REGULAR ARTICLE

Phonological competition in Mandarin spoken word recognition

Qing Yang* and Yiya Chen*

*Centre for Linguistics, Leiden University, Leiden, The Netherlands; †Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands

ABSTRACT

Most of the world’s languages use both segment and lexical tone to distinguish word meanings. However, the few studies on spoken word recognition in tone languages show conflicting results concerning the relative contribution of (sub-)syllabic constituents, and the time course of how segmental and tonal information is utilised. In Experiments 1 & 2, participants listened to monosyllabic Mandarin words with the presence of a phonological competitor, which overlaps either segmental syllable, onset and tone, rhyme and tone, or just tone. Eye movement results only confirmed the segmental syllable competition effect. Experiment 3 investigated the time course of segmental vs. tonal cue utilisation by manipulating their point of divergence (POD) and found that POD modulates the look trajectories of both segmental and tonal phonological competitors. While listeners can use both segmental and tonal information incrementally to constrain lexical activation, segmental syllable plays an advantageous role in Mandarin spoken word recognition.

ARTICLE HISTORY

Received 20 June 2021
Accepted 21 December 2021

KEYWORDS

Mandarin spoken word recognition; visual world paradigm; phonological competition; Mandarin Chinese; models of lexical tone processing

The majority of the world’s languages are tonal, in which pitch variation, known as lexical tone, distinguishes word meanings (Yip, 2002). For example, in Mandarin Chinese, the same segmental syllable ma means “mother” with a high-level tone but “horse” with a low (dipping) tone. Thus, it is expected that speakers of tonal languages such as Mandarin Chinese need to utilise tonal information effectively for successful and efficient spoken word recognition. Despite the importance of tone in the lexicon of the majority of the world’s languages, existing models of spoken word recognition (SWR) have only begun to investigate the role of lexical tone. Understanding lexical processing in tonal languages would provide insights into the potential universal and diverse patterns of SWR across languages of the world and benefit the development of existing SWR models, which have based mainly on data from Indo-European non-tonal languages (e.g. Gareth Gaskell & Marslen-Wilson, 2002; Luce & Pisoni, 1998; Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994).

One broad consensus in current models of SWR is that the process of recognising a word is incremental. Listeners activate several possible word candidates as the incoming speech signal unfolds. Sub-lexical phonemic features influence online lexical processing (e.g. Dahm et al., 2001; McMurray et al., 2008; Salverda et al., 2003). There is some evidence from tonal languages, mainly limited to Mandarin Chinese (cf. Burnham et al., 2011 for Tai tones), that suggests incremental activation and competition of sub-lexical phonologically similar word candidates (e.g. Lee, 2007; Liu & Samuel, 2007; Malins & Joanisse, 2010; Sereno & Lee, 2010; Zhao et al., 2011). Despite possible similarities of SWR processes across tonal and non-tonal typologically different languages, several issues, as detailed out below, have remained outstanding and need to be clarified for SWR in tonal languages. Briefly speaking, in tonal languages, it is commonly recognised that segmental and suprasegmental tonal information both play a role during SWR (e.g. Malins & Joanisse, 2010, 2012a; Zhao et al., 2011). What has remained open is how exactly segmental and tonal cues are taken up and processed during SWR. At the segmental level, the non-tonal syllable seems to play a critical role as a functional unit of processing (e.g. Sereno & Lee, 2015; Zhao et al., 2011). Relatedly, overlaps in sub-syllabic constitutes (segmental syllable onset and rhyme) have been found to exert no influence on Mandarin lexical competition (see Malins & Joanisse, 2010 for a null rhyme competition effect; see Zou, 2017 for a null onset competition effect). Given the different experimental paradigms/designs and their different levels of
sensitivity to the time course of speech processing, it remains debatable whether segmental syllables are processed incrementally or holistically. The present study aimed to employ the eye-tracking technique to address the following issues by seeking answers to three specific research questions: (1) Do segmental syllables have a special status in Mandarin lexical processing? (2) What are the relative contributions of sub-syllabic segmental constituents (such as onsets and rhyme) and suprasegmental lexical tone? (3) What is the time course of segmental and suprasegmental processing and cue utilisation during online lexical processing?

**Issue one: the role of segmental syllable in spoken word recognition**

One issue to be resolved is the role of the non-tonal segmental syllable as a primary and holistic processing unit during SWR. Thus far, Mandarin has served as the main empirical base in the extant literature. It is well-known that Mandarin syllables differ from syllables of Indo-European languages in several aspects. First, Mandarin syllables consist of both segmental and suprasegmental information, i.e. lexical tone. The segmental syllables in Mandarin are simple in structure and have a relatively small number of syllable types. For example, they do not have complex consonant, and only two nasal consonants (/n/ and /ŋ/) are allowed as codas. The total number of syllables is also rather limited (i.e. about 1,200 tonal syllables and 400 segmental syllables). Second, most morphemes in Mandarin are monosyllabic (i.e. segmental syllable plus tone), rendering syllables as a unit of meaning. Last but not least, the writing system in Mandarin is based on syllable-sized characters, reinforcing the notion of syllable as a holistic unit. These unique properties have motivated researchers to entertain the idea that Mandarin syllables serve as an ideal lexical processing unit.

The evidence on the role of syllable and sub-syllabic units in SWR, however, has been mixed. Zhao et al. (2011) proposed that Mandarin SWR is “syllable-based holistic processing rather than phonemic segment-based processing”. In Zhao et al. (2011), Mandarin speakers made semantic judgments on pictures while listening to an auditory distractor word. Event-related potentials (ERPs) showed that when the distractor mismatched the name of the picture in either onset, rhyme, tone, or the whole syllable (see Table 1 for sample stimuli used in the study), N400 (a negative ERP component elicited with semantic or phonological violations of expectations; Kutas & Hillyard, 1984) was elicited. Crucially, the earliest and highest amplitude was elicited by the whole syllable (i.e. segmental and tonal) violation. Sereno and Lee (2015) reached a similar conclusion with two auditory lexical decision tasks. In their study, participants’ responses were only facilitated when the primes had identical segmental syllables or syllables to targets, and no priming effect was found for those with only partial segmental overlaps (i.e. onset and tone overlaps; rhyme and tone overlaps).

Counter evidence against segmental syllable as the basic unit of processing in SWR has also been reported. With EEG recording, Malins and Joanisse (2012a) asked participants to make judgments on whether an auditory word and a simultaneously presented picture match or not. The picture names overlapped with the auditory words in either segmental syllable, onset, rhyme, tone, or unrelated (see Table 1 for sample stimuli). Results showed that all conditions modulated the phonological mapping negativity effects (PMN; an ERP component associated with pre-lexical processing; Connolly & Phillips, 1994; Newman & Connolly, 2009) and N400 effects (associated with lexical word meaning processing; Kutas & Hillyard, 1984). Moreover, the PMN effects did not differ between the rhyme, tone, and the unrelated condition, suggesting that neither syllable nor segmental syllable in Mandarin merits any special status as a holistic processing unit. More recently, Ho et al. (2019) investigate the role of syllable in Mandarin word processing in sentence context with the cross-modal priming paradigm. In this task, prime words were embedded in the middle of a visually presented sentence, while target words were embedded in a following aural sentence. The targets and primes were mismatched in onset, tone, or syllable (see Table 1 for sample stimuli). Compared with identical sentences, all three critical mismatching conditions modulated PMN and N400 components. Crucially, the smallest amplitudes for PMN and N400 components were elicited by the whole syllable mismatching condition. This was interpreted as due to the lack of phonological competition between target and prime by Ho et al., who further suggested that Mandarin listeners process spoken words segment by segment rather than by the whole syllable.

It is clear that the above studies differed in the experimental paradigms employed, the specific behavioural and neural measurements taken, and the exact conditions compared. More research on the topic is necessary to clarify the role of segmental syllable as a holistic unit of lexical processing. Experiment 1 aimed to address this issue.

**Issue two: relative weighting of sub-syllabic constituents in spoken word recognition**

A second issue on which the existing studies of SWR differ is whether, and if so, to what extent sub-syllabic
segmental constituents (i.e. onset and rhyme) and lexical tone affect lexical activation. Continuous mapping models such as TRACE (McClelland & Elman, 1986) predicted that word candidates with the same onset are activated earlier and greater than word candidates which share the same rhyme. Such a prediction was confirmed for English by a seminal eye-tracking study by Allopenna et al. (1998) with the visual world paradigm. In this study, participants were asked to follow instructions (e.g. *Pick up the beaker*) and move objects around on a computer screen. They looked at both the target *beaker* and its phonological competitors (i.e. the cohort competitor *beetle* and the rhyme competitor *speaker*). Moreover, participants’ eye fixations towards cohort competitors were significantly earlier than those of rhyme competitors.

However, the effect of sub-syllabic units (i.e. cohort and rhyme) reported for English seems less reliable in

