# Phrasal stress in Beijing Mandarin disyllabic phrases

Hao Yi *

## 1 Introduction

Mandarin Chinese has been claimed to have phrasal stress which falls on a nonhead constituent: on the modifier in a modifier-noun phrase, and on the object in a verb-object phrase (MODN$_h$ and V$_h$OBJ, respectively; the subscript *h* stands for head, and the stressed constituent is underlined). This NONHEAD STRESS RULE is motivated by the greater information load carried by the nonhead than its syntactic head (Duanmu 2007). Taking NONHEAD STRESS RULE as a point of departure, the current study investigated Mandarin phrasal stress by using focus as a diagnostic tool. Fifteen pairs of homophonous disyllabic phrases, each consisting of a MODN$_h$ phrase and a V$_h$OBJ phrase, were elicited under both BROADFOCUS and NARROWFOCUS. The phonetic correlates of phrasal stress—duration and $F_0$—were measured. The hypotheses tested was that the nonheads have phrasal stress. Accordingly, the predictions were that (i) the nonheads will have greater duration and greater $F_0$ measurements under both focus conditions, and that (ii) the increase of duration and $F_0$ measurements on the nonheads will be greater under NARROWFOCUS. The results showed that at the phrase level, a MODN$_h$ and a homophonous V$_h$OBJ differed significantly in duration ratio and $F_0$ measurements, consistent with the interpretation that MODN$_h$ exhibits initial stress and V$_h$OBJ exhibits final stress. However, there also existed cross-stimulus variation, which is argued to be idiosyncratic rather than random. In sum, it is concluded that NONHEAD STRESS RULE, despite being a weak universal, is an important component to Mandarin Prosody, and underlies the contrastive stress patterns of MODN$_h$ and V$_h$OBJ.

## 2 Background

### 2.1 NONHEAD STRESS RULE

While primarily a tone language, Mandarin Chinese has been claimed to have phrasal stress. The distribution of stress, according to Duanmu 2007, is governed by NONHEAD STRESS RULE: phrasal stress falls on the nonhead constituent of a phrase.

(1) NONHEAD STRESS RULE (Duanmu 2007)

> In the syntactic structure [X XP] (or [XP X]), where X is the syntactic head and XP the syntactic nonhead, XP should be stressed.

Therefore, stress falls on the object in a verb-object phrase V$_h$OBJ, and on the modifier (adjective or noun) in a modifier-noun phrase MODN$_h$.

| | | Example | Gloss | Structure |
|---|---|---|---|---|
| (2) | a. | [ʂɑŋ]-1 [ɻən]-2 | *'business person'* | MODN$_h$ |
| | b. | [ʂɑŋ]-1 [ɻən]-2 | *'to hurt people'* | V$_h$OBJ |
| (3) | a. | [t$^h$əʊ]-2 [t$^h$aɪ]-1 | *'first born'* | MODN$_h$ |
| | b. | [t$^h$əʊ]-2 [t$^h$aɪ]-1 | *'to reincarnate'* | V$_h$OBJ |

*I thank Draga Zec for his guidance and feedback. I also thank Sam Tilsen and Abby Cohn for their feedback.

The homophonous pairs in (2-3) differ only in syntactic structure, as indicated by the last column: (2a) and (3a) are MODN$_h$, where the MOD modifies the N$_h$; (2b) and (3b) are V$_h$OBJ, where the V$_h$ takes action on the OBJ. According to Duanmu 2007, the MOD is more prominent than N$_h$, whereas the OBJ is more prominent than the V$_h$.

NONHEAD STRESS RULE is motivated by INFORMATION-STRESS PRINCIPLE: stress falls on syntactic nonheads because nonheads carry more information than their corresponding heads. This principle can be further accounted for by linking the information load of a form with its predictability: the more predictable a form is, the less information it carries.

(4) INFORMATION-STRESS PRINCIPLE (Duanmu 2007)

A word or phrase that carries more information than its neighbor(s) should be stressed.

INFORMATION-STRESS PRINCIPLE (as well as NONHEAD STRESS RULE) arises out of communicative effectiveness: we want the least predictable form to be conveyed with the most prominence, because it carries the most information. Therefore, the default stress should fall on the less predictable member in a disyllabic form, i.e., the MOD of a MODN$_h$, and the OBJ of a V$_h$OBJ.

### 2.2 Phonetic studies on the distribution of Mandarin phrasal stress

The distribution of phrasal stress has been addressed by several acoustic and perceptual studies, all taking NONHEAD STRESS RULE as their point of departure. Lai et al. (2010) showed in a corpus study that V$_h$OBJ patterned like MODN$_h$ in terms of both $F_0$ and duration. Specifically, the OBJ of a V$_h$OBJ was no longer than the V$_h$, and the $F_0$ measurement (the height for level tones or the slope for contour tones) for the OBJ of V$_h$OBJ was no larger than that for the V$_h$. It was found that despite the fact that V$_h$OBJ itself did not exhibit final stress, the OBJ in a V$_h$OBJ was stronger than the N$_h$ in a MODN$_h$ in that the duration was longer and the $F_0$ measurement was larger (the height of level tones was higher and the slope of contour tones was steeper). It was concluded that there was no difference between disyllabic MODN$_h$ and V$_h$OBJ in stress pattern on the basis of acoustic measurements, therefore could not confirm NONHEAD STRESS RULE.

Shen et al. (2013) investigated the acoustic correlates of contrastive stress between MODN$_h$ and V$_h$OBJ. In line with Lai et al. (2010), their results demonstrated that the OBJ of V$_h$OBJ was stronger than the N$_h$ of MODN$_h$: the duration was longer and the $F_0$ measure was larger. Moreover, Shen et al. (2013) also claimed that V$_h$OBJ exhibited final stress. That is, the absolute duration of the OBJ was longer than that of the V$_h$. However, the syllable position did not have so large an effect in MODN$_h$ as in V$_h$OBJ : no initial stress was found in MODN$_h$. Their study was partly in line with NONHEAD STRESS RULE. However, a closer look into the methodology renders their results problematic. The study used identical MODN$_h$ and V$_h$OBJ disyllabic pairs. The difference between a MODN$_h$ and a V$_h$OBJ was overt indicated by the part of speech in the carrier sentence. For example, "*I didn't not say the noun MODN$_h$ but the verb V$_h$OBJ.*" would prompt the participants to produce a MODN$_h$ and subsequently a V$_h$OBJ. The problem with such elicitation procedure is that native Mandarin speakers are normally unaware of the part of speech of the majority of the words in the lexicon because there are no overtly morphological markers in Mandarin. Therefore, with no morphological markers, a disyllabic phrase like [pjɛn]-Tone1 [hɑʊ]-Tone4 can either mean 'to number' or 'numbers', depending on the context within which it

occurs. Thus, the problematic elicitation procedure could have interfered with the purpose of their study.

