If a sequence was based on repetitions of *k^how-214. k^hɤ-214*, the syllable *k^hɤ* would be coded as syllable2.

In the labeling of the conditions, loud-soft means e.g. that syllable1 was loud and syllable2 was soft. In the labeling of tones, *23* means that syllable1 had tone 2 and syllable2 and tone 3. In the labeling of the responses, we will say that syllable1 was heard as initial or prominent, rather than reporting the specific syllable. Of course, that syllable1 for a given sequence may have been any of the 8 syllables, e.g. it may have been $k^h ɤ$ or $k^h ow$, depending on the underlying order of these two syllables in the sequence.[3] The underlying order was counterbalanced within each participant.

3. Results

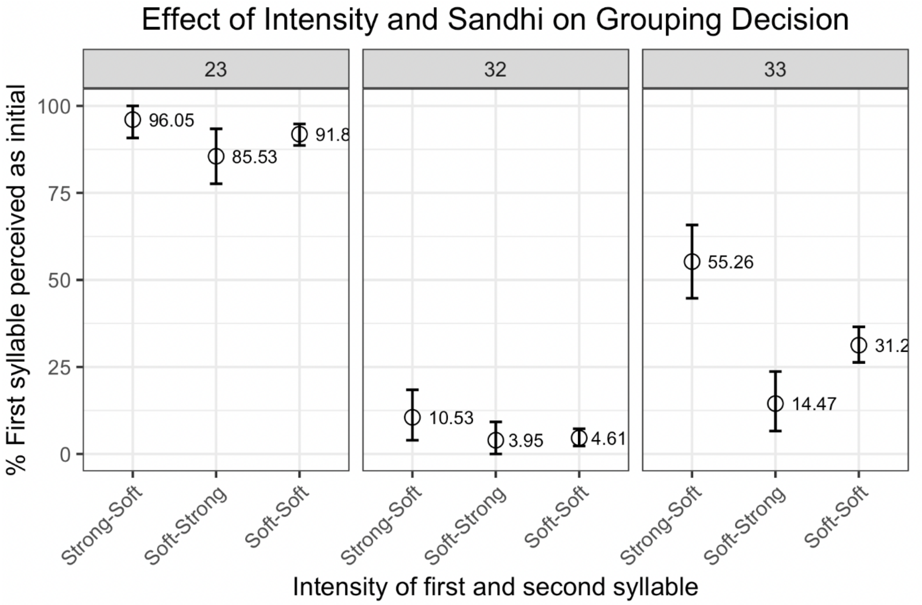

Figure 1. Perceived grouping depending on tone sandhi and intensity.

Figure 1 and Figure 2 summarize the results of the grouping question (*What word did you hear?*). When tone sandhi occurred on one of the syllables (i.e. tone conditions *23* and *32*), this cue dominated the response. Listeners heard the word with T2 on the first syllable, and T3 on the second, as expected. In these cases, intensity and duration only had a small influence on the grouping decision. Intensity and duration influenced the decision in the predicted direction when they did, most clearly visible in *33* cases. Intensity was interpreted as a cue for initiality, as shown in Figure 1. Duration correlated with finality, although to a much smaller degree, as presented in Figure 2.

Table 2 summarizes ME logistic regression models for each decision, which we fit using the package *lme4* (Bates et al., 2014). Categorical predictors were Helmert-coded, so that contrasts were orthogonal. All predicted variables were converted to z-scores, so that effect sizes can be meaningfully compared. The model for the grouping decision shows that the main effects of intensity (Loud.vs.Soft, $\beta = 0.91$, $p < 0.001$) and tone sandhi (Tone23.vs.32, $\beta = 6.38$, $p < 0.001$) were significant, but the effect of duration was not (Long.vs.Short, β = -0.22). The effect of tone sandhi was almost an order of magnitude

[3] The full set of stimuli is posted in the OSF project associated with this study, at https://osf.io/u84br/.