---

### Table 1. Experimental conditions and sample stimuli of previous studies and the present study.

| Study                  | Task                                      | Conditions         | Sample Stimulus (Pinyin)        | Shared information                  | Divergent information       |
|------------------------|-------------------------------------------|--------------------|---------------------------------|--------------------------------------|-----------------------------|
| Zhao et al. (2011)     | Picture/spoken-word/picture task          | Match              | bi2-bi2                         | All phonemes and tone                | None                        |
|                        |                                            | Onset Mismatch     | bi2-li2                         | Rhyme and tone                       | Onset                       |
|                        |                                            | Rime Mismatch      | bi2-bo2                         | Onset and tone                       | Rhyme                       |
|                        |                                            | Tone Mismatch      | bi2-bi3                         | All phonemes                         | Tone                        |
|                        |                                            | Syllable Mismatch  | bi2-ge1                         | None                                 | All phonemes and tone       |
| Sereno and Lee (2015)  | Auditory priming paradigm, Experiment 1   | Tone-segment Overlap| ru4-ru4                         | All phonemes and tone                | None                        |
|                        |                                            | Segment-only Overlap| ru4-ru3                         | All phonemes                         | Tone                        |
|                        |                                            | Tone-only Overlap  | sho4-ru4                        | Tone                                 | All phonemes                |
|                        |                                            | Unrelated          | qin1-ru4                        | None                                 | All phonemes and tone       |
|                        | Auditory priming paradigm, Experiment 2   | Tone-segment Overlap| ru4-ru4                         | All phonemes and tone                | None                        |
|                        |                                            | Only-onset segment Overlap| ru4-re4                         | Onset and tone                       | Rhyme                       |
|                        |                                            | Only-rime Overlap  | ru4-pu4                         | Rhyme and tone                       | Onset                       |
|                        |                                            | Unrelated          | ru4-qin1                        | None                                 | All phonemes and tone       |
| Malins and Joanisse (2012a) | Picture/spoken-word matching task | Segmental Cohort | hua1-hua4                       | All phonemes                         | Tone                        |
|                        |                                            | Rhyme              | hua1-gua1                       | Onset, glide and tone                | Rhyme                       |
|                        |                                            | Tonal              | hua1-jing1                      | Tone                                 | All phonemes                |
|                        |                                            | Unrelated          | hua1-lang2                      | None                                 | All phonemes and tone       |
| Ho et al. (2019)       | Cross-modal priming paradigm              | Match              | jia1-jia1                       | All phonemes and tone                | None                        |
|                        |                                            | Onset Violation    | jia1-xia1                       | Rhyme and tone                       | Onset                       |
|                        |                                            | Tone Violation     | jia1-jia4                       | All phonemes                         | Tone                        |
|                        |                                            | Syllable Violation | jia1-tang2                      | None                                 | All phonemes and tone       |
| Malins and Joanisse (2010) | Visual World Paradigm                     | Segmental Cohort | chuang2-chuang1                 | All phonemes                         | Tone                        |
|                        |                                            | Rhyme              | chuang2-chuang2                 | Onset, glide, and tone               | Rhyme                       |
|                        |                                            | Tonal              | chuang2-huang2                  | Rhyme and tone                       | Onset                       |
|                        |                                            | segmental          | chuang2-ji1                     | All phonemes                         | Tone                        |
| Zou (2017)             | Visual World Paradigm                     | Segmental Cohort | chuang1-chuang2                 | All phonemes                         | Tone                        |
|                        |                                            | Rhyme              | chuang1-chuang1                 | Onset and tone                       | Rhyme                       |
|                        |                                            | Tonal              | chuang1-chuang2                 | Tone                                 | All phonemes                |
| Connell (2017)         | Visual World Paradigm                     | Segmental Cohort | ban1-ban3                       | All phonemes                         | Tone                        |
|                        |                                            | Rhyme              | ban1-ban3                       | Onset, glide, and tone               | Rhyme                       |
|                        |                                            | Tonal              | ban1-ban3                       | None                                 | All phonemes and tone       |
| The present study      | Visual World Paradigm, Experiment 1&2     | Segmental syllable Cohort | chuang2-chuang1                 | All phonemes                         | Tone                        |
|                        |                                            | Rhyme              | chuang2-chuang2                 | Onset and tone                       | Rhyme                       |
|                        |                                            | Tonal              | chuang2-di1                     | Tone                                 | Onset                       |
|                        | Visual World Paradigm, Experiment 3       | Segmental Early    | dian4-dou4                      | All phonemes                         | Tone                        |
|                        |                                            | Segmental Late     | xue3-xuan3                      | Tone                                 | Onset                       |
|                        |                                            | Tonal Early        | ying1-ying3                     | All phonemes                         | Tone                        |
|                        |                                            | Tonal Late         | chou3-chou2                     | Tone (onset and offset)              |                             |

All listed previous studies were discussed in the Introduction in the same order.
Mandarin SWR. With the same visual world paradigm, Malins and Joanisse (2010) examined the effect of phonological similarity on Mandarin word recognition. Their results showed that given a target such as *chuang2* “bed”, both segmental syllable (e.g. *chuang1* “window”) and cohort (e.g. *chuau2* “boat”) competitors distracted fixations towards target pictures significantly, with no difference between the two conditions in terms of effect size and time course. However, in contrast to the findings of Allopenna et al. (1998), rhyme competitors (e.g. *huang2* “yellow”) did not influence participants’ gaze patterns more than unrelated distractors. These findings led Malins and Joanisse (2010) to propose that sub-syllabic constitutes weigh differently in Mandarin and English SWR.

Results reported in Malins and Joanisse (2010), however, are not fully replicated. Zou (2017) used a similar design and investigated phonological competition effects in Mandarin SWR. Although the goal of the study was to examine SWR by second language learners of Mandarin (with Dutch as the first language), native Mandarin listeners were also included as a control group. Zou (2017) showed that the presence of rhyme competitors distracted participants’ looks to targets the most. In contrast, the cohort competitors did not differ from Malins & Joanisse’s results. Another difference between Malins and Joanisse (2010) and Zou (2017) is the role of lexical tone in SWR. Malins and Joanisse (2010) reported an early interference effect of tonal competitors. Zou (2017), however, did not observe this effect. Similar to Zou (2017), Connell (2017) examined the process of word recognition in L1 and L2 Mandarin listeners with a visual world eye-tracking experiment. Unlike Malins and Joanisse (2010), in which comparable effects of segmental syllable and cohort competition were found, Connell (2017) found significantly more target eye-fixations with segmental overlapping distractors than with cohort competitors (see Table 1 for sample stimuli used in the study). Overall, these different results raise further questions about the role of all sub-syllabic units (i.e. onset, rhyme, tone) in Mandarin spoken word processing.

It is worth noting that discrepant results between Malins and Joanisse (2010) and Zou (2017) are likely to lie in two major differences in their methods. One is the stimuli used for different competitor conditions, and the other is the different preview times for participants to view pictures before listening to the auditory stimuli.

About the stimuli, there are two differences. One concerns the cohort competitors. Malins and Joanisse (2010) defined the cohort competitors as sharing onset, tone, and the glide or rhyme with the targets (e.g. *hua1* “flower”- *hui1* “grey”; *tu3* “dirt”- *tui3* “leg”). In Zou (2017), cohort competitors were controlled more consistently as sharing only the lexical tone and the first phoneme in the onset (e.g. *tang2* “candy”- *tou2* “head”). The other concerns the repeated items. In Malins and Joanisse (2010), a few items were used repeatedly, especially in the tonal condition. For example, all tonal competitors were also presented as segmental/rhyme competitors; the word *mi3* (rice) was not only used as a tonal competitor for both target words *xin1* (heart) and *tu3* (dirt), but also a segment competitor for the target word *mi4* (honey). This led to an overall unequal number of occurrences for various phonological competitors and increased familiarity with tonal competitors. Zou (2017) avoided using the same stimuli for different phonological competitors.

As for the preview time difference, Malins and Joanisse (2010) allowed for a preview time of 1500 ms while Zou (2017) presented the pictures and auditory stimuli simultaneously. Preview time has been shown to affect phonological competition effects in the visual word paradigm (Huettig & McQueen, 2007; Huettig et al., 2011). Huettig and McQueen (2007) found that when participants viewed pictures at sentence onset (with an estimation of 700 ms–1000 ms preview time), substantial online phonological competition effects were found during Dutch spoken word recognition. However, no phonological competition effect was found when participants viewed pictures with a preview time of 200 ms. Huettig and McQueen (2007) thus proposed that a 200 ms preview may not be sufficient for participants to retrieve the names of the displayed objects and associate them with locations in their visuospatial working memory. It is worth noting that Huettig and McQueen (2007) adopted a modified version of the visual world paradigm in which no target but three different types of competitors were presented. Also, their participants were not instructed to give any explicit response. Given that how participants approached this task is still unclear (Magnuson, 2019), it leaves open the question of the impact of preview time on phonological competition effects. Specifically, is a 200 ms preview a prerequisite for observing phonological competition using a standard visual world paradigm? More importantly, to what extent the length difference of preview time could help to account for the discrepant results in Mandarin SWR?

To summarise, the different findings on the role of sub-syllabic constituents in Mandarin SWR may have resulted from different preview times and the unequal occurrences of the same stimuli as various phonological competitors. Therefore, new experiments with stricter control of stimuli (Experiment 1) and different preview time (Experiment 2) would illuminate and resolve the conflicting results.
Issue three: segment and lexical tone processing in Mandarin spoken word recognition

Regardless of whether the primary processing unit in Mandarin is a segmental syllable or a sub-syllabic unit, a third unresolved issue is whether exactly segmental information and tonal information are recognised and utilised during SWR. Existing studies on the role of lexical tone in spoken word processing have mainly focused on whether lexical tone plays a similar role as segments. Using various behavioural tasks, a perceptual disadvantage of lexical tone, compared with segmental information, has been reported in earlier studies (Cutler & Chen, 1997; Taft & Chen, 1992; Yip, 2001; Ye & Connine, 1999, experiment 1). Such a view has also been supported by a recent study (e.g. Hu et al., 2012 with comparison to vowels; Sereno & Lee, 2015 with comparison to segmental syllables; Gao et al., 2019 with comparison to segmental syllables). This body of literature reasoned that tonal information plays a weaker role during lexical processing because such information “often arrives later than does information about the vowel that bears the tone” (Cutler & Chen, 1997) and is “less informative than segmental information” (Tong et al., 2008).

An increasing number of studies, many with experimental techniques that are more sensitive to the time course of speech processing, have provided evidence that lexical tonal information is processed timely and can play an essential role during SWR. For example, Schirmer et al. (2005) showed that mismatched tonal and segmental (rhyme) targets induce comparable ERPs in Cantonese word processing with a sentence completion task. They thus argued that tone and segment play comparable roles and are accessed with a similar time course during spoken word processing. Using the visual world paradigm, Malins and Joanisse (2010) found comparable competition effects between cohort and segmental competitors (in terms of amplitude and time course). This was interpreted as evidence that tonal and segmental information is accessed concurrently during online SWR.

Connell et al. (2016) tapped into this debate by examining the low-level perceptual difference between tone and segments with a gated AX-discrimination task. Their native Chinese listeners showed a delay of about 28 ms to perceive tonal contrast compared to segmental contrasts even when they have comparable acoustic divergent points. This raises a question: is there an inherent delay of tonal perception, why is it not reflected in time-sensitive online experiments? Connell (2017) examined this issue further by conducting an eye-tracking visual world experiment. With the acoustic and perceptual divergence points strictly controlled, Connell found that lexical tones are used no later and even more rapidly than segments in constraining word activation. One possible explanation is that lexical tones are more efficient in eliminating potential lexical candidates than vowels. If this is the case, tonal information must be used in lexical access even before the tone can be recognised. Qin (2017) looked into this issue by conducting an eye-tracking experiment with tone pairs that either has early pitch height overlapping (T1–T2) or not (T1–T4). Qin found a larger target-over-competitor activation when there was an early pitch height difference. This suggests that pitch height information can be used early to constrain word recognition. Overall, findings of Connell (2017) and Qin (2017) and previous online studies, have provided evidence that lexical tone can be used before being recognised in lexical access.

Nevertheless, no study has systematically examined and compared how the point of divergence (hereafter POD) affects tone and segments processing in Mandarin SWR with paradigms such as the visual world paradigm that are sensitive to the time course of speech processing. Experiments are needed to address the following open questions. First, given that there is evidence for holistic processing of syllable in Mandarin (Zhao et al., 2011), whether and how does POD affect the spoken word recognition process? Second, does lexical tone (with early/late diverging pitch contours) constrain word recognition more than vowels? Answers to these questions would lend strong evidence to the exact time course of how the two tiers of information (i.e. segmental vs. tonal) are utilised for SWR. Experiment 3 was designed to fill this knowledge gap, which also serves to replicate existing findings in Qin (2017) and Connell (2017).

To summarise, the present study consists of three experiments and aims to clarify the role of segmental syllable and sub-syllabic constituents in Mandarin SWR, as well as to investigate the time course of when segmental and suprasegmental tonal information is utilised during lexical processing. All three experiments were conducted within the visual world paradigm (Allopenna et al., 1998; Tanenhaus et al., 1995).