Jia (2011) circumvented the ambiguity arising out of native speakers' unawareness of parts of speech while controlling the segmental influences by making use of homophones that differed in terms of morphosyntactic structure. Each homophonous pair consisted of one $\underline{\text{MOD}}\text{N}_\text{h}$ and one $\text{V}_\text{h}\underline{\text{OBJ}}$. Target phrases were elicited in isolation. Admittedly, using different words could raise the issue of frequency effects, which nonetheless cannot be entirely averted in the case of identical words, either. Although that some homophonous pairs did show different phrasal stress patterns, the majority showed final stress. It was concluded that morphosyntactic structure did not govern the allocation of phrasal stress in Mandarin, which refuted NONHEAD STRESS RULE. The problem of this study is that the stimuli were elicited in isolated form, therefore the pattern of final stress may arise out of pre-pause lengthening.

While these studies lend great insight into the distribution of Mandarin phrasal stress, none of them confirmed NONHEAD STRESS RULE. Moreover, they raise methodological problems, such as the potential complications due to the unfounded reliance on Mandarin speakers' judgement of parts of speech or due to phrase final lengthening. These concerns render the results suspicious. As a result, there is no consensus on the distribution of phrasal stress in Mandarin. Moreover, there is also no consensus on the acoustic cues of stress in a tone language like Mandarin; different acoustic cues (such as duration, various $F_0$ measurements, and intensity) have been claimed to be relevant to phrasal stress in different studies.

### 2.3 Focus as a diagnostic

Chen and Gussenhoven (2008) investigated the effect of emphasis (induced by corrective focus) on the duration and tonal implementation of monosyllabic words in Mandarin. Their results demonstrated that as the discourse context changed from NoEmphasis to Emphasis, there was a significant increase in the duration and in the $F_0$ range. Furthermore, they showed that under emphasis, lexical tones were realized with magnified $F_0$ contours, which were adapted to the durational increase of the tone-bearing syllables, therefore maximally contrasting with each other. They suggested that the effect of emphasis can be accounted for by appealing to an abstract notion of metrical prominence. The focus-introduced metrical prominence applies to the focused constituent, rendering it more prominent. However, because they only investigated monosyllabic words as the focused constituents, it is unclear how emphasis (or focus) will affect polysyllabic words/phrases with different syntactic structures, which is the task of the current study.

The current study makes use of focus to look for prosodic regularities in different stress patterns in Mandarin disyllabic words. In English, focus-introduced metrical prominence leads to the association of the nuclear pitch accent (H*L) with the focused constituent. Moreover, only the stressed syllable can coincide with the pitch accent. Given that a disyllabic word is the focused constituent, the first syllable of initial-stress words and the second syllable of final-stress words should be associated with focus-introduced metrical prominence. For instance, under narrow focus, the first syllable of *ˈproduce* (n.) and the second syllable of *proˈduce* (v.) should be associated with the nuclear pitch accent (H*L). These syllables thus exhibit greater durational increase and $F_0$ range expansion than their unstressed counterparts.

If Mandarin displays phrasal stress patterns that vary with different morphosyntactic structures, focus could function as a diagnostic tool. This study investigates the phonetic

correlates of phrasal stress in Mandarin Chinese by measuring the duration and $F_0$ contours under both BROADFOCUS and NARROWFOCUS. The effect of focus on duration and $F_0$ measurements are tested in fifteen homophonous pairs of one $\underline{\text{MOD}}\text{N}_\text{h}$ and one $\text{V}_\text{h}\underline{\text{OBJ}}$. If a homophonous pair of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ displays different phrasal stress patterns, focus-introduced prominence should apply differently: the acoustic changes of duration and $F_0$ for the stressed constituents (the MOD of $\underline{\text{MOD}}\text{N}_\text{h}$ and the OBJ of $\text{V}_\text{h}\underline{\text{OBJ}}$) will be of greater magnitude than for their unstressed counterparts (the $\text{N}_\text{h}$ of $\underline{\text{MOD}}\text{N}_\text{h}$ and the $\text{V}_\text{h}$ of $\text{V}_\text{h}\underline{\text{OBJ}}$).

## 3 Methods

### 3.1 Participants

Two female speakers (F01 and F02) and one male speaker (M01) who are native speakers of Beijing Mandarin participated in this experiment. All three speakers were born and raised in Beijing, and were graduate students at Cornell University at the time of recording. From their self-report, all three speakers are free from any speech and hearing problems. The recording took place in the sound-proof booth in Cornell Phonetics Lab in Department of Linguistics at Cornell University. The participants were naïve to the purpose of the study.

### 3.2 Test materials and data collection

The stimulus set consisted of 15 homophonous pairs of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$. Homophones were chosen because segmental variation within each minimal pair can be controlled. The stimulus set exhausted the possible combinations of four lexical tones (i.e. Tone1, Tone2, Tone3, and Tone4) in Mandarin Chinese to the exclusion of the Tone3+Tone3 combination due to third tone sandhi (see Appendix I). The target stimuli were elicited in two discourse contexts: BROADFOCUS and NARROWFOCUS.

One frame sentence was used throughout the experiment, as shown in (5). The disyllabic target stimulus is represented as $\sigma_1\ \sigma_2$. The frame sentence ensured the target stimuli would not appear in the sentence final position so as to avoid phrase final lengthening.

(5) tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 $\sigma_1\ \sigma_2$ ʂuən-4 hən-3 tuɔ-1.
*He think say* $\sigma_1\ \sigma_2$ *fluent a lot*
*'He thinks it's a lot more fluent to say* σ1 σ2*.'*

The target stimuli were elicited in three discourse contexts: BROADFOCUS, NARROWFOCUS, and PREFOCUS. The BROADFOCUS elicitations served as the baseline for the NARROWFOCUS elicitations. The PREFOCUSelicitations served as the fillers.

(i) In each trial, the speaker was first presented with a sentence in Chinese characters as the background information. The information was presented in black.

(ii) Five seconds later, the speaker was presented with a related question based on the above background information. The question was presented in red.

(iii) The speaker was instructed to answer the prompted question based on the given information.

Elicitation examples in IPA are given (11–13): (11) illustrates the BROADFOCUS elicitation, (12) illustrated the NARROWFOCUS elicitation, and (13) illustrated the PREFOCUS elici-

tation. In each type of elicitations, "I" stands for the background information, upon which the answer should be based; "Q" stands for the question, which was presented in red in the experiment; "A" stands for the intended answer with the focused constituent underlined, which was not presented in the experiment. The experimenter would ask the speakers to repeat the answer if the experimenter failed to perceive the intended focus.

(11) **BroadFocus elicitation**

**I:** tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

**Q:** tʰa-1 tɕyœ-2 tɤ-0 ʂən-2 mə-0
*He think what*
*'What does he think?'*

**A:** tʰa-1 tɕyœ-2 *t7*-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

(12) **NarrowFocus elicitation**

**I:** tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

**Q:** tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 ʂən-2 mə-0 ʂuən-4 hən-3 tuɔ-1.
*He think say what fluent a lot*
*'What does he think is fluent to say?'*

**A:** tʰa-1 tɕyœ-2 *t7*-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

(13) **PreFocus elicitation (filler)**

**I:** tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

**Q:** tʰa-1 tɕyœ-2 tɤ-0 ʂuɔ-1 σ1 σ2 ʂən-2 mə-0 hən-3 tuɔ-1.
*He think say σ1 σ2 what a lot*
*'What does he think of saying σ1 σ2?'*

**A:** tʰa-1 tɕyœ-2 *t7*-0 ʂuɔ-1 σ1 σ2 ʂuən-4 hən-3 tuɔ-1.
*He think say σ1 σ2 fluent a lot*
*'He thinks it's a lot more fluent to say σ1 σ2.'*

In every block of elicitation, there were 30 (= 15 tone combinations × 2 syntactic types) BroadFocus trials, 30 NarrowFocus trials, and 10 PreFocus trials. The trials were presented in a random order. The blocks were separated by five-minute breaks. The two female speakers (F01 and F02) each completed six blocks, while the male speaker (M01) only completed four blocks. A portion of F02's data was taken out due to later modifications to the stimuli set that applied consistently to the other two speakers (F01 and M01). Since

the elicitation procedure was the same for all three speakers, the applied changes should not affect the intended elicitations. Therefore, the rest of F02's data, together with all of F01 and M01's data, were included in the analysis. In total, 833 trials were collected.