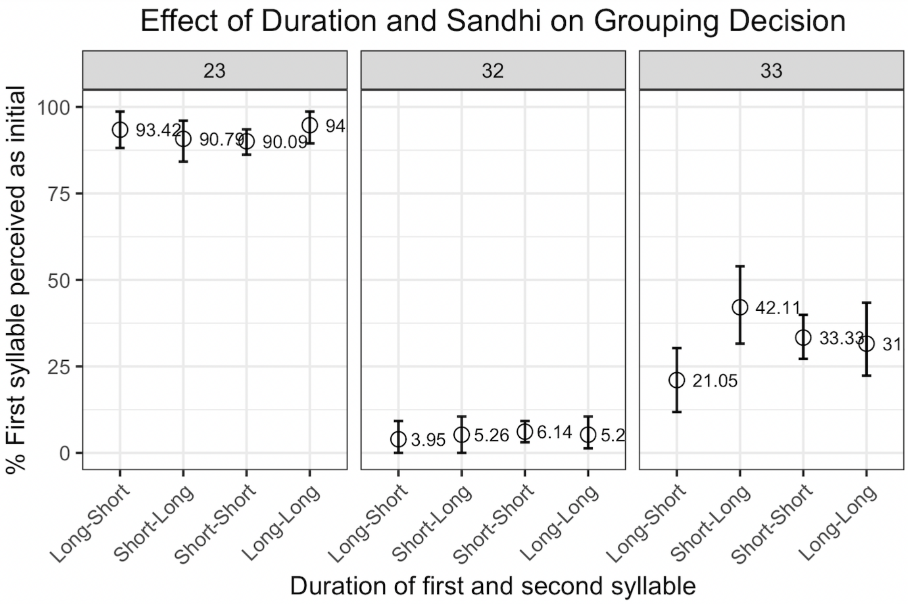

Figure 2. Perceived grouping depending on tone sandhi and duration.

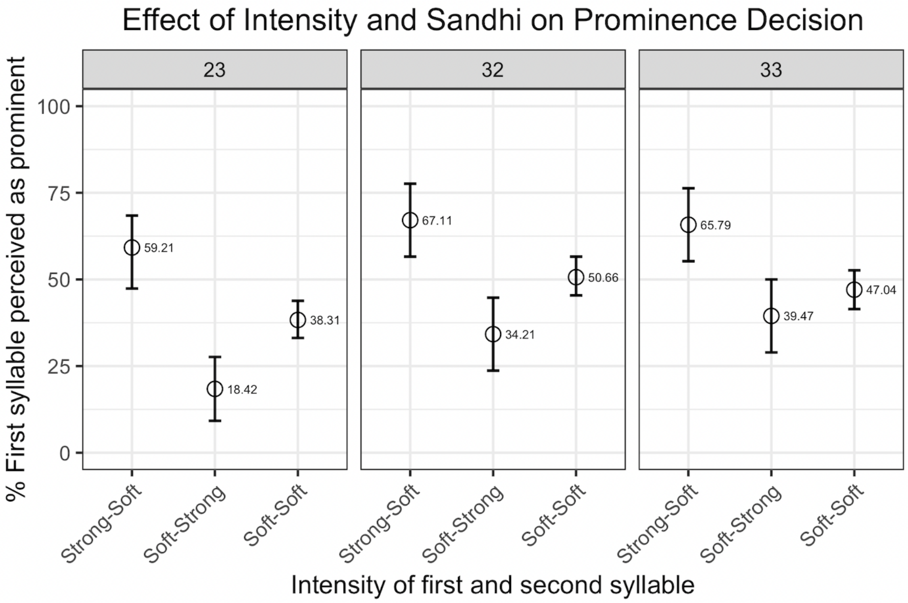

Figure 3. Perceived prominence depending on tone sandhi and intensity.

bigger than the effect of intensity. Interestingly, the interactions between the presence and absence of tone sandhi (Different.vs.SameTone) and the effect of duration/intensity was not significant, despite the clear pattern in the figure.