Experiment 1

Experiment 1 examined the role of segmental syllable, sub-syllabic segmental constituent (onset and rhyme), and lexical tone in Mandarin SWR, as indexed by how much participants’ visual attention on the target word is disrupted by the presence of a phonological
competitor (with an overlapping segmental syllable, onset, rhyme or tone) when they listen to a target Mandarin word.

Given the debates in the existing literature, particularly the discrepancies between Malins and Joanisse (2010) and Zou (2017), our goal was to replicate some of the findings conceptually to resolve the discrepancies. We followed Malins and Joanisse (2010) for most of the design, but made some necessary modifications as motivated earlier:

First, we avoided using the same stimuli as different phonological competitors and kept the number of occurrences of tonal competitors the same as other competitors. This would ensure that the tonal competitor effect reported in Malins and Joanisse (2010) is introduced by lexical tone overlap and not due to the effect of familiarity. We made sure that there was an equal number of reciprocal trials in which the role of target and competitor in critical trials was reversed. Thus, participants’ chances of hearing the target and competitor in a trial remained the same. Additionally, we also made sure that participants’ chances of seeing the target and competitor pictures were the same by arranging competitors of one target as unrelated distractors of another. In this way, participants’ chances of predicting the targets and developing strategic responses were controlled to be the same.

Second, we changed a subset of the stimuli used in the cohort condition. Following Zou (2017), we defined the cohort competitor as sharing only the first onset phoneme with the target. This is because our stimuli are monosyllabic words. In SC, mono-syllables either constitute a word or at least a morpheme; the syllable structure (C)V(C) (with optional onset and coda) which serves as the bearing unit of lexical tone is also relatively simple. Based on such characteristics, previous studies on Chinese lexical access often examined the role of onset and rhyme respectively (e.g. Ho et al., 2019; Yip, 2001; Zhao et al., 2011; Zou, 2017). With an auditory priming lexical decision task, Yip (2001) observed that onset and tone overlapping between target and prime elicited an inhibitory effect whereas rhyme and tone overlapping introduced a facilitatory effect in Cantonese. The first onset phoneme (plus lexical tone) likely plays an independent and rather different role from rhyme (plus lexical tone) in Chinese. Thus, to better compare the relative contribution of sub-syllabic constituents in SC, we selected words that share the first onset phoneme and tone with targets as cohort competitors, despite that traditionally cohort words for studies in Germanic and Romance languages have been defined as sharing two or more phonemes (Marslen-Wilson, 1987).

Method

Participants

Twenty (mean age: 20, standard deviation: 0.8; 12 females, 8 males) native Mandarin speakers participated in the experiment. All participants were college students from Shaanxi Normal University. All of them reported normal hearing and no history of speech or language disorders. All participants identified Standard Chinese as their first language, and none of them speak other regional Chinese dialects. This study was approved by the Ethics Committee at Leiden University Centre for Linguistics. All participants provided informed consent before participation and were paid 30 RMB in compensation for their time.

Stimuli

The stimuli consisted of 60 monosyllabic Mandarin words which are easily picturable nouns (see supplementary material). Among the stimuli, 12 were critical targets. For each critical target, competitors of four conditions were defined based on their phonological overlap with the target. Segmental syllable competitor shared all phonemes but differed in tone with the target; cohort competitor shared the initial consonant and tone with the target; rhyme competitor shared rhyme and tone with the target; tonal competitor shared tone alone with the target. See Table 1 for sample stimuli and their comparison with previous studies. No item was used in more than one competitor condition.

Word frequency, as computed with SUBTLEX-CH (Cai & Brysbaert, 2010), was balanced across target words and the four competitor conditions [F(4, 55) = 0.83, p > 0.5]. All stimuli were recorded through a Sennheiser MKH416T microphone (44.1 kHz, 16 bit) at the Phonetics Lab of Leiden University, produced by a female native speaker of Standard Chinese who was born and grew up in Beijing. Each word was read four times in isolation using a randomised list. One token of each word was chosen based on its clarity. All stimuli were normalised for intensity at 70 dB. The matching pictures were real object pictures selected with the assistance and consensus of three native Chinese speakers who did not participate in the experiment.

Procedure and design

To ensure participants were familiar with all stimuli, a naming task was assigned preceding the eye-tracking recording. During the naming session, participants were shown the pictures and asked to name them with appropriate Standard Chinese words. If the name produced was not the intended word, participants were provided with the intended name.
During the subsequent experiment, participants were tested in a sound-attenuated booth at the Psychology Lab of Shaanxi Normal University. While performing the task, participants’ eye movements were recorded with SR Eyelink Portable DUO eye-tracker at a sampling rate of 500 Hz. For visual stimuli display, a 24-inch DELL U2412M monitor was located behind the eye-tracker. The camera of the eye-tracker was at a distance of about 52 cm from the participants’ eye, which was fixed with the help of a chin rest. The auditory stimuli were played over a Beyer DT-770 Pro dynamic headphone at a constant and comfortable hearing level.

Before the test, participants’ eye gaze position was validated and calibrated with a 9-point grid. At the beginning of each trial, a central cross appeared on the screen for 500 ms. Participants were asked to look directly at the fixation for a drift check. Four pictures then appeared on the screen for 1500 ms before an auditory word. The four pictures (300 × 300 pixels) were placed top-left, top-right, bottom-left, and bottom-right; each comprising a distinct quadrant of the display (see a sample display in the supplementary material). Participants were required to click on the picture that matches the auditory word with a mouse. The next trial appeared 1000 ms after the click. The target picture’s position was counterbalanced so that the target picture appeared an equal number of times in each location, and did not appear in the same location in two consecutive trials.

All the instructions were given in Standard Chinese. Participants were first asked to complete a practice block of four trials. In total, there were 360 trials for four blocks of 90 trials. The block order was counterbalanced across participants. Between each block, participants were given time to rest and proceed as they wish. Each of the syllable, cohort, rhyme, and tonal conditions has 36 trials, in which the participants listened to the targets with corresponding phonological competitors in the display. Additionally, there was a baseline condition in which no competitor but only distractors were presented along with the target. Half of the trials (180) were fillers, in which the role of target and competitors were reversed (i.e. the phonological competitors were played as auditory targets) so that the chances of hearing the target and competitors with the same display were equal. Furthermore, to balance the overall occurrences of target and competitor items as picture displays, competitors were taken as unrelated distractors in another set of stimuli. The same target did not appear in three consecutive trials. After the test, participants were asked to fill in a language background questionnaire.

Data analysis

Analysis of behaviour data. Reaction time and response accuracy for mouse clicks were collected for statistical analysis. Reaction time was calculated with respect to the onset of the auditory word. Trials for which the reaction time was shorter than 250 ms were excluded for both accuracy and RT analyses. Furthermore, only correct responses were considered for RT analyses. RTs were analyzed using the generalised linear mixed-effects model (GLMM) to account for the skewed distribution without transforming the raw data (Lo & Andrews, 2015). A backward algorithm was used to select the model (Barr et al., 2013). A maximum model including fixed effects of experimental conditions, by-subject and by-item random intercept, by-subject and by-item random slopes for experimental conditions was constructed first. If a model failed to converge, we first increased the number of iterations, then simplified the model by removing correlation parameters and the random structure’s main effects (Brauer & Curtin, 2018). Fixed effects and the random structure were tested by comparing the likelihood ratio test with the simpler model. Response accuracy was modelled using the same approach using GLMM. All the analyses were run in the R software (R Core Team, 2017) with the package lme4 (Bates et al., 2015).

Analysis of eye-tracking data. We excluded trials for which the target was not correctly identified and trials for which the reaction time was shorter than 250 ms. The time window of 200–980 ms post auditory stimuli onset was chosen as our interest period. The lower boundary was chosen because there is a 200 ms delay to launch an eye movement (Hallett, 1986), while the upper boundary reflects when there were approximately maximum looks towards the targets. As the gaze position and duration of participants’ eye fixation were recorded, looks toward targets, competitors, and distractors during the interest period were collected. The collected eye-tracking data were first downsampled to 50 Hz (a 20 ms bin), following the tutorial of Porretta et al. (2018). Then, the proportions of fixations to target, competitor, and distractors at each time point were calculated by dividing the sum of fixations on the four pictures (target, competitor, and two distractors) by the number of fixations toward each picture type. The eye-fixation data in the visual world paradigm is intrinsically binary, i.e. participants are either looking at the target/competitor or not. It has been questioned that treating the eye-tracking data as a ratio variable on a linear scale averaging across conditions may cause problems such as data distortion and the violation
of the assumptions of parametric statistics (Huang & Snedeker, 2020). To avoid these issues, we performed empirical logit transformation with weights for variance estimation on eye-fixation proportions following the advice of Mirman (2014) and Porretta et al. (2018).

We used generalised additive mixed modelling (GAMM; Wood, 2017) to analyse the eye-tracking data.1 GAMM is a type of generalised mixed-effects model that uses smooth functions to model the non-linearity between predictor(s) and the dependent variable. The smooth function (e.g. the thin plate regression spline) combines a number of pre-defined basic functions by multiplying them with individual coefficients. With cross-validation or maximum likelihood estimation, GAMM adds a penalisation to the estimation of the coefficients to avoid over-fitting and minimise errors. GAMM is well-established and has been applied to eye movement data of the visual world paradigm (e.g. Nixon & Best, 2017; Nixon et al., 2016; Porretta et al., 2018). Compared with polynomial models such as growth curve analysis, GAMM is not only more flexible in shapes, but also offers the option to take into account the auto-correlation in time-series data (Porretta et al., 2018).

We used the mgcv package (Wood, 2017) in R (R Core Team, 2017) to implement GAMM. The model was fit by first entering all predictors of interest. Model comparison was conducted by means of $\chi^2$ tests of fREML scores, using the “compareML” function in the itsadug package (Van Rij et al., 2020). Model residuals were examined to check for non-normality, heteroscedasticity, and auto-correlation. The model summary of GAMM includes parametric coefficients and smooth terms. The parametric coefficients can be interpreted the same way as linear models, with the intercepts indicating the overall heights of the trajectories. The smooth terms capture the shape of the looking trajectories. To test the statistical difference between each experimental condition, we used ordered factors to model the difference smooth. The $p$-value in the smooth terms thus indicates the statistical difference between the trajectories in terms of shape. To control the family-wise error rate, the Holm–Bonferroni method was applied to adjust the $p$-values (Holm, 1979). We also plotted the difference smooths with tidyedm (Coretta, 2021) to show when and how the look trajectories differ.

Plots of the raw eye-tracking factors to model the differences between each experimental condition, we used ordered looking trajectories. To test the statistical differences, the smooth terms thus indicates the statistical difference between the trajectories in terms of shape. To control the family-wise error rate, the Holm–Bonferroni method was applied to adjust the $p$-values (Holm, 1979). We also plotted the difference smooths with tidyedm (Coretta, 2021) to show when and how the look trajectories differ. Plots of the raw eye-tracking proportions (without empirical logit transformation) are also included in the supplementary material.