### 3.3 Acoustic analysis and statistical analysis

#### 3.3.1 Acoustic analysis

The start and the end of both the first syllable ($\sigma_1$) and the second syllable ($\sigma_2$) were manually labelled in Praat (Boersma and Weenink 2015). Durations on the syllable level were obtained in MATLAB$^{TM}$. Since only real words stimuli were included to avoid unfamiliarity, hesitation and/or non-fluent speech, durations of the target syllables vary inherently based on the syllable structure of the segments, thus are not comparable between one another. For instance, the average duration of $\sigma_1$ is 40 ms longer than that of $\sigma_2$ in [ʂɑŋ]-1 [ɻən]-2. This does not necessarily mean both $V_h\underline{OBJ}$ and $\underline{MOD}N_h$ of the tone sequence Tone1+Tone2 exhibit initial stress. Because the duration of [ɻən] is inherently shorter than that of [ʂɑŋ], which arises out of differences in syllable structure. The same holds for [ɕi]-3 [tɕ$^h$jɛn]-2, where the duration of $\sigma_1$ is inherently shorter than that of $\sigma_2$. In order to render the duration measurements comparable among different pairs, the DURATIONRATIO—the ratio between the duration of $\sigma_1$ and the duration of $\sigma_2$—of each disyllabic phrase was derived:

$$\text{DURATIONRATIO} = \frac{Duration(\sigma_1)}{Duration(\sigma_2)}$$

Because the majority of of Tone3-bearing syllables exhibited a high level of creakiness or glottalization, $F_0$ measurements were not obtained for Tone3. Therefore, $F_0$ measurements of nine (= 3 tones × 3 tones) tone combinations were first obtained in MATLAB$^{TM}$ with Pitch Tracking Tool developed by the Cornell Phonetics Lab. The tool incorporated VOICEBOX, a third-party speech processing toolbox (Brookes 2005). For each trial, the $F_0$ values in Hz were measured every 5 ms within the disyllabic interval. The pitch tracking parameters were set differently for different speakers in order to get the best-fit $F_0$ tracks.

The $F_0$ values in Hz were then converted into semitones to reduce cross-speaker variation. The minimum frequency in Hz ($F_{0\_min}$) was searched for across all of the productions by each speaker. The following formula relates frequency in semitone ($F_{st}$) to frequency in Hz ($F_0$):

$$F_{st} = 12\log_2\left(\frac{F_0}{F_{0\_min}}\right)$$

For Tone1, a high level tone, $F_{st\_mean}$ is the mean $F_{st}$ value of the measurable part of the $F_0$ contour of a Tone1-bearing syllable, regardless of the syllable position:

$$F_{st\_mean} = \frac{1}{k}\sum^{k} F_{st},$$

where k is the number of the measurable points within the Tone1-bearing syllable.

For Tone2, a rising tone, $S_{rise}$ is the linear slope of $F_{st}$ rise:

$$S_{rise} = \frac{F_{st\_max} - F_{st\_min}}{t_{max} - t_{min}},$$

where $t_{min}$ is the time point at which the $F_{st}$ contour starts to rise; $t_{max}$ is the time point at which the $F_{st}$ contour reaches its maximum; $F_{st\_min}$ and $F_{st\_max}$ are the minimum and maximum $F_{st}$ values of the Tone2-bearing syllable, respectively.

When Tone2 follows a high-offset tone (Tone1 or Tone2), Tone2 does not start to rise until at least halfway into the tone-bearing syllable due to carryover effects. When Tone2 follows Tone4, a low-offset tone, there is a low elbow point at which $t_{min}$ could be determined.

When a Tone2-bearing $\sigma_1$ is followed by a non-obstruent-initial $\sigma_2$, $t_{max}$ could potentially locate outside the $\sigma_1$ boundary, because Tone2 often reaches its peak after the acoustic offset of the tone-bearing syllable. Therefore, in the cases of Tone2-Tone2 and Tone2-Tone4, $t_{max}$ of Tone2 on $\sigma_1$ is located in $\sigma_2$ (also in part because $\sigma_2$ has a non-obstruent onset in both cases). When the Tone2-bearing $\sigma_1$ is followed by an obstruent-initial $\sigma_2$, the $F_0$ contour is discontinued, and $t_{max}$ is determined as the time point at which $F_0$ contour reaches its maximum within the $\sigma_1$ boundary, as in the case of Tone2-Tone1. When a $\sigma_2$ bears Tone2, $t_{max}$ also locates within the $\sigma_2$ boundary, because the syllable following the target stimuli always starts with the sibilant [s], which discontinues the $F_0$ contour.

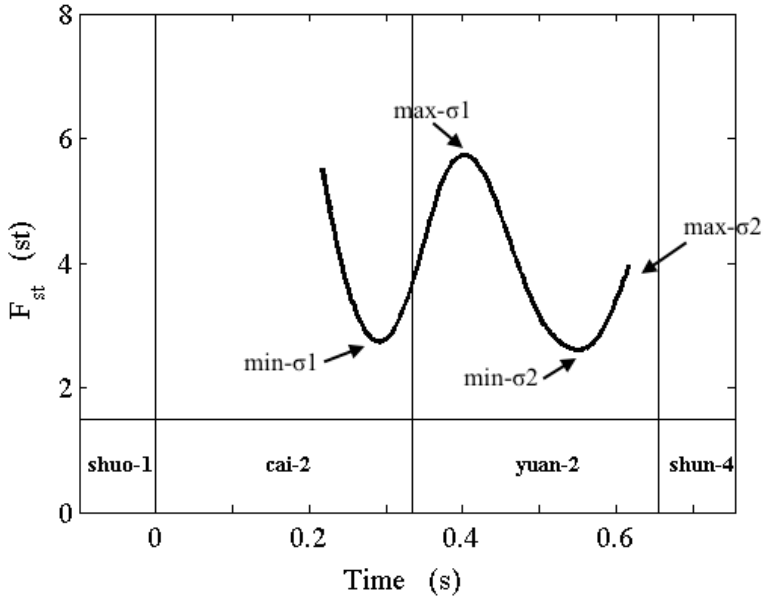



Figure 1: $F_{st}$ measurements in Tone2-Tone2 target stimulus produced by speaker F01.