There was also a significant effect of the prominence decision on the grouping decision ($\beta = 0.61,\ p < 0.01$). It is possible that the influence of the prominence decision

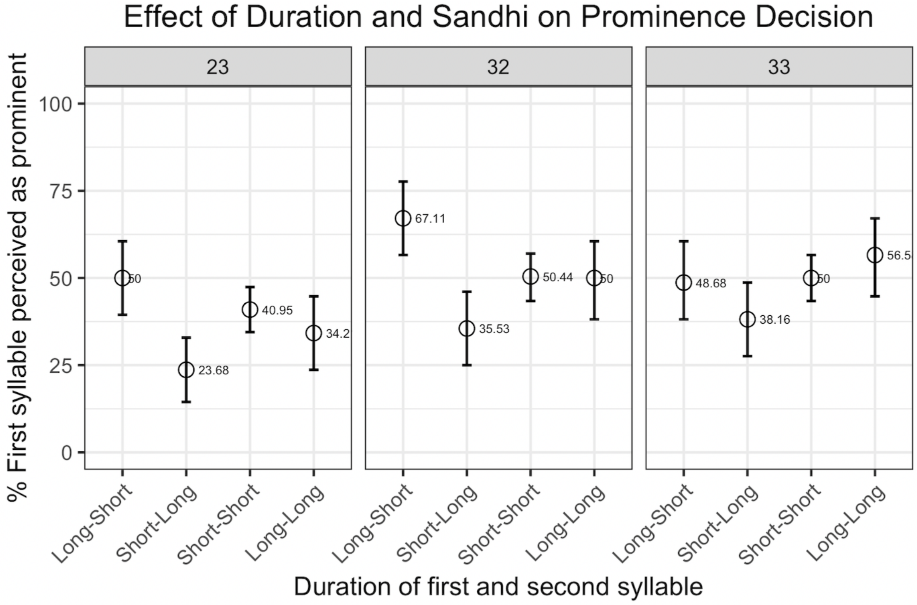

Figure 4. Perceived prominence depending on tone sandhi and duration.

	Grouping	Prominence
(Intercept)	$0.48(0.15)^{**}$	$0.21(0.10)^{*}$
ProminenceDecision	$0.61(0.19)^{**}$	
Tone23.vs.32	$6.38(0.94)^{***}$	$-0.96(0.55)$
Different.vs.SameTone	$0.57(0.22)^{*}$	$-0.35(0.13)^{**}$
Loud.vs.Soft	$0.91(0.21)^{***}$	$0.96(0.13)^{***}$
Different.vs.SoftSoft	$0.21(0.26)$	$0.00(0.18)$
Long.vs.Short	$-0.22(0.21)$	$0.75(0.13)^{***}$
Different.vs.ShortShort	$0.15(0.24)$	$-0.15(0.17)$
Other.vs.LongLong	$-0.14(0.24)$	$-0.12(0.17)$
Different.vs.SameTone:Loud.vs.Soft	$0.28(0.38)$	$-0.68(0.26)^{**}$
Different.vs.SameTone:Long.vs.Short	$-0.64(0.36)$	$-0.62(0.26)^{*}$
GroupingDecision		$0.61(0.19)^{**}$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2. Logistic ME models for the grouping and the prominence decision.

accounts for the apparent interaction between tone sandhi and duration in explaining the grouping decision.

Figure 3 and Figure 4 summarize the results for the prominence question. Both intensity and duration are used as robust cues to prominence regardless of the presence of T3 sandhi. Tone sandhi did not have clear overall effects on the prominence decision, although the figure suggests that the absence of tone sandhi amplified the effect of duration and intensity. This makes sense if tone sandhi determines the grouping, and thereby has the effect that all additional variation in duration/intensity will be attributed to prominence.

The prominence model in Table 2 shows that intensity ($\beta = 0.96$, $p < 0.001$) and duration ($\beta = 0.75$, $p < 0.001$) were significant predictors for the prominence decision. The effects of intensity/duration interacted with the presence/absence of tone sandhi (Different.vs.SameTone), such that both effects were significantly amplified by the absence of tone sandhi. The grouping decision was also a significant predictor for the prominence decision.

Overall, tone sandhi (the difference between *23* and *32*) did not significantly affect prominence, except that the presence/absence of tone sandhi (Different.vs.SameTone) had a significant main effect on the prominence decision.

For reasons of space, we do not report on the result of the naturalness rating. The biggest effect of the naturalness ratings is that listeners preferred sequences in which one of the two syllables had undergone tone sandhi, confirming that in the examples used here, tone sandhi application is preferred.