**Results**

**Behavioural data**

Reaction time and response accuracy for mouse click are shown in Table 2. For reaction time, the maximum likelihood estimation of the maximum model and the simplified random slope models failed to reach convergence. The best-fit model included fixed effects of experimental condition, by-subject and by-item random intercepts (note that random-intercepts-only models may inflate Type-I error rate; Barr et al., 2013). The fixed effects of experimental conditions ($\chi^2 (4) = 81.221, p < 0.001$) suggested that participants’ reaction time differed across conditions. Post-hoc analysis revealed that only when segmental syllable competitors were present, participants took longer time to identify the targets (segmental syllable condition: $p < 0.001$; cohort condition: $p = 0.454$; rhyme condition: $p = 0.775$; tonal condition: $p = 0.075$). The error rate was low in each condition (all approximately under 1%). Thus, no further analyses were conducted on the response accuracy.

**Eye movement data**

**Looks to target.** The final model of target fixations includes a fixed effect of condition, a smooth term of time, a smooth over time for each level of condition, and a non-linear random effect of subject by condition. The final model explains 98.4% of the deviance. The summary of model fit is provided in Table 3. The upper half of this table presents the parametric coefficients of the model. The first row presents the intercept of the baseline condition. The following rows indicate the changes in the intercept for the other four experimental conditions. As shown in Table 3, no condition was found

**Table 2.** Mean Reaction time (ms) and mean percent response accuracy of Experiment 1.

| Condition            | Reaction time (SE) | Percent Accuracy (SE) |
|----------------------|--------------------|-----------------------|
| Baseline             | 1053 (25.4)        | 99.7 (2.32)           |
| Cohort               | 1067 (29.4)        | 98.9 (8.65)           |
| Rhyme                | 1056 (26.9)        | 99.8 (2.05)           |
| Segmental syllable   | 1116 (31.2)        | 99.4 (3.41)           |
| Total                | 1088 (23.6)        | 99.7 (2.76)           |

Standard Errors are in parentheses.

**Table 3.** GAMM analysis of fixation proportions to targets in Experiment 1 with 1500 ms preview time.

| Estimate | SE   | t value | p-value |
|----------|------|---------|---------|
| Intercept| 0.467| 0.126   | 3.716 | 0.000 *** |
| Cohort – Baseline | –0.017 | 0.179 | –0.097 | 0.923 |
| Rhyme – Baseline | 0.007 | 0.181 | 0.039 | 0.969 |
| Segmental syllable – Baseline | –0.161 | 0.179 | –0.906 | 0.397 |
| Tone – Baseline | 0.014 | 0.179 | 0.079 | 0.937 |
| s(Time) | 8.658 | 8.739 | 175.435 | 0.000 *** |
| s(Time): Cohort – Baseline | 1.001 | 1.001 | 0.043 | 0.836 |
| s(Time): Rhyme – Baseline | 1.001 | 1.001 | 1.027 | 0.311 |
| s(Time): Segmental syllable – Baseline | 4.037 | 4.389 | 4.197 | 0.002 ** |
| s(Time): Tone – Baseline | 1.001 | 1.001 | 0.615 | 0.433 |
to be significantly different from the baseline condition in the intercept.

The second half of Table 3 describes the thin plate regression spline smooths for different levels of conditions over time. The first smooth presents the trajectory of the (empirical logit transformed) proportions of eye fixations over time for the baseline condition. The next four smooths evaluate the curves’ difference with respect to the baseline condition. The model summary indicates that there was a significant difference between the segmental syllable and the baseline conditions ($p < 0.005$).

The smooths for all levels of conditions are visualised in Figure 1A. Figure 1B plots the difference between the two smooths comparing the segment and baseline condition.

**Looks to competitors.** The final model of competitor fixations includes a fixed effect of condition, a smooth term of time, a smooth over time for each level of condition and a non-linear random effect of subject by condition. The final model explains 97.2% of the deviance. The summary of model fit is provided in Table 4.

![Figure 1](image-url)

**Table 4.** GAMM analysis of fixation proportions to competitors in Experiment 1 with 1500 ms preview time.

| Estimate | SE   | t value | p-value |
|----------|------|---------|---------|
| Intercept| 3.179| 0.096   | 33.283  | 0.000 *** |
| Cohort – Baseline| –0.009| 0.140 | –0.063| 0.950 |
| Rhyme – Baseline| –0.054| 0.127 | –0.422| 0.673 |
| Segmental syllable – Baseline| 0.536| 0.169 | 3.170| 0.002 ** |
| Tone – Baseline| –0.030| 0.135 | –0.224| 0.823 |
| s(Time) | 7.675| 8.050 | 32.905| 0.000 *** |
| s(Time): Cohort – Baseline| 1.000| 1.000 | 0.738| 0.390 |
| s(Time): Rhyme – Baseline| 1.001| 1.001 | 0.114| 0.736 |
| s(Time): Segmental syllable – Baseline| 5.396| 5.811 | 5.230| 0.000 *** |
| s(Time): Tone – Baseline| 1.000| 1.000 | 0.075| 0.785 |

The green background in B and D indicates that the shaded confidence band is significantly different from zero.
The parametric coefficients of the model indicate that only the segmental syllable condition was significantly different from the baseline condition in intercept \((p < 0.005)\). In the segmental syllable condition, the empirical logit of eye-fixation proportions towards competitors was higher than that of the baseline condition by 0.536. The smooth terms of the GAMMs (as shown in Table 4) indicate that there was a significant difference between the segmental syllable and the baseline conditions over time \((p < 0.001)\). The smooths for all levels of conditions are visualised in Figure 1C. Figure 1D plots the smooth difference between the segmental syllable and baseline condition.

**Discussion**

Results of Experiment 1 showed a significant segmental syllable competitor effect, confirming findings reported in Malins and Joanisse (2010) and Zou (2017). Different from findings in Malins and Joanisse (2010) but confirming Zou (2017), there were no cohort and tonal competition effects. Note that the different cohort effects are likely due to the different definitions of cohort (see further discussion below). Furthermore, different from Zou (2017), no rhyme competition effect was observed, confirming the lack of rhyme competition effect reported in Malins and Joanisse (2010). The different findings in the rhyme condition may be in part due to the different preview times. The possible effects of preview time on spoken word processing were investigated in Experiment 2.

To summarise, our study confirmed that segmental syllable competitors exhibit a larger competition effect over cohort, rhyme, and tonal competitors. The results thus lend further support that segmental syllable has an overall advantage over sub-syllabic segmental constituents and lexical tone during SWR. The effects of sub-syllabic units seem much more variable and may be subject to the influence of factors such as preview time.

**Experiment 2**

Experiment 2 aimed to investigate the impact of preview time on the phonological interference effects during lexical processing. In this experiment, we changed the preview time to 200 ms (from the 1500 ms in Experiment 1) while keeping everything else the same across the two experiments. If the amount of preview time given to participants is indeed a critical factor for some of the inconsistent findings, we should observe different results from Experiment 1, in similar ways as some of the results of Malins and Joanisse (2010) differ from that of Zou (2017).

We opted for a 200 ms preview instead of no preview for two main reasons. First, preview time allows listeners to perform object recognition, visual search, and other non-lexical processes before the onset of the spoken word; without it, listeners must attend to visual properties simultaneously, which has been found to add noise to the phonological competition effect (Apfelbaum et al., 2021). Second, as mentioned earlier, 200 ms has been found to be insufficient for observing phonological competition with a non-standard visual world paradigm (Huettig & McQueen, 2007). Whether such a short preview time would delay or even cancel phonological competition effects with a standard visual word paradigm has been questioned since and is worthy of further investigation (Magnuson, 2019).

**Methods**

**Participants**

23 (mean age: 19, standard deviation: 1.8; 14 females, nine males) new native Mandarin speakers participated in the experiment. As in Experiment 1, all participants were college students from Shaanxi Normal University, with normal hearing and no history of speech or language disorders. All participants speak Standard Chinese and no other Chinese varieties. This study was approved by the Ethics Committee at Leiden University Centre for Linguistics. All participants provided informed consent before participation and were paid 30 RMB in compensation for their time.

**Stimuli**

The same stimuli of Experiment 1 were used.

**Procedure and design**

The same procedure of Experiment 1 was used, except that the amount of time given to participants for viewing the pictures before the auditory stimuli was shortened from 1500 ms (in Experiment 1) to 200 ms.

**Table 5.** Mean Reaction time (ms) and mean percent response accuracy of Experiment 2.

| Condition         | Reaction Time (SE) | Percent Accuracy (SE) |
|-------------------|--------------------|-----------------------|
| Baseline          | 975 (41.9)         | 96.9 (2.12)           |
| Cohort            | 976 (39.9)         | 97.6 (1.29)           |
| Rhyme             | 988 (46.3)         | 99 (0.81)             |
| Segmental syllable| 1054 (43.8)        | 98.2 (1.05)           |
| Tonal             | 995 (46.5)         | 98.0 (1.25)           |

Standard Error are in parentheses.
Results

Behavioural data

Reaction time and response accuracy for mouse click are shown in Table 5. The best-fit reaction time model included fixed effects of experimental condition, by-subject, and by-item random intercepts. There was significant effect for fixed effects of experimental condition ($\chi^2(4) = 37.521, p < 0.001$). Post-hoc analysis revealed that only in the segmental syllable condition, reaction time was significantly different from the baseline condition (segmental syllable condition: $p < 0.005$; cohort condition: $p = 0.175$; rhyme condition: $p = 0.464$; tonal condition: $p = 0.445$). For the best-fit accuracy model, the fixed factor of condition did not improve model fit, which suggests that participants’ response accuracy did not differ across conditions ($\chi^2(4) = 4.6957, p = 0.32$).

Eye movement data

Looks to target. The model of target fixations includes the main effect of condition, a smooth term of time, a

Table 6. GAMM analysis of fixation proportions to targets in Experiment 2 with 200 ms preview time.

|                      | Estimate | SE    | t value | p-value |
|----------------------|----------|-------|---------|---------|
| Intercept            | 0.116    | 0.206 | 0.565   | 0.572   |
| Cohort – Baseline    | -0.076   | 0.300 | -0.254  | 0.800   |
| Rhyme – Baseline     | 0.007    | 0.273 | 0.027   | 0.979   |
| Segmental syllable – Baseline | -0.212  | 0.274 | -0.776  | 0.438   |
| Tone – Baseline      | -0.001   | 0.280 | -0.005  | 0.996   |
| edf                  |          |       | F       | p-value |
| s(Time)              | 8.609    | 8.686 | 121.067 | 0.000   ***
| s(Time):Cohort – Baseline | 1.000    | 1.000 | 1.910   | 0.167   |
| s(Time):Rhyme – Baseline | 1.000    | 1.000 | 0.988   | 0.320   |
| s(Time):Segmental syllable – Baseline | 5.445     | 5.863 | 5.050   | 0.000   **
| s(Time):Tone – Baseline |          |       | F       | p-value |

Figure 2. Estimated smooths for all conditions and smooth differences in Experiment 2. (A) Smooths of target fixations for all conditions; (B) Difference between the two smooths comparing the segment and baseline condition of target fixations model; (C) Smooths of competitor fixations for all conditions; (D) Difference between the two smooths comparing the segmental syllable and baseline condition of competitor fixations model. The pointwise 95%-confidence intervals are shown by shaded bands. The green background in B. and D. indicates that the shaded confidence band is significantly different from zero.
smooth over time for each level of condition and a non-linear random effect of subject by condition. The final model explains 98.5% of the deviance. The summary of model fit is provided in Table 6.