As shown in Figure 1, the target stimulus [ts$^{h}$aɪ]-2 [yɛn]-2 is preceded by [ʂuɔ]-1 and followed by [ʂuən]-4. Because [ʂuɔ]-1 has a high offset, Tone2 on $\sigma_1$ does not start to rise until near the offset of $\sigma_1$, as indicated by *min-σ1*. For the same reason, Tone2 on $\sigma_2$ also does not start to rise until halfway through $\sigma_2$, as indicated by *min-σ2*. For $\sigma_1$, $t_{max}$, as indicated by *max-σ1*, occurs after the offset of $\sigma_1$, because $\sigma_2$ starts with a glide [j], a non-obstruent onset. For $\sigma_2$, $t_{max}$, as indicated by *max-σ2*, occurs before the offset of $\sigma_2$, because the following syllable [ʂun]-4 starts with a sibilant [s] that discontinues the $F_{st}$ contour.

For Tone4, a falling tone and a mirror image of Tone2, $S_{fall}$ is the linear slope of $F_{st}$ fall:

$$S_{fall} = \frac{F_{st\_max} - F_{st\_min}}{t_{min} - t_{max}},$$

where $t_{max}$ is the time point at which the $F_{st}$ contour starts to fall, whereas $t_{min}$ is the time point at which the $F_{st}$ contour reaches its minimum; $F_{st\_max}$ and $F_{st\_min}$ are the maximum and minimum $F_{st}$ values of the Tone4-bearing syllable.

### 3.3.2 Statistical analysis

The effects of morphosyntactic structure (TYPE) and discourse context (DISCOURSE) on DURATIONRATIO and $F_{st}$ measurements were tested using Linear Mixed Models (lme4 Bates et al. (2015) in R version 3.2.0). Other variables of fixed effects included tone types of both syllables ($\text{TONE}_1$ and $\text{TONE}_2$). Stimuli (STIM) and speakers (SPK) were included in the mixed model as variables of random effects.

- TYPE: morphosyntactic type. Two levels: $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$.
- DISCOURSE: discourse context. Two levels: BROADFOCUS and NARROWFOCUS.
- $\text{TONE}_1$: tone of $\sigma_1$. Four levels: Tone1, Tone2, Tone3, and Tone4.
- $\text{TONE}_2$: tone of $\sigma_2$. Four levels: Tone1, Tone2, Tone3, and Tone4.
- TONECOMB: tone combination. Fifteen levels: from Tone1+Tone1 to Tone4+Tone4, except Tone3+Tone3.
- STIM: stimulus. Thirty different items (see Appendix I).
- SPK: speaker. Three different speakers: F01, F02, and M01.

## 4 Hypotheses and predictions

**Hypothesis i:** The nonheads have phrasal stress.

**Prediction i:** a) The nonheads will have greater durations under both focus conditions. Therefore, the DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ will be larger than that of $\text{V}_\text{h}\underline{\text{OBJ}}$. b) Tonal targets of the nonheads will be realized with magnified $F_0$ contours. Therefore, the MOD and the OBJ will respectively have larger $F_{st}$ measurements than the $\text{V}_\text{h}$ and the $\text{N}_\text{h}$. Consequently, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ will exhibit different stress patterns.

**Hypothesis ii:** Under NARROWFOCUS, focus-introduced prominence applies only to the stressed constituent, leading to stronger production of the nonheads in both $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ phrases.

**Prediction ii:** Under NARROWFOCUS, the increase of both DURATIONRATIO and the $F_{st}$ measurements of the nonheads will be greater than that of their syntactic heads. Therefore, from BROADFOCUS to NARROWFOCUS, a) the DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ will increase and that of $\text{V}_\text{h}\underline{\text{OBJ}}$ will decrease; b) the $F_{st}$ measurements on the MOD and OBJ will exhibit significant increases whereas those on the $\text{N}_\text{h}$ and the $\text{V}_\text{h}$ will not exhibit significant increases or even exhibit significant decreases.

# 5 Results

## 5.1 DURATIONRATIO

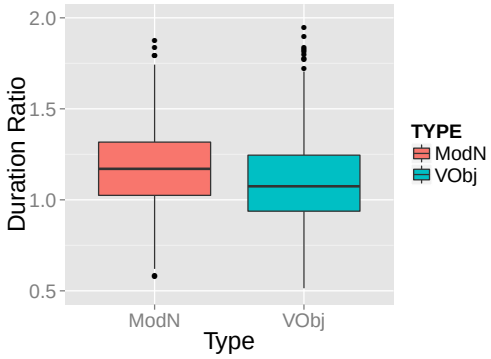



Figure 2: DURATIONRATIO of MODN$_h$ and V$_h$OBJ. Globally, the DURATIONRATIO of MODN$_h$ was larger than that of V$_h$OBJ.

Globally, there was an effect of TYPE on DURATIONRATIO. The DURATIONRATIO of MODN$_h$ was significantly larger than that of V$_h$OBJ ($t(822) = 4.3767$, $p < 0.00001$) (Figure 2).

In particular, under BROADFOCUS, the DURATIONRATIO of V$_h$OBJ was significantly larger than that of V$_h$OBJ ($t(420) = 2.3043$, $p < 0.05$); under NARROWFOCUS, the DURATIONRATIO of V$_h$OBJ was significantly larger than that of V$_h$OBJ ($t(394) = 3.9462$, $p < 0.0001$) (Figure 3). Moreover, the DURATIONRATIO difference between MODN$_h$ and V$_h$OBJ was more pronounced under NARROWFOCUS (0.097) than under BROADFOCUS (0.057).

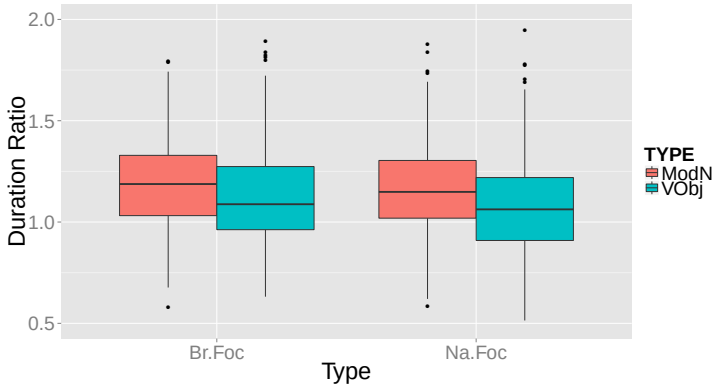



Figure 3: DURATIONRATIO of MODN$_h$ and V$_h$OBJ, grouped by DISCOURSE (BROADFOCUS and NARROWFOCUS ). The DURATIONRATIO difference between MODN$_h$ and V$_h$OBJ was more pronounced under NARROWFOCUS than under BROADFOCUS.

Figure 4 shows the DURATIONRATIO grouped by SPK. While there were some consistent global patterns indicative of the TYPE effect, there also existed speaker-specific patterns. Under NARROWFOCUS, both female speakers (F01 and F02) produced MODN$_h$ with significantly larger DURATIONRATIO than V$_h$OBJ ($t(167) = 3.6001$, $p < 0.001$; $t(113) = 2.2586$, $p < 0.01$). However, the male speaker (M01) did not differentiate between MODN$_h$ and V$_h$OBJ with DURATIONRATIO under NARROWFOCUS ($t(106) = 0.4983$, $p > 0.05$). Out of three

speakers, only F01 differentiated between $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ with DURATIONRATIO under BROADFOCUS ($t(173) = 3.4725$, $p < 0.01$).