4. Discussion and Conclusion

The results show that both acoustic and phonological are important in parsing the signal. Tone sandhi seems like a much more powerful cue for grouping than gradient manipulations of duration and intensity. This is not too surprising, given that we looked at cases in which the application of T3 sandhi is nearly obligatory, other than in very slow or careful speech. However, it is still interesting that this study shows how the effect of acoustic cues can be compared with the effect of phonological cues in this type of experimental paradigm.

The results suggest several similarities, but also some differences, in how intensity and duration contribute to speech segmentation in Mandarin compared to English. Intensity is a significant cue to initiality, just as in English (Wagner, 2022). Duration, however, was not overall a good cue for grouping the signal into words in Mandarin, in contrast to English. When it comes to the prominence decision, however, both duration and intensity are significant predictors in Mandarin just, as was reported in earlier studies in English. The present study thus replicates the pattern observed for Mandarin listeners based on nonce words stimuli designed for English in Wagner et al. (2021). The present results show that these results generalize to sequences based on actual Mandarin words, thereby validating the method of using nonce words to explore speech segmentation.

The presence/absence of sandhi had one important effect on the interpretation of acoustic cues with respect to prominence. In the presence of tone sandhi, which effectively settles the grouping percept single handedly, all additional acoustic variation was attributed to prominence, leading to greater acoustic effects on the prominence decision when T3 sandhi had applied. This makes sense if grouping and prominence in general compete with each other for explaining the properties of the stimulus, and listeners attribute acoustic effects to prominence if the grouping decision has already been decided by the much stronger grouping cue of tone sandhi.

In those cases where T3 sandhi applied, the choice of syllables that underwent tone sandhi, however, had no influence on the prominence percept. Given the results in Yiu (2019) on tone sandhi in Southern Min, and given earlier analyses relating T3 sandhi with prominence, one might have expected that syllables with T3, the citation form, might be more likely perceived as prominent than the sandhi syllable with T2, but this turned out not to be the case, favoring models that account for T3 sandhi as a cue to phrasing/grouping.

Overall, listeners try to explain the signal based on different possible causes, as predicted by accounts of perception in terms of auditory scene analysis (Bregman, 1994). Making a grouping decision will inform the prominence decision and vice versa. Listeners attribute aspects of the signal to two in principle orthogonal hierarchical levels of organization of the same acoustic events, the dimension of prominence and the dimension of grouping.

Perceptual decisions about grouping and prominence inform and mutually constrain each other since they account for overlapping sensual information. The perceptual decisions are hence similar to decisions in the visual domain that are mutually informative and constraining, such as the decisions about the size and distance of an object, or the decisions about the hue of the background light and the color of an object. This mutual influencing of the grouping and prominence decisions is not just evidenced by the significant interactions with the acoustic effects we observe, but also by the significant main effects that each decision has on the other.

While both this study and Wagner et al. (2021) found that in Mandarin, duration did not influence the grouping decision, it would be premature to conclude that duration does not cue grouping in Mandarin. It may just be that, for reasons yet to be understood, excess duration is attributed for the stimuli here and Wagner et al. (2021) a different cause, maybe segmental or tonal content. More studies will be necessary to fully explore the full range of cues and their full range of interpretations when parsing speech.

5. Acknowledgements

This research was funded by NSERC Grant RGPIN-2018-06153: *Three dimensions of sentence prosody* to the second author. Many thanks to Jiayi Huang and Yijie Yan, who together with the first author were involved in running a pilot study of the present one as part of a class project in LING 450 Laboratory Linguistics in Fall 2019, taught by the second author. Thanks for all the audience and their comments and questions at CLA 2022, as well as the feedback from UBC Speech in Context Lab. More information about this project, including the full set of stimuli, is available on OSF at https://osf.io/u84br/.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Boersma, Paul, and David Weenink. 1996. Praat: a system for doing phonetics by computer. Report 132. Institute of Phonetic Sciences of the University of Amsterdam.

Bolton, Thaddeus L. 1894. Rhythm. *The American journal of psychology* 6(2): 145–238.