The parametric coefficients of GAMM analysis indicate that no condition was significantly different from the baseline condition in the intercept. The smooth terms indicate that there was a significant difference in target fixations between the syllable and baseline conditions over time (p < 0.001). The smooths for all levels of conditions are visualised in Figure 2A. Figure 2B plots the smooth difference between the segmental syllable and the baseline condition.

**Looks to competitors.** The final model of competitor fixations includes a fixed effect of condition, a smooth term of time, a smooth over time for each level of condition and a non-linear random effect of subject by condition. The final model explains 97.1% of the deviance. The summary of model fit is provided in Table 7.

The parametric coefficients of the model indicate that only the segmental syllable condition was significantly different from the baseline syllable condition in intercept (p < 0.005). In the segmental syllable condition, the empirical logit of eye-fixation proportions towards competitors was higher than that of the baseline condition by 0.408.

The smooth terms of the GAMMs (as shown in Table 7) indicate that there was a significant difference between the segmental syllable and the baseline conditions over time (p < 0.001). The estimated smooths for all levels of conditions are visualised in Figure 2C. Figure 2D plots the estimated smooth difference between the segmental syllable and baseline condition.

**The effect of preview time differences between Experiment 1 and 2.** We examined further the effect of preview time on Mandarin phonological competition effects. Similar general additive modelling procedures described above were applied to the combined data of Experiment 1 and Experiment 2. The interaction of preview time and experimental condition was added to the model and tested for exclusion.

Compared with models including the factor of condition, adding preview time significantly improved model fit of target fixations (p < 0.001). The interaction between preview and condition did not significantly improve model fit. Coefficients of parameter estimations (see Table 8) showed significant differences in intercept and smooth terms for different preview times (all p < 0.001). Figure 3A shows estimated smooths of target fixations of both preview times in baseline condition. Figure 3B shows the estimated smooth difference between the two preview times. As can be seen from Figure 3A, target eye-fixations reach the peak around 700 ms post stimuli onset in Experiment 1; around 600 ms in Experiment 2. Moreover, the target fixation peak in Experiment 1 has higher empirical logit transformed proportion than in Experiment 2. The estimated difference smooth in Figure 3B shows a consistent pattern. Overall, with a short preview time (200 ms), participants’ target fixation proportions reached the peak earlier with a relatively lower proportion compared with a long preview (1500 ms).

As for the model of competitor fixations, the interaction of preview time and condition also significantly improved model fit (p < 0.001). Five ordered factors each presenting the difference between two preview times of one condition was created. Table 9 shows the estimations of parametric coefficients and smooth terms of the final model. The results show that while there was no significant difference between Experiment 1 and Experiment 2 in the baseline condition (intercept: p = 0.910; smooth term: p = 0.720), there were significant differences in the cohort condition (intercept: p < 0.05; smooth term: p < 0.001), the segmental syllable condition (smooth term: p < 0.001), the rhyme condition (intercept: p < 0.05; smooth term: p < 0.001), and the tonal condition (intercept: p < 0.05; smooth term: p < 0.001). Figure 4 shows the estimated smooth differences between the two preview times.

**Table 7.** GAMM analysis of fixation proportions to competitors in Experiment 2 with 200 ms preview time.

| Estimate | SE  | t value | p-value |
|----------|-----|---------|---------|
| Intercept| −3.144| 0.085 | −57.067| 0.000***|
| Cohort – Baseline| 0.089| 0.099| 0.907| 0.365|
| Rhyme – Baseline| −0.129| 0.122| −1.058| 0.290|
| Segmental syllable – Baseline| 0.408| 0.142| 2.880| 0.004**|
| Tone – Baseline| −0.125| 0.110| −1.132| 0.258|
| s(Time): Segmental syllable – Baseline| 0.525| 0.525| 4.616| 0.000***|
| s(Time): Rhyme – Baseline| 1.000| 1.000| 0.025| 0.874|
| s(Time): Cohort – Baseline| 1.000| 1.000| 0.153| 0.696|
| s(Time): Preview P200-1500| 8.042| 8.352| 211.766| 0.000***|

**Table 8.** GAMM analysis of fixation proportions to targets in Experiment 1 & 2 (1500 ms & 200 ms).

| Estimate | SE  | t value | p-value |
|----------|-----|---------|---------|
| Intercept| 0.493| 0.130| 3.798| 0.000***|
| Segmental syllable| −0.347| 0.029| −12.022| 0.969|
| Cohort| −0.408| 0.029| −13.860| 0.969|
| Rhyme| −0.462| 0.029| −15.789| 0.314|
| Tonal| −0.316| 0.027| −11.567| 0.969|
| Preview P200-1500| −0.292| 0.029| −10.165| 0.969|
| s(Time): Intercept| 8.636| 8.712| 68.094| 0.496***|
| s(Time): Segmental syllable| 3.522| 3.818| 2.988| 0.813|
| s(Time): Cohort| 5.340| 5.747| 7.028| 0.969|
| s(Time): Rhyme| 4.289| 4.654| 3.455| 0.000***|
| s(Time): Tonal| 4.958| 5.358| 5.729| 0.969|
| s(Time): Preview P200-1500| 8.042| 8.552| 211.766| 0.000***|
between the two preview times for each experiment condition. Compared with Experiment 1, the segmental syllable competitors in Experiment 2 have more competitor fixations around 440–600 ms post stimuli onset, but fewer competitor fixations before 380 ms and after 800 ms post stimuli onset (see Figure 4B); the cohort condition has more competitor fixations around before 300 ms and after 640 ms, but less during around 320–560 ms (see Figure 4C); the rhyme condition has more competitor fixations around 220–340 ms and 480–560 ms (see Figure 4D); the tonal condition has more competitor fixations before around 380 ms, 580–600 ms, but less during around 400–520 ms (see Figure 4E). As for the baseline condition, there is no significant difference between Experiment 1 & 2 (see Figure 4A).

**Discussion**

As in Experiment 1, we found a robust competition effect when the segmental syllable competitors were present but null effects for the cohort, rhyme, and tonal competitors in Experiment 2. Nevertheless, with a close look at the effect of preview time on each experiment condition, we found that different preview times did affect participants’ visual attention to targets and phonological competitors differentially. With a short preview time, target eye-fixations reached the peak sooner. Moreover, the peak of target fixation proportions was lower than that with a long preview. These indicate that participants completed the visual search faster. It is likely that the short preview time created an overall faster rhythm of the task. Furthermore, under different preview times, there were also slightly different look trajectories for phonological competitors. Participants seemed to pay more attention to different phonological competitors at different time points over the time course of the target recognition. Such differences in competitor fixations between the two preview times need to be further verified.

Overall, regardless of having a short or a long preview time (i.e. 200 ms vs. 1500 ms), the segmental syllable competitor exhibited a significant phonological competition effect. Unlike in Huettig and McQueen (2007), which found reduced phonological competition with a

---

**Table 9.** GAMM analysis of fixation proportions to competitors in Experiment 1 vs. Experiment 2.

|                     | Estimate | SE  | t value | p-value |
|---------------------|----------|-----|---------|---------|
| Intercept           | -3.130   | 0.074 | -42.544 | 0.000 *** |
| Baseline: Exp.1 – Exp.2 | -0.004   | 0.034 | -0.113  | 0.910   |
| Segmental syllable: Exp.1 – Exp.2 | 0.055 | 0.040 | 1.383   | 0.170   |
| Cohort: Exp.1 – Exp.2  | 0.104 | 0.037 | 2.806   | 0.020 * |
| Rhyme: Exp.1 – Exp.2   | -0.100 | 0.031 | -3.212  | 0.007 **|
| Tonal: Exp.1 – Exp.2   | 0.121 | 0.036 | 3.348   | 0.005 **|
| s(Time)               | 6.403   | 7.101 | 42.484  | 0.000 ***|
| s(Time):Baseline: Exp.1 – Exp.2 | 2.206 | 2.605 | 1.167   | 0.220   |
| s(Time): Segmental syllable: Exp.1 – Exp.2 | 6.976 | 7.705 | 15.890  | 0.000 ***|
| s(Time):Cohort: Exp.1 – Exp.2 | 7.668 | 8.234 | 5.769   | 0.000 ***|
| s(Time):Rhyme: Exp.1 – Exp.2 | 4.480 | 5.312 | 15.325  | 0.000 ***|
| s(Time):Tonal: Exp.1 – Exp.2 | 7.802 | 8.372 | 17.400  | 0.000 ***|
200 ms preview, our results indicate that the length of preview does not affect the general phonological competition patterns in Mandarin SWR. Possible explanations for such discrepancy and its implications are discussed in the general discussion.

It is necessary to note that both Experiment 1 and 2 have a relatively small sample size with each having around 20 participants. Brysbaert (2019) recommended at least 50 participants using repeated measures and warned that studies underpower are more likely to miss genuine effects or increase false-positive results in the long run. We recognise the size limitation of our experiments and hereby remind the readers to interpret the results with caution. Given that Experiment 1 and 2 consistently generate the same eye-tracking pattern, however, we also feel that our findings on the segmental syllable condition are unlikely to be false outcomes.

Experiment 3

Experiment 3 took a closer look at the time course of how listeners utilise tonal and segmental information during online spoken word processing. We manipulated the timing of the point of divergence (POD: early vs. late) for acoustic cues in two information tiers (segmental vs. tonal) and set up five conditions accordingly. To bring the reader’s attention to the divergent information, we named the conditions of Experiment 3 by the component that diverged. This differed from Experiment 1 and 2, in which the conditions are named by the shared component. The five conditions are early segmental (diverging) condition, which has word pairs with early diverging segmental information; early tonal (diverging) condition, which has word pairs with early diverging tonal information; late segmental (diverging) condition, which has word pairs with late diverging segmental information; late tonal (diverging) condition, which has word pairs with late diverging tonal information; and baseline condition, which has unrelated word pairs. Participants’ gaze patterns across conditions would effectively inform us when and how Mandarin listeners use tonal and segmental information during SWR. We hypothesised that if both lexical tone and segment are utilised during online lexical processing, phonological competition effects (indexed by more eye fixations towards targets and fewer eye fixations towards competitors compared with the baseline condition) should be observed for both tonal and segmental diverging word pairs. In case the utilisation of segmental and tonal cues is time-locked to the presence of the cues, significant differences between the early and late diverging word pairs’ competition effects should be observed.

Figure 4. Estimated smooths difference of competitor fixations between Experiment 1 & 2 (1500 ms & 200 ms). (A) Smooths difference between preview times in the baseline condition; (B) Smooths difference between preview times for competitor fixations in the segmental syllable condition; (C) Smooths difference of cohort competitor fixations between preview times; (D) Smooths difference of rhyme competitor fixations between preview times; E. Smooths difference of tonal competitor fixations between preview times. The pointwise 95%-confidence intervals are shown by shaded bands. The green background indicates that the shaded confidence band is significantly different from zero.
Specifically, the late conditions should show larger cumulative competition effects than the early ones regardless of the information tier.