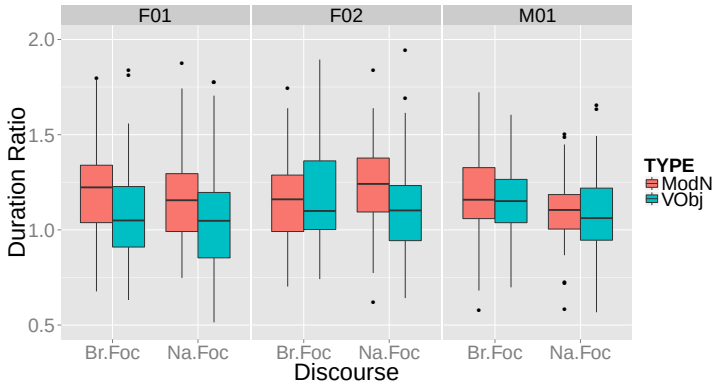



Figure 4: DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ grouped by SPK. Global TYPE effect was observed across speakers; with DURATIONRATIO, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS than under BROADFOCUS, though there existed cross-speaker variation.

Figure 5 shows the DURATIONRATIO grouped by tone combination ($\text{TONE}_1$ + $\text{TONE}_2$). Consistent with the previous results, for the majority of the tone combinations, the global patterns were: 1) the DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ was larger than that of $\text{V}_\text{h}\underline{\text{OBJ}}$; 2) $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS than under BROADFOCUS. However, there were also anomalies: $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were not differentiated under either DISCOURSE condition in terms of DURATIONRATIO (e.g., Tone1+Tone1), and the DURATIONRATIO difference was larger under BROADFOCUS than under NARROWFOCUS (e.g. Tone2+Tone3).

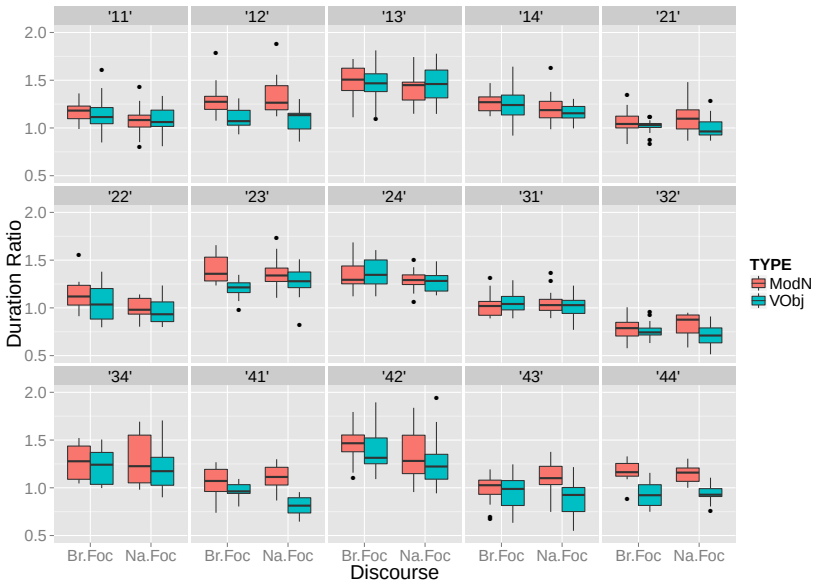



Figure 5: DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ grouped by tone combination ($\text{TONE}_1$ + $\text{TONE}_2$). Global TYPE effect was observed for the majority of the tone combinations; with DURATIONRATIO, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS than under BROADFOCUS, though there existed variation across tone combinations.

The above results shown in Figures 2-5 are in line with **Prediction i** in that the nonheads have greater duration, therefore the DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ is larger than that of $\text{V}_\text{h}\underline{\text{OBJ}}$, under both BROADFOCUS and NARROWFOCUS.

In Figure 6, DURATIONRATIO was grouped by TYPE to better examine the DURATIONRATIO change from BROADFOCUS to NARROWFOCUS. A two-way ANOVA (factors: TYPE and DISCOURSE) showed that DURATIONRATIO was conditioned by both TYPE ($F(1, 829) = 19.232$, $p < 0.0001$) and DISCOURSE ($F(1, 829) = 4.13$, $p < 0.05$). Tukey's HSD post-hoc tests showed that for $\text{V}_\text{h}\underline{\text{OBJ}}$, the DURATIONRATIO decrease (0.056) from BROADFOCUS to NARROWFOCUS was marginally significant ($p < 0.1$), which is consistent with **Prediction ii**. However, for $\underline{\text{MOD}}\text{N}_\text{h}$, the DURATIONRATIO decrease (0.015) from BROADFOCUS to NARROWFOCUS was not only non-significant ($p > 0.1$), but also departs from **Prediction ii**, which suggests a significant DURATIONRATIO increase. Consequently, as also observed in Figures 3-5, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS: the DURATIONRATIO difference between $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ was more pronounced under NARROWFOCUS.

Linear mixed model analysis (Table 1) confirmed that there was a global effect of TYPE that on average the DURATIONRATIO of $\text{V}_\text{h}\underline{\text{OBJ}}$ was 0.06 smaller than that of $\underline{\text{MOD}}\text{N}_\text{h}$($t(22.7) = -2.55$, $p < 0.05$). No significant effect of DISCOURSE was found. However, the interaction effect between TYPE and DISCOURSE bordered on the level of marginal significance ($t(799) = -1.024$, $p = 0.1115$). Given that $\underline{\text{MOD}}\text{N}_\text{h}$ and BROADFOCUS were assigned the value of 0, i.e., they were the dummy variables, and that $\text{V}_\text{h}\underline{\text{OBJ}}$ and NARROWFOCUS were assigned the value of 1 in the mixed-effects model, such an interaction effect suggested that the DURATIONRATIO decrease from BROADFOCUS to NARROWFOCUS for $\text{V}_\text{h}\underline{\text{OBJ}}$ was (marginally) significant, whereas for $\underline{\text{MOD}}\text{N}_\text{h}$ the DURATIONRATIO change was non-significant. This is consistent with Tukey's HSD post-hoc tests. Also note that when the second syllable ($\sigma_2$)

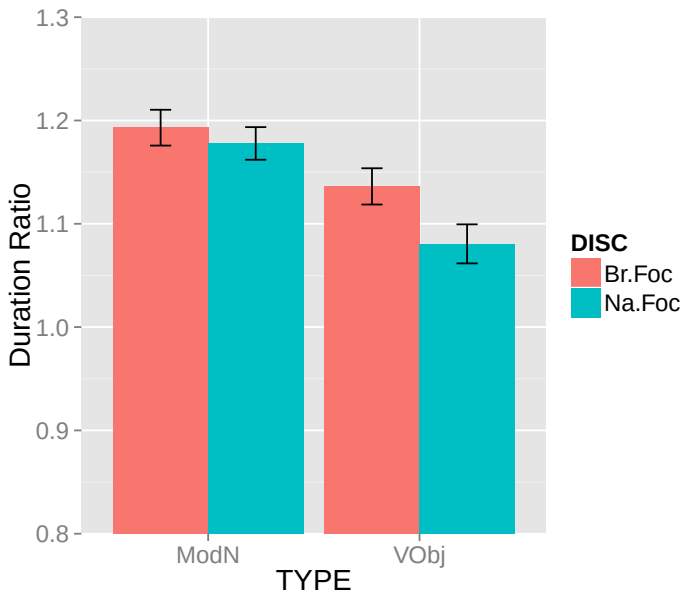



Figure 6: Mean DURATIONRATIO $\pm$ 1 standard error by TYPE and by DISCOURSE. The DURATIONRATIO of $V_h$OBJ significantly decreased from BROADFOCUS to NARROWFOCUS, whereas no significant DURATIONRATIO change was found for MOD$N_h$.

bore Tone3, the DURATIONRATIO significantly increased by 0.36 ($t(13) = 5.257$, $p < 0.0001$). This can be accounted for by the idiosyncrasy induced by Tone3-bearing syllables in that they have shorter durations.