Bregman, Albert S. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Chao, Yuen Ren. 1968. *Language and symbolic systems*, vol. 260. Cambridge University Press Cambridge.

Chen, Matthew Y. 2000. *Tone sandhi: Patterns across chinese dialects*, vol. 92. Cambridge University Press.

De Leeuw, Joshua R. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods* 47(1): 1–12.

Fu, Boer. 2022. UR underlearning of mandarin tone 3 sandhi words. In *The 58th annual meeting of the chicago linguistics society*.

Hay, Jessica, and Randy L Diehl. 2007. Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception & psychophysics* 69(1): 113–122.

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Kilbourn-Ceron, Oriana. 2017. Speech production planning affects variation in external sandhi. Doctoral dissertation, McGill University.

Kilbourn-Ceron, Oriana, Meghan Clayards, and Michael Wagner. 2020. Predictability modulates pronunciation variants through speech planning effects: A case study on coronal stop realizations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1).

Klatt, Dennis H. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 3: 129–140.

Lai, Wei, and Aini Li. 2022. Integrating phonological and phonetic aspects of mandarin tone 3 sandhi in auditory sentence disambiguation. *Laboratory Phonology* 13(1).

Lin, M-C., J-Z. Yan, and G-H. Sun. 1984. Beijinghua liangzizu zhengchang zhongyin de chubu shiyan [a preliminary experiment on normal stress in beijing mandarin bisyllabic expressions]. *Fangyan* 1: 57–73.

Liu, Chin-Ting, and Li-mei Chen. 2020. Testing the applicability of third tone sandhi at the intonation boundary. *Language and Linguistics* 21(4): 636.

Oller, D Kimbrough. 1973. The effect of position in utterance on speech segment duration in english. *The journal of the Acoustical Society of America* 54(5): 1235–1247.

Qu, Chen. 2013. Representation and acquisition of the tonal system of mandarin chinese. Doctoral dissertation, McGill University.

Shih, Chi-lin. 1986. *The prosodic domain of tone sandhi in chinese*. University of California, San Diego.

Shih, Chilin. 1997. Mandarin third tone sandhi and prosodic structure. In *Studies in chinese phonology*, ed. Jialing Wang and Norval Smith, *Linguistics Models*, vol. 20, 81–124. De Gruyter.

Tanner, James, Morgan Sonderegger, and Michael Wagner. 2017. Production planning and coronal stop deletion in spontaneous speech. *Laboratory Phonology* 8 (1): 15: 1–39.

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

Wagner, Michael. 2021. Prosodylab experimenter. Retrieved from Github in spring 2021.

Wagner, Michael. 2022. Two-dimensional parsing explains the iambic-trochaic law. *Psychological*

Review 129(2): 268–288.

Wagner, Michael, Alvaro Iturralde Zurita, and Sijia Zhang. 2021. Parsing speech for grouping and prominence, and the typology of rhythm. In *Proceedings of interspeech*, 2656–2660.

Wagner, Michael, and Michael McAuliffe. 2017. Three dimensions of sentence prosody and their (non-) interactions. In *Proceedings of interspeech*, 3196–3200.

Wagner, Michael, and Michael McAuliffe. 2019. The effect of focus prominence on phrasing. *Journal of Phonetics* 77: 1–26.

Woods, Kevin JP, Max H Siegel, James Traer, and Josh H McDermott. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* 79(7): 2064–2072.

Yiu, Suki. 2019. Iambic and trochaic rhythm in jieyang (southern min). In *Proceedings of the annual meetings on phonology*, vol. 6.

Zhang, Jie. 2022. Tonal processes defined as tone sandhi. In *The Cambridge Handbook of Chinese Linguistics*, ed. Chu-Ren Huang, Yen-Hwei Lin, I-Hsuan Chen, and Yu-YinEditors Hsu, Cambridge Handbooks in Language and Linguistics, 291–312. Cambridge University Press. doi: 10.1017/9781108329019.017.

Zhang, Jie, and Yuwen Lai. 2010. Testing the role of phonetic knowledge in mandarin tone sandhi. *Phonology* 27(1): 153–201.