Unlike in Experiment 1 and 2, we used Chinese characters as visual displays in Experiment 3. Huettig and McQueen (2007) have reported a stronger phonological competition effect in Dutch when using printed words than pictures as a visual display. They suggested that using printed words in visual world paradigm is “more sensitive to phonological manipulations than the version using pictures”. We tapped into how subtle phonetic cues are used during auditory word recognition in Experiment 3. Suppose Chinese characters serves the same function as alphabetic scripts in the visual world paradigm. Using Chinese characters should help to zoom into subtle phonological competition effects that otherwise may not be easily detected. Another benefit of using printed words is that it makes our experiment feasible. This is because we adopted a between-subject design (i.e. participants of Experiment 1 and 2 also participated in Experiment 3). To avoid using the same stimuli, we had only a limited number of picturable nouns available as stimuli, making the design practically infeasible.

Methods
Participants
37 native Mandarin speakers (mean age: 19, standard deviation: 1.5; 21 females, 16 males) who participated in Experiment 1 or 2 also participated in Experiment 3. The order of participating in Experiment 1 and 3 (or Experiment 2 and 3) was counterbalanced.

Stimuli
A total of 96 Mandarin monosyllable words, grouped into two sets of stimuli, were used in Experiment 3 (see supplementary material). One set consists of 24 tonal pairs, of which one word differs from the other only in the lexical tone; the other set consists of segmental pairs, of which one word differs from the other only in the segment. Based on the POD, both groups were further classified as with early POD or late POD. The early tonal POD word pairs either had a nasal onset or no onset, so the entire syllable carries tonal information from the beginning of the syllable. Their lexical tones contrast with each other from the beginning of their tonal pitch contours (e.g. Tone1, high-level tone vs. Tone3, low rising-falling dipping tone; Tone 4, high falling tone vs. Tone 3, low rising-falling dipping tone). The late tonal POD pairs have obstruent onsets that do not carry tonal information. Lexical tones either both start low (e.g. Tone 2, rising tone vs. Tone 3, rising-falling dipping tone) or both start high (Tone 1, high-level tone vs. Tone 4, high-falling tone), so their tonal divergence point occurs late. For word pairs diverging in segmental information, the early POD word pairs share the same onset which contains only one segment and diverges in rime (e.g. pa2 – ping2), while the late POD word pairs share not only the one onset segment but also the following glide (e.g. xue3 – xuan3), which has been analyzed either as part of the onset or part of the rhyme (for further discussion on the treatment of glides, see e.g. Chen & Gussenhoven, 2015). Table 1 provides sample stimuli in Pinyin, an alphabetic writing system of Standard Chinese.

As discussed earlier, we used printed characters instead of real object pictures as a visual display in this experiment. Word frequency, as computed with SUBTLEX-CH (Cai & Brysbaert, 2010), was balanced across conditions $[F(3, 92) = 1.871, p > 0.1]$. Also, the number of components and strokes of the characters were controlled across conditions (Strokes $[F(3, 92) = 0.538, p > 0.5]$; Component $[F(3, 92) = 1.564, p > 0.1]$). All stimuli were recorded in the Phonetics Lab of Leiden University through a Sennheiser MKH416T microphone (44.1 kHz, 16 bit). The speaker is a male native Standard Chinese speaker who was born and grew up in Beijing. Each word was read four times in isolation using a randomised list. One token of each word was chosen based on its clarity and then normalised for intensity at 70 dB.

Procedure and design
The procedure in Experiment 3 was the same as that in Experiment 1. Each of the early segmental, early tonal, late segmental, late tonal, and baseline conditions have 24 trials. Another 72 trials were included as fillers in which no phonological-related items were presented. In total, there were 192 trials distributed in four blocks. As in Experiment 1 & 2, the order of blocks was counterbalanced between participants.

Results
Behavioural data
Reaction time and response accuracy for mouse click are shown in Table 10. The same criteria used for data

| Information | Timing | Reaction (SE) | Percent Accuracy (SE) |
|-------------|--------|---------------|-----------------------|
| Segmental   | Early  | 1103 (26)     | 97.6 (1)              |
|             | Late   | 1256 (38.2)   | 96.5 (0.8)            |
| Tonal       | Early  | 1160 (25.9)   | 97.5 (1.5)            |
|             | Late   | 1202 (31.7)   | 97.8 (1.46)           |
| Baseline    |        | 1086 (26)     | 99.2 (0.8)            |

Standard Errors are in parentheses.
analysis (as described in Experiment 1) were adopted. To enable better comparison with baseline, we used experimental conditions with five levels (i.e. early segmental, late segmental, early tonal, late tonal, and baseline) as fixed effects to model fixation proportions to target. The best-fit reaction time model included fixed effects of experimental conditions ($\chi^2(4) = 41.802, p < 0.001$), by-subject and by-item random intercepts, and by-subject random slope for testing conditions. Post-hoc analysis showed that participants’ reaction times in all critical conditions were significantly longer than baseline (early segmental: $p < 0.05$; late segmental, early tonal, late tonal: $p < 0.001$). Furthermore, the reaction time in the early segmental condition was significantly shorter than that in the late segmental condition ($p < 0.001$). So did the early tonal condition, compared with the late tonal condition ($p < 0.05$). There was no significant improvement in model fit for the best-fit accuracy model after adding fixed effects of the experiment conditions ($\chi^2(4) = 7.627, p = 0.106$), suggesting that response accuracy did not differ across experimental conditions.

**Eye movement data**

**Looks to target.** Generalised additive modelling (Wood, 2017) was also employed to model participants’ eye fixations. The same modelling procedure as in Experiment 1 and 2 was applied. The resulting model of target fixations includes a fixed effect for condition, a smooth over time for each level of condition (i.e. the baseline, early segmental, late segmental, early tonal, and late tonal conditions), and a non-linear random smooth of subject by condition. This final model explains 97.8% of the deviance. Pairwise comparisons between each level were conducted with ordered factors. The final model explains 96% of the deviance. The estimated smooths for all levels of conditions are visualised in Figure 5A. The estimated smooths for all conditions are summarised in Table 11. The estimated smooths for all conditions are visualised in Figure 5A.

Smooth term differences between each level are plotted in Figure 6. As shown in Figure 6, compared with the baseline condition, there are fewer target fixations in early segmental, late segmental, early tonal, and late tonal conditions about 250 ms after the auditory stimuli onset. However, according to the estimated parameters of GAMMs (see Table 11), all the differences against baseline were not statistically different.

As for the effect of POD (early vs. late), although proportions of target fixations in both early conditions were slightly higher than that of late conditions (Figure 6), the differences were not statistically different. The estimated parameters of GAMMs (see Table 11) indicate no significant differences between the information tiers (tonal vs. segmental) either.

Overall, target fixation proportions show two trends: (1) target pictures were less frequently looked at in the early segmental, late segmental, early tonal, and late tonal conditions than in baseline; (2) the late POD conditions generally has a larger effect on target fixations than the early conditions. Although these trends are observable with visual inspection, our GAMM analyses did not yield any statistical significance.

**Looks to competitors.** The final model of competitor fixations includes a fixed effect for condition, a smooth over time for each level of condition, and a non-linear random smooth of subject by condition. Pairwise comparisons between each level were conducted with ordered factors. The final model explains 96% of the deviance. The estimated smooths for all levels of conditions are visualised in Figure 5B. The estimates for the parametric and smooth terms are summarised in Table 12.

As we can see from Figure 7, all experimental competitors attract more fixations than the baseline condition after 250 ms post auditory stimuli onset. As GAMMs parameters indicate (see Table 10), model fits of the late segmental, early tonal and late tonal

| Table 11. GMM analysis of targets’ fixation proportions in Experiment 3. |
| -------- | --- | --- | --- | --- |
| s(Time) | -0.096 | 0.129 | -0.746 | 1.000 |
| s(Time):Early Segmental – Baseline | -0.047 | 0.118 | -0.400 | 1.000 |
| s(Time):Late Segmental – Baseline | 3.995 | 4.335 | 3.865 | 0.053 |
| s(Time):Early Tonal – Baseline | 2.825 | 3.062 | 0.627 | 1.000 |
| s(Time):Late Tonal – Baseline | 3.617 | 3.927 | 3.168 | 0.268 |
| s(Time):Early Tonal – Early Segmental | 1.000 | 1.000 | 1.857 | 1.000 |
| s(Time):Early Tonal – Late Segmental | 3.048 | 3.307 | 2.207 | 1.000 |
| s(Time):Early Tonal – Late Segmental | 2.410 | 2.596 | 1.524 | 1.000 |
| s(Time):Early Tonal – Early Segmental | 3.047 | 3.307 | 2.145 | 1.000 |
| s(Time):Early Tonal – Late Segmental | 2.409 | 2.595 | 0.615 | 1.000 |
| s(Time):Late Tonal – Early Segmental | 1.453 | 1.520 | 0.186 | 1.000 |
| s(Time):Late Tonal – Late Segmental | 1.453 | 1.520 | 0.186 | 1.000 |

* *** p-value
Figure 5. Estimated smooths for all conditions in Experiment 3. (A) Smooths of target fixations for all conditions; (B) Smooths of competitor fixations for all conditions. The pointwise 95%-confidence intervals are shown by shaded bands.

Figure 6. Estimated smooths difference between experimental conditions for target fixations in Experiment 3. (A) Smooths difference between the early tonal condition and the baseline condition; (B) Smooths difference between the early segmental condition and the baseline condition; (C) Smooths difference between the late tonal condition and the baseline condition; (D) Smooths difference between the late segmental condition and the baseline condition; (E) Smooths difference between the early tonal condition and the late tonal condition; (F) Smooths difference between early segmental condition and the late segmental condition. The pointwise 95%-confidence intervals are shown by shaded bands. The green background indicates that the shaded confidence band is significantly different from zero.
conditions were significantly different from the baseline condition in intercept ($p < 0.01$; $p < 0.05$; $p < 0.001$). As for differences in the estimated smooth terms, only late segmental and late tonal conditions were significantly different from the baseline condition ($p < 0.001$; $p < 0.001$). The early segmental condition did not significantly differ from the baseline in either intercept or smooth term.

As for the effect of POD (point of divergence in segmental and tonal information), model fits of the early and late segmental conditions significantly differed in intercept ($p < 0.01$) and smooth term ($p < 0.001$), while the early and late tonal conditions significantly differed in smooth term ($p < 0.001$). As Figure 7E and F show, participants looked more frequently at the late segmental and tonal competitors than the early competitors. Moreover, participants’ competitor fixations in the early segmental condition seem to be overall less frequent than that in the early tonal condition, but the difference was not statistically significant. For word pairs of late POD, segmental competitors had a slightly higher proportion of eye fixations than the tonal competitors at the late time window. The difference was again not statistically significant.