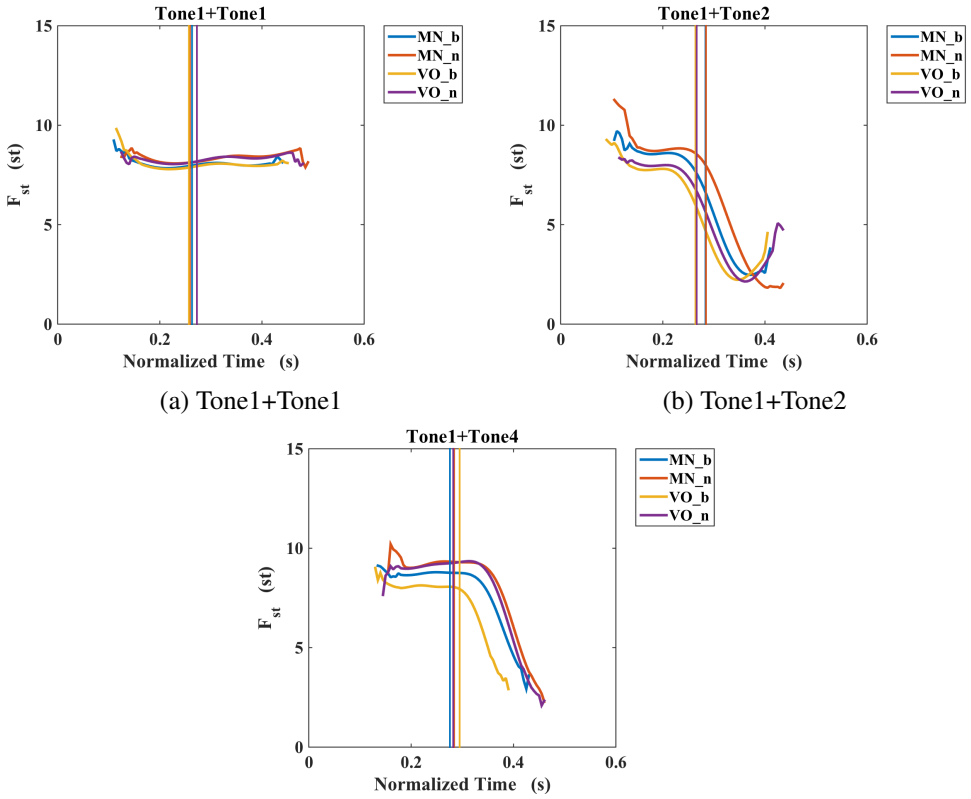
| **Fixed effects:** | | | | |
|---|---|---|---|---|
| | Estimate | df | Pr(>\|t\|) | |
| (Intercept) | 1.17 | 14.5 | 0.0000 | *** |
| **TypeV$_h$Obj** | -0.06 | 22.7 | 0.0180 | ** |
| DiscourseNarrowFocus | -0.02 | 798 | 0.3062 | |
| **TypeV$_h$Obj** : **DiscourseNarrowFocus** | -0.04 | 798 | 0.1115 | ' |
| Tone$_1$Tone2 | -0.08 | 13 | 0.1966 | |
| Tone$_1$Tone3 | -0.09 | 13 | 0.1616 | |
| **Tone$_1$Tone4** | -0.13 | 13 | 0.0409 | * |
| Tone$_2$Tone2 | 0.08 | 13 | 0.1850 | |
| **Tone$_2$Tone3** | 0.36 | 13 | 0.0001 | *** |
| Tone$_2$Tone4 | 0.09 | 13 | 0.1312 | |
| **Random effects:** | | | | |
| Groups | Variance | | St.Dev. | |
| Spk | 0.0014 | | 0.0378 | |
| Stim | 0.0024 | | 0.0495 | |
| Residual | 0.0275 | | 0.1660 | |
| Number of observations: 833, groups: Stim, 30; Spk, 3 | | | | |

Table 1: Results of the mixed model analysis on DurationRatio. Significant factors are shown in bold. Interaction between tones were not shown.

### 5.2 $F_{st}$ measurements

#### 5.2.1 Graphic comparisons of $F_{st}$ contours

This section presents graphic comparison of mean $F_{st}$ contours for all nine combinations (excluding those containing Tone3). For each tone combination, the $F_{st}$ contours were averaged across speakers and repetitions for four conditions (MN_b = Mod$N_h$ under BroadFocus, MN_n = Mod$N_h$ under NarrowFocus, VO_b = $V_h$Obj under BroadFocus, and VO_n = $V_h$Obj under NarrowFocus). For each condition, the duration of the target syllable was normalized to the median duration across speakers and repetitions.