Overall, competitors’ eye fixations confirmed the general trends observed with target fixations. First, both tonal (early and late) and segmental (late) competitors attract participants visual attention; Second, POD affects the proportion of eye-fixations on competitors regardless of the information tier: the later the information diverges, the more frequent eye-fixations on the competitors.

**Discussion**

Through careful control of the phonological overlap and timing of acoustic information (i.e. the Point of Divergence; POD), we confirmed that phonologically similar words, be it segmental or tonal, drew participants’ visual attention more than unrelated distractors. This indicates that similar words with divergent phonemes or lexical tones are co-activated and compete for recognition in Mandarin lexical access. Moreover, the size and the time course of the phonological competition effects were modulated by the timing of the point of divergence in the acoustic signal. The later the information disambiguates, the larger the competition effects. These findings lend solid support to the view that Mandarin listeners use both tonal and segmental information incrementally during Mandarin SWR.

**General discussion**

The present study examined the role of segmental syllable, sub-syllabic constituents, and lexical tone in Mandarin SWR and the time course of how segmental phoneme and suprasegmental lexical tone are utilised during lexical processing. Our findings suggest that segmental syllable plays a dominant role in Mandarin lexical processing while the effects of onset, rhyme, and lexical tone are more subtle and variable. Moreover, when all else is controlled, both segmental information and suprasegmental information can be used to constrain word competition as soon as their respective acoustic cues are present. Results of the three experiments have implications for both models of word recognition in tonal languages and methodological issues of using the visual world paradigm for SWR.

More specifically, Experiments 1 & 2 examined the relative contribution of segmental syllable, onset, rhyme, and lexical tone. We investigated when and to what extent participants’ gazes are distracted by the presence of phonological competitors in recognising Mandarin monosyllabic words. While Experiment 1 allowed participants to preview the pictures for 1500 ms before listening to the target, Experiment 2 only allowed a short preview of 200 ms. Both experiments consistently found that only competitors with the same segmental syllable significantly distract participants’ visual attention towards the target word. Cohort competitors (with onset and lexical tone overlaps), rhyme competitors (with rhyme and lexical tone overlaps),
overlaps), or tonal competitors (with lexical tone overlap) do not introduce more fixations than unrelated distractors.

Results of Experiments 1 & 2 thus replicate Malins and Joanisse’s (2010) findings of the segmental (syllable) competitors and rhyme competitors, but not of the cohort and tonal competitors. Note that we have followed Zou (2017) and defined the cohort competitors as having only onset and lexical tone overlap with the targets. Our results confirmed Zou’s (2017) finding of the null cohort effect. The robust cohort competition effect reported in Malins and Joanisse (2010) is likely due to the more extended overlap beyond a single onset phoneme (e.g. hua1 “flower”- hui1 “grey”; tu3 “dirt”- tui3 “leg”). Furthermore, by assigning tonal competitors an equal number of occurrences as other types of competitors and thereby avoiding a potential familiarity effect within the experiment, our results confirmed the lack of tonal competition effect as reported in Zou (2017). This suggests that lexical tone alone does not have an impact on Mandarin SWR.

With a short preview time of 200 ms, Experiment 2 replicated the results of Experiment 1. This finding differs from that reported in Huetting and McQueen (2007), which showed a lack of phonological competitor activation with a short preview (200 ms). Huetting and McQueen (2007) showed that 200 ms may be insufficient for Dutch participants to pre-activate the object names and consequently bias phonologically guided eye-fixations. With a series of eye-tracking experiments, Apfelbaum et al. (2021) have argued that phonological competition is not contingent upon pre-naming or pre-activating names during the preview. Instead, the preview allows participants some time to recognise visual objects so that their eye movements can better reflect lexical processing. Note that one particular design of Huetting and McQueen (2007) is that three types of competitors, namely the visual, semantic, and phonological competitors, were all presented in one display. It is possible that when the preview is short, the visual search is delayed such that listeners may have fixated first and primarily on the visual and
Our results are consistent with Zou (2017) and Apfelbaum et al., (2021), both of which found evidence of phonological competition even without preview. These studies suggest that the length differences in preview time (in our case a difference of 200 ms vs. 1500 ms) do not influence the general pattern of phonological competition in Mandarin lexical access. Although the length of preview time is not a determining factor in phonological competition, it does influence how participants distribute their visual attention. We found that, with a shorter preview, participants located the target picture faster with fewer fixations and showed different fixation patterns for phonological competitors when the preview was short. For example, there were slightly more frequent fixations on the rhyme competitors at a later processing stage than in Experiment 1. Future research needed to fully understand the effect of preview time length on phonological competition in SWR.

Experiment 3 zoomed further into the time course of SWR and in particular, listeners’ sensitivity to the acoustic details of segmental and tonal information. Word pairs (target and competitor) with divergent segmental information or tonal information were contrasted. With all else controlled, we were interested in whether the lexical co-activation and competition effect is modulated by the timing of the POD (i.e. the point of information divergence; early vs. late) along both the segmental and tonal dimensions. Results showed that, while both early and late tonal competitors significantly attracted participants’ visual attention, the late tonal competitors (which have the same segments with the target but a tonal pitch contour disambiguated later from the target tone) exhibited a significantly larger effect than the early tonal competitors (which share segments and with a pitch contour disambiguated earlier from the target tone). The results are consistent with the findings reported in Qin (2017), confirming that lexical tone can be used early to constrain word activation before it is recognized. Segmental competitors only exhibited a significant effect when the segmental information diverges late (i.e. with target and competitor sharing onset, glide, and tone) but not early (i.e. with target and competitor sharing the first onset phoneme and tone). No statistically significant difference was found between either early or late tonal and segmental conditions, suggesting comparable weighting of tonal and segmental information during lexical access. For both tonal and segmental conditions, we found that the competition effects were less persistent and weaker when the information diverges early. Taken together, our results provide further evidence against the view of holistic processing in Mandarin lexical access. Both tonal and segment phonemic cues are incrementally processed as soon as they arrive and with more informative cues (be it tone or segment) accumulated, listeners showed a higher level of confidence in arriving at the right percept of the incoming stimuli.

In sum, the results of Experiments 1 & 2 indicate an advantageous role of segmental syllable over onset, rhyme, and lexical tone in activating word candidates. While Experiment 3 shows that, both tonal and segmental information can be used incrementally to constrain word candidates’ activation during the process of Mandarin SWR.

How to model such effects? Previous studies have proposed several accounts of SWR in tonal languages (Gao et al., 2019; Malins & Joanisse, 2012b; Ye & Connine, 1999; Shuai & Malins, 2017; Tong et al., 2014; Yue, 2016; Zhao et al., 2011). The classic TRACE model (McClelland & Elman, 1986) posits a three-layer (word, phoneme, feature) architecture and bidirectional interconnections between layers. Existing models of SWR in tonal languages such as Mandarin typically add the “toneme” (Malins & Joanisse, 2012b; Ye & Connine, 1999; Zhao et al., 2011) or “tone” node (Gao et al., 2019; Yue, 2016) as the representation of lexical tone.

One disagreement among these existing models is whether an extra level of (tonal) syllable (Zhao et al., 2011) or segmental syllable (Gao et al., 2019; Yue, 2016) is necessary. The syllable node in Zhao et al. (2011) incorporates syllabic morpheme (which includes both segmental syllable and tone) as a phonological representation to store morphemic syllables. By “hiding” phonemes and tonemes, the Reverse Accessing Model (RAM) in Gao et al. (2019) treats atonal syllable (i.e. segmental syllable) as “the earliest and smallest unit of phonological information immediately available for mental operations”. In line with the proposal of RAM (Gao et al., 2019), our results also argue for the inclusion of segmental syllable at the sub-lexical level to account for its advantageous role in Mandarin lexical access. However, we remain skeptical about “hidding” phonemes and tonemes. The RAM proposes that tones and segmental phonemes are “hidden”, i.e. can only be accessed when the information at the (atonal) syllable level is insufficient for the task at hand. This assumption accounts for the findings of the speeded discrimination tasks in Gao et al. (2019), which led them to propose that it is easier for participants to make identical/different judgments on (segmental) syllables than phonemes or tones because the latter requires re-activation of the phonemic and tonal information as a mental replay.
Following Gao et al. (2019), we would expect that within the visual world paradigm, only words with the same or similar segmental syllable would be co-activated and compete for recognition. It has been, however, repeatedly observed that despite the robust competition effect of segmental syllables, effects of sub-syllabic components are present (e.g. late segmental competition effect in our Experiment 3; Malins & Joanisse, 2010; Zou, 2017). We take them as evidence that all information is maintained without “hidden” structures and utilized incrementally to aid SWR, which agrees more with the assumption of the TRACE model.

Another disagreement in the current models of spoken word recognition with lexical tone is whether the processing of segment and tone are integrated (e.g. the TTRACE model; Tong et al., 2014) or separated (e.g. the TRACE-T model; Shuai & Malins, 2017). Zou et al. (2017) showed that native Mandarin listeners found it difficult to attend only to one of these two tiers of information, suggesting that at a certain level of processing, segmental and tonal information are integrated. Furthermore, it is also relatively easier for them to attend only to segments (compared to only to tone), suggesting an asymmetrical relationship between the processing of these two tiers and, therefore the need for separate processing at other levels. There are also data from neural processing to shed light on this issue. Choi et al. (2017) examined the pre-attentive and phonological perceptual integration of vowels and tones in Cantonese using the oddball paradigm; the mismatch negativity (MMN) suggests the integration of vowel and tone processing at the phonological level. With the violation paradigm, Zou et al. (2020) reported different ERP effects for the rhyme and tone violation conditions, indicating different roles of tone and vowel at different stages of speech processing. Our study was not designed to explicitly test the integration or separation of segment and tone processing. However, in Experiment 3, we do see clear substantial time course differences between the tonal and segmental diverging conditions. For example, the tonal condition had a significant early competition effect while the early segmental condition did not. Also, considering previous findings of tonal and segmental processing differences in terms of timing (e.g. Ye & Connine, 1999), speed (e.g. Connell, 2017), and relative weighting (e.g. Zou et al., 2020), it is more prudent for us to posit that at some levels of processing, tone and segments are processed independently, rather than integrated and holistically throughout the SWR process.

Based on our findings and data reported in the literature, we suggest a revised TRACE model for Mandarin SWR with a four-layer structure: syllable (i.e. segmental syllable and tone), segmental syllable, phonemes, and toneme, as well as their respective features. The extra level of segmental syllable accounts for the overall larger and more stable phonological competition effects of segmental syllable over a combination of sub-lexical phonological components (e.g. onset plus tone; rhyme plus tone) during Mandarin SWR. Moreover, with independent representations of phonemes and tonemes, both phonemic and tonal information can be used to resolve phonological competition when the context introduces enhanced sensitivity to such phonological information.