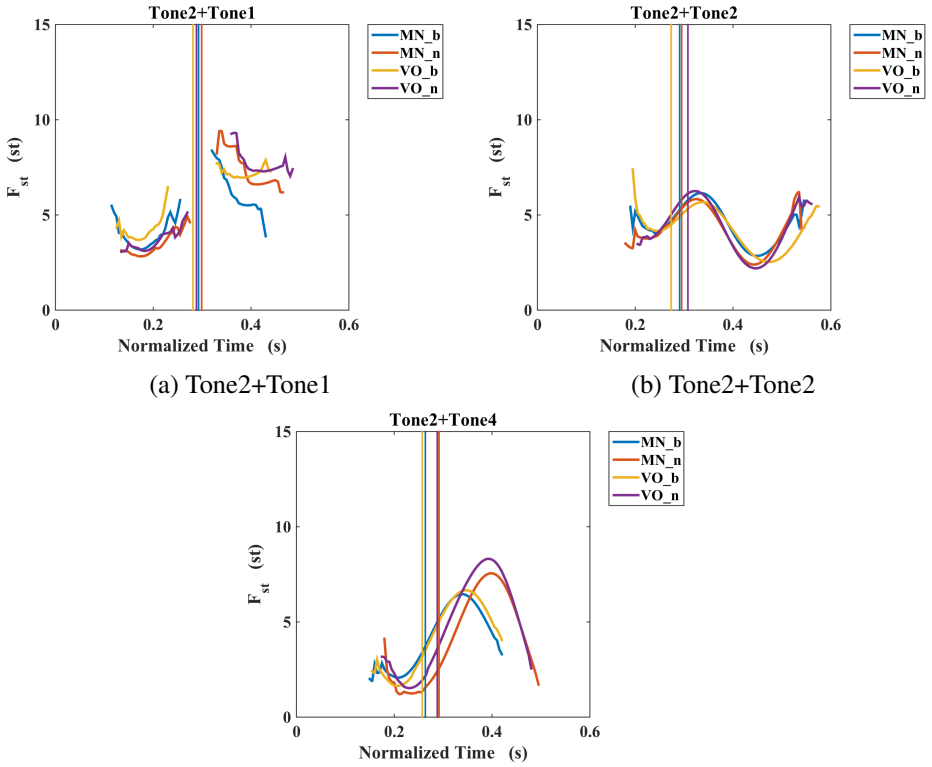

(a) Tone1+Tone1

(b) Tone1+Tone2

(c) Tone1+Tone4

Figure 7: Mean $F_{st}$ contours of Tone1-initial target stimuli. The $F_{st}$ contours were averaged across speakers and repetitions for four conditions: MN_b = MODN$_h$ under BROADFOCUS, blue; MN_n = MODN$_h$ under NARROWFOCUS, green; VO_b = V$_h$OBJ under BROADFOCUS, red; and VO_n = V$_h$OBJ under NARROWFOCUS, cyan. The duration of the target syllable was normalized to the median duration across speakers and repetitions. The vertical lines indicate the acoustic onset of $\sigma_2$ as well as the acoustic offset of $\sigma_1$.

Figure 7 shows mean $F_{st}$ contours of Tone1-initial target stimuli. For Tone1+Tone1, $F_{st}$ contours for four conditions were nearly identical. For Tone1+Tone2, the $F_{st}$ contours of Tone1 ($\sigma_1$) for both MN_b and MN_n were higher than those for VO_b and VO_n. Moreover, the $F_{st}$ contours of Tone2 ($\sigma_2$) show a pronounced rise for MN_b, VO_b and VO_n, whereas such a rise was not found for MN_n. This indicates the tone target of Tone2 (rising) was not fully realized on $\sigma_2$ (N$_h$) of MODN$_h$ under NARROWFOCUS. For Tone1+Tone4, MN_n and VO_n have higher overall $F_{st}$ contours than MN_b and VO_b, which can be accounted for by DISCOURSE.



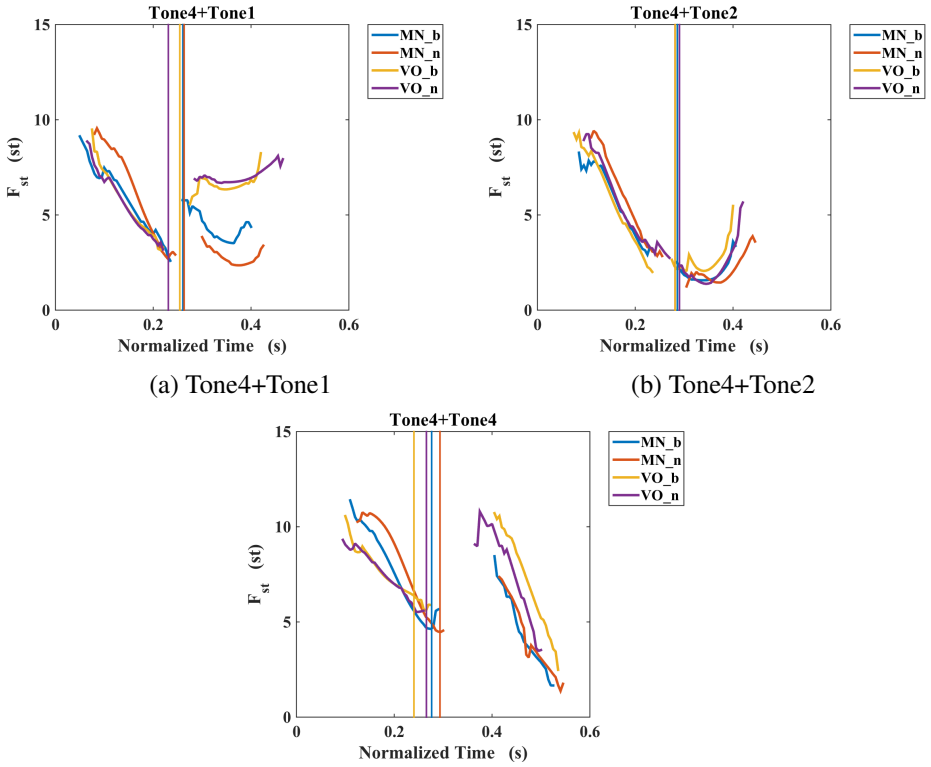

(a) Tone2+Tone1

(b) Tone2+Tone2

(c) Tone2+Tone4

Figure 8: Mean $F_{st}$ contours of Tone1-initial target stimuli. The $F_{st}$ contours were averaged across speakers and repetitions for four conditions: MN_b = MODN$_h$ under BROADFOCUS, blue; MN_n = MODN$_h$ under NARROWFOCUS, green; VO_b = V$_h$OBJ under BROADFOCUS, red; and VO_n = V$_h$OBJ under NARROWFOCUS, cyan. The duration of the target syllable was normalized to the median duration across speakers and repetitions. The vertical lines indicate the acoustic onset of $\sigma_2$ as well as the acoustic offset of $\sigma_1$.

Figure 8 shows mean $F_{st}$ contours of Tone2-initial target stimuli. For Tone2+Tone1, the $F_{st}$ differences on $\sigma_1$ (Tone2) were not noticeable among four conditions. The $F_{st}$ contours on $\sigma_2$ (Tone1) for VO_b and VO_n were respectively higher than those for MN_b and MN_n. Moreover, the overall $F_{st}$ contours under NARROWFOCUS were higher than those under BROADFOCUS. For Tone2+Tone2, $F_{st}$ contours for four conditions were nearly identical. For Tone2+Tone4, MN_n and VO_n have higher overall $F_{st}$ contours than MN_b and VO_b, which can be accounted for by DISCOURSE.




(a) Tone4+Tone1

(b) Tone4+Tone2

(c) Tone4+Tone4

Figure 9: Mean $F_{st}$ contours of Tone1-initial target stimuli. The $F_{st}$ contours were averaged across speakers and repetitions for four conditions: MN_b = MODN$_h$ under BROADFOCUS, blue; MN_n = MODN$_h$ under NARROWFOCUS, green; VO_b = V$_h$OBJ under BROADFOCUS, red; and VO_n = V$_h$OBJ under NARROWFOCUS, cyan. The duration of the target syllable was normalized to the median duration across speakers and repetitions. The vertical lines indicate the acoustic onset of $\sigma_2$ as well as the acoustic offset of $\sigma_1$.

Figure 9 shows mean $F_{st}$ contours of Tone4-initial target stimuli. For Tone4+Tone1, the $F_{st}$ contour on $\sigma_1$ (Tone4) for VO_n was substantially steeper than those for MN_b, MN_n, and VO_b. The $F_{st}$ contours for on $\sigma_2$ (Tone1) for VO_b and VO_n were higher than those for MN_b and MN_n. For Tone4+Tone2, $F_{st}$ contours for four conditions were nearly identical. For Tone4+Tone4, MN_b and MN_n have steeper $F_{st}$ contours on $\sigma_1$ (Tone4) but less steep $F_{st}$ contours on $\sigma_1$ (Tone4) than VO_b and VO_n.