Having an extra unit of segmental syllable also echoes findings during online speech production. With classic paradigms such as implicit priming (Meyer, 1991), previous studies on Mandarin word production found effects of the atonal syllable (i.e. segmental syllable) but not of the initial onsets, which clearly differed from Indo-European languages (e.g. Chen & Chen, 2013; Chen et al., 2003, 2016; Wang et al., 2018). O’Seaghdha et al. (2010) therefore proposed that, whereas the proximate phonological encoding units in Indo-European languages are phonemic segments, it is the atonal syllable that is proximate in Mandarin. Roelofs (2015) incorporated the proximate unit principle into the WEAVER+ Model (Levelt et al., 1999). Computational simulation results successfully explained the divergent findings between Mandarin and English, confirming cross-language differences in terms of the phonological planning units. Also adopting the proximate unit principle, Alderete et al., (2019) proposed a two-stage model for tone word production that not only incorporates the primary role of atonal syllables but also an early selection process of lexical tone, similar to the model structure we proposed for spoken word recognition. To further explore the relationship between tone word production and recognition, future studies are needed.

Two potential limitations of our findings of this study are to be noted. First, as we mentioned earlier, according to recent statistical advice (Brysbaert, 2019; Brysbaert & Stevens, 2018), our sample sizes in the three experiments are relatively small, which might have reduced the power and increased the margin of error. Second, due to the difficulty of finding sufficient items, we followed the design of Malins and Joanisse (2010) in using the targets repeatedly without dividing them into counter-balancing lists. How this practice may affect the data is still unclear, but we believe that it is important to note this limitation and also take a cautionary note in interpreting the results.

In summary, results of this study confirm that during spoken word recognition, Mandarin listeners are
sensitive to the unfolding segmental information and suprasegmental tonal information and utilise both to constrain the recognition of words which is time-locked to the available information and incremental. Unlike in English or other West-Germanic languages, segmental syllable (i.e. syllabic constituents without specifying lexical tone) plays a more advantageous role in Mandarin lexical access, in comparison to sub-syllabic constituents. This study has contributed more empirical data to help adjudicate current and future models of spoken word recognition in lexical tone languages and to shed new insights into the universal and diverse patterns of spoken word recognition across languages.

Notes
1. For the ease of comparison to the existing findings (i.e. Malins & Joanisse, 2010; Zou, 2017), We have also analyzed the eye-tracking data with the growth curve analysis (GCA; Mirman, 2014). The results converge for most analyses except for Experiment 3, in which the results of GAMM are more conservative. Given the discussion in Huang and Snedeker (2020), we report our results based on the GAMM analysis.
2. While Table 6 shows a significant difference between the segmental syllable and baseline condition, the plot of the difference smooth in Figure 2B does not show any difference over time. This discrepancy is most likely due to the use of different R packages ("mgcv" for the model summary; "tidymv" for visual inspection; Coretta, 2021). Given that model summary using ordered factors as significance testing are generally more reliable than visual inspections in GAMM (Söskuthy, 2021), we refer readers to the model summary as the final results of significance testing.
3. Note that the Pinyin system is designed for spelling out the Standard Chinese syllables, not for phonetic transcription or phonological analysis as the international phonetic alphabets would imply.
4. While Table 11 shows no significant difference between conditions, the plots of estimated smooth in Figure 6 does show some significant difference over time. The results shown in Table 11 are more conservative because the p-values were corrected with the Holm–Bonferroni method (Holm, 1979) to avoid family-wise errors.

Acknowledgments
We thank Dr. Xuhai Chen for helping with data collection, and Dr. Marjolein Fokkema for sharing her statistical expertise during our consultation. We thank the reviewers for pointing out these limitations.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research was supported by the China Scholarship Council (201707110012) and the nederlandse organisatie voor wetenschappelijk onderzoek (VI.C.181.040).

ORCID
Yiya Chen http://orcid.org/0000-0003-0009-5882

References
Alderete, J., Chan, Q., & Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. Cognition, 191, 103952. https://doi.org/10.1016/j.cognition.2019.04.021
Alloppenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38(4), 419–439. https://doi.org/10.1006/jmla.1997.2558
Apfelbaum, K. S., Klein-Packard, J., & McMurray, B. (2021). The pictures which shall not be named: Empirical support for benefits of preview in the Visual World Paradigm. Journal of Memory and Language, 121, 104279. https://doi.org/10.1016/j.jml.2021.104279
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01
Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of non-independent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychological Methods, 23(3), 389–411. https://doi.org/10.1037/met0000159
Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. Journal of Cognition, 2(1). https://doi.org/10.5334/joc.72
Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1(1), 1–20. https://doi.org/10.5334/joc.10
Burtnach, D., Kim, J., Davis, C., Ciocca, V., Schoknecht, C., Kassissopa, B., & Lukanevanawin, S. (2011). Are tones phones? Journal of Experimental Child Psychology, 108(4), 693–712. https://doi.org/10.1016/j.jecp.2010.07.008
Cai, Q., & Brysbaert, M. (2010). SUBTLEX—CH: Chinese word and character frequencies based on film subtitles. PLOS ONE, 5(6), e10729. https://doi.org/10.1371/journal.pone.0010729
Chen, J. Y., Lin, W. C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese speech production. Zhonghua Xinxili Xuekan, 45(1), 107–120.
Chen, J.-Y., O’Séaghdha, P. G., & Chen, T.-M. (2016). The primacy of abstract syllables in Chinese word production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(5), 825–836. https://doi.org/10.1037/a0039911
Marder, E., & McCarley, R. W. (2000). Paradoxical regulation of thalamic neuronal excitability by excitatory and inhibitory afferents. Science, 289(5475), 829–832. https://doi.org/10.1126/science.289.5475.829

Marder, E., & Pinsky, P. (2014). Dynamic and plastic features of neuronal excitability. Annual Review of Neuroscience, 37, 345–366. https://doi.org/10.1146/annurev-neuro-061913-084133

Marder, E., & Zadra, A. (2009). Evolutionary and phylogenetic perspectives on neural oscillations. Frontiers in Neural Circuits, 3, 1–14. https://doi.org/10.3389/neucirc.2009.00002

Marder, E., & Yarar, M. (2014). How neural oscillations shape behavior. Trends in Neurosciences, 37(5), 224–233. https://doi.org/10.1016/j.tins.2014.03.001

Marder, E., & Yarar, M. (2017). The evolution of neural oscillations. Trends in Neurosciences, 40(1), 66–76. https://doi.org/10.1016/j.tins.2016.09.005

Marder, E., & Yarar, M. (2018). The evolution of neuronal oscillations. Current Opinion in Neurobiology, 53, 1–8. https://doi.org/10.1016/j.conb.2018.03.011

Marder, E., & Yarar, M. (2019). The role of neural oscillations in behavior. Current Opinion in Neurobiology, 57, 12–19. https://doi.org/10.1016/j.conb.2019.02.003

Marder, E., & Yarar, M. (2020). The evolution of neural oscillations. Trends in Neurosciences, 43(9), 737–748. https://doi.org/10.1016/j.tins.2020.04.006

Marder, E., & Yarar, M. (2021). The role of neural oscillations in behavior. Current Opinion in Neurobiology, 66, 103–110. https://doi.org/10.1016/j.conb.2020.10.008
Acoustic cue *variability* affects eye movement movement *during non-native speech perception: A GAMM model*. The 14th International Conference on Auditory-Visual Speech Processing (AVSP2017), Stockholm, Sweden.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, 90, 103–125. https://doi.org/10.1016/j.jml.2016.03.005

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

O’Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, 115(2), 282–302. https://doi.org/10.1016/j.cognition.2010.01.001

Porretta, V., Kyröläinen, A.-J., van Rij, J., & Järvikivi, J. (2018). Visual word paradigm data: From preprocessing to non-linear time-course analysis. In I. Czarnowski, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent decision technologies 2017* (pp. 268–277). Springer International Publishing. https://doi.org/10.1007/978-3-319-59424-8_25

Qin, Z. (2017). How native Chinese listeners and second-language Chinese learners process tones in word recognition: An eye-tracking study [Doctoral dissertation]. University of Kansas. https://kus.kuscholarworks.ku.edu/handle/1808/26474

R Core Team. (2017). *R: A language and environment for statistical computing* [Manual]. https://www.R-project.org/

Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From Germanic languages to Mandarin Chinese and Japanese. *Japanese Psychological Research*, 57(1), 22–37. https://doi.org/10.1111/jpr.12050

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90(1), 51–89. https://doi.org/10.1016/S0010-0277(03)00139-2

Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., & Chen, H.-C. (2005). Brain responses to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience*, 17(1), 1–12. https://doi.org/10.1162/0898929052880057

Sereno, J. A., & Lee, H. (2010). Temporal changes in Mandarin tone due to speaking rate variation. *The Journal of the Acoustical Society of America*, 127(3), 2023–2023. https://doi.org/10.1121/1.3385295

Sereno, J. A., & Lee, H. (2015). The contribution of segmental and tonal information in Mandarin Spoken word processing. *Language and Speech*, 58(2), 131–151. https://doi.org/10.1177/0023830914522956

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in JTRACE: A simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, 49(1), 230–241. https://doi.org/10.3758/s13428-015-0690-0

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, 84, 101017. https://doi.org/10.1016/j.wocn.2020.101017

Taf, M., & Chen, H.-C. (1992). Judging homophony in Chinese: The influence of tones. *Acoustical Science and Technology*, 90, 151–172.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634. https://doi.org/10.1126/science.777863

Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes*, 23(5), 689–708. https://doi.org/10.1080/01690960701728261

Tong, X., McBride, C., Lee, C.-Y., Zhang, J., Shuai, L., Maurer, U., & Chung, K. K. H. (2014). Segmental and suprasegmental features in speech perception in Cantonese-speaking second graders: An ERP study. *Psychophysiology*, 51(11), 1158–1168. https://doi.org/10.1111/psyp.12257

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020). (itsadug): Interpreting time series and autocorrelated data using GAMMs (R package version 2.4) [Computer software]. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=itsadug

Wang, J., Wong, A. W.-K., & Chen, H.-C. (2018). Time course of syllabic and sub-syllabic processing in Mandarin word production: Evidence from the picture-word interference paradigm. *Psychonomic Bulletin & Review*, 25(3), 1147–1152. https://doi.org/10.3758/s13423-017-1325-5

Wood, S. N. (2017). Generalized additive models: an introduction with R, second edition. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes*, 14(5–6), 609–630. https://doi.org/10.1080/016909999386202

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

Yip, M. C. W. (2001). Phonological priming in Cantonese spoken word processing. *Psychologia*, 44(3), 223–229. https://doi.org/10.2117/psycsoc.2001.223

Yue, J. (2016). *Tone-word recognition in Mandarin Chinese: Influences of lexical-level representations* [Doctoral dissertation]. University of Groningen. https://www.rug.nl/research/portal/files/27655768

Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, 49(7), 1761–1770. https://doi.org/10.1016/j.neuropsychologia.2011.02.054

Zou, T. (2017). *Production and perception of tones by Dutch learners of Mandarin* [Doctoral dissertation]. Leiden University. https://hdl.handle.net/1887/52977

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

Zou, Y., Lui, M., & Tsang, Y.-K. (2020). The roles of lexical tone and rime during Mandarin sentence comprehension: An event-related potential study. *Neuropsychologia*, 147, 107578. https://doi.org/10.1016/j.neuropsychologia.2020.107578