In sum, the $F_{st}$ contours of the nine tone combinations fall into three types: (a) those that had identical $F_{st}$ contours across four conditions (Tone1+Tone1, Tone2+Tone2, and Tone4+Tone2); (b) those that had higher overall $F_{st}$ contours under NARROWFOCUS (Tone1+Tone4 and Tone2+Tone4); (c) those that exhibited differences in $F_{st}$ among four conditions induced by the interaction between DISCOURSE and TYPE (Tone1+Tone2, Tone2+Tone1, Tone4+Tone1, and Tone4+Tone4). In the next section, the tone combinations that belong to Type (c) will be examined in detail.

### 5.2.2 Quantitative analysis of $F_{st}$ contours

This section examines three tone combinations, i.e., Tone1+Tone2, Tone4+Tone1, and Tone4+Tone4, for which the graphic comparisons of $F_{st}$ contours in four conditions (MN_b,

MN_n, VO_b, and VO_n) exhibited notable differences. For each tone combination, $F_{st}$ measurements of both $\sigma_1$ and $\sigma_2$ were respectively fitted into a mixed-effects regression model with two fixed factors DISCOURSE and TYPE, and a random effect factor SPK. The interaction effect between DISCOURSE and TYPE was also included. The mixed-effects model is shown below, where the specific measurements of $F_{st}$ depend on tonal categories:

$$F_{st} \sim TYPE + DISC + TYPE * DISC + (1 + SPK)$$

**Tone1+Tone2**

**Tone1 ($\sigma_1$)**

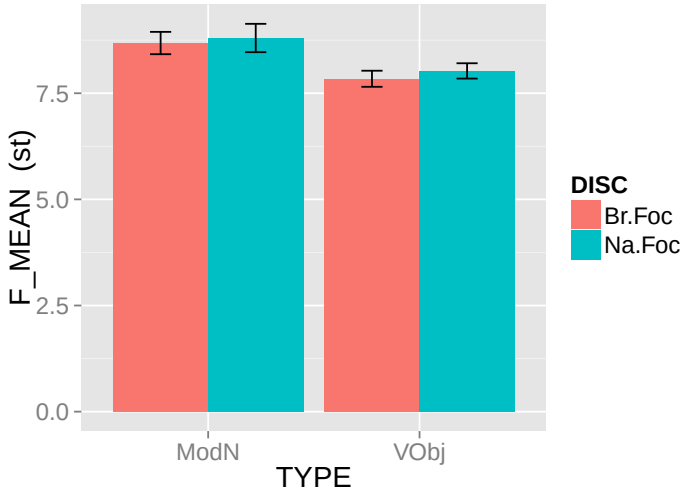



Figure 10: Mean $F_{st\_mean}$ ($\pm$1 standard error) on $\sigma_1$ (Tone1) of Tone1+Tone2 combination by TYPE and by DISCOURSE.

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 8.71 | 0.31 | 5 | 27.710 | 0.0000 | *** |
| **TYPEV$_h$OBJ** | -0.86 | 0.33 | 46 | -2.599 | 0.0125 | * |
| DISCOURSENARROWFOCUS | 0.15 | 0.33 | 46 | 0.438 | 0.6634 | |
| TYPEV$_h$OBJ: DISCOURSENARROWFOCUS | 0.10 | 0.48 | 46 | 0.199 | 0.8428 | |
| **Random effects:** | | | | | | |
| Groups | Name | Variance | | Std. Dev. | | |
| SPK | (Intercept) | 0.1365 | | 0.3694 | | |
| Residual | | 0.7432 | | 0.8621 | | |
| Number of observations: 52, groups: SPK , 3 | | | | | | |

Table 2: Results of the mixed-effects linear regression of $F_{st\_mean}$ on $\sigma_1$ (Tone1) of Tone1+Tone2 combination. Significant factors are shown in bold.

Figure 10 shows the mean values of $F_{st\_mean}$ across speakers and repetitions on $\sigma_1$ (Tone1) of Tone1+Tone2 for four conditions. $F_{st\_mean}$ on $\sigma_1$ (Tone1) for $\underline{\text{MOD}}\text{N}_\text{h}$ was significantly larger than that for $\text{V}_\text{h}\underline{\text{OBJ}}$ [$t_{\text{V}_\text{h}\underline{\text{OBJ}}}(46) = -2.599$, $p < 0.05$], which indicates that $F_{st}$ contours on $\sigma_1$ (Tone1) for $\underline{\text{MOD}}\text{N}_\text{h}$ were higher than that for $\text{V}_\text{h}\underline{\text{OBJ}}$, as shown in Figure 7b. However, no significant effect of either DISCOURSE or interaction between DISCOURSE and TYPE was found on $F_{st\_mean}$ on $\sigma_1$ (Tone1).

**Tone2 ($\sigma_2$)**

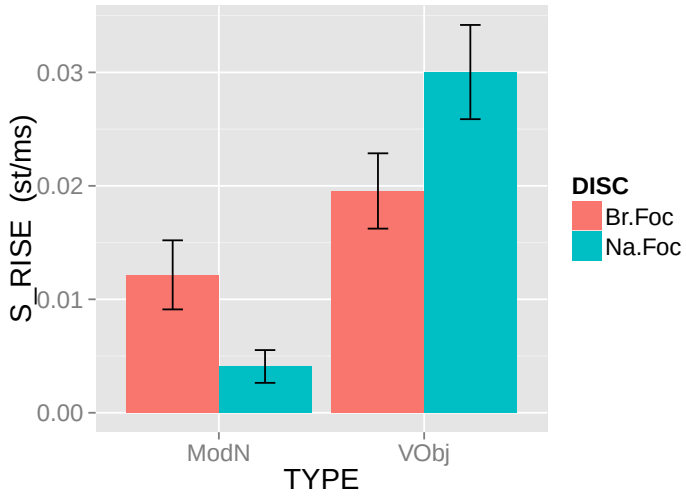



Figure 11: Mean $S_{rise}$ ($\pm 1$ standard error) on $\sigma_2$ (Tone2) of Tone1+Tone2 combination by TYPE and by DISCOURSE.

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 0.01 | 0.00 | 7 | 3.416 | 0.0105 | * |
| **TYPE$\text{V}_\text{h}\underline{\text{OBJ}}$** | 0.01 | 0.00 | 46 | 1.843 | 0.0718 | . |
| **DISCOURSENARROWFOCUS** | -0.01 | 0.00 | 46 | -1.944 | 0.0579 | . |
| **TYPE$\text{V}_\text{h}\underline{\text{OBJ}}$ : DISCOURSENARROWFOCUS** | 0.02 | 0.01 | 46 | 3.173 | 0.0027 | ** |
| **Random effects:** | | | | | | |
| Groups | Name | Variance | Std. Dev. | | | |
| SPK | (Intercept) | 1.363e-05 | 0.0037 | | | |
| Residual | | 1.147e-04 | 0.0107 | | | |
| Number of observations: 52, groups: SPK , 3 | | | | | | |

Table 3: Results of the mixed-effects linear regression of $S_{rise}$ on $\sigma_2$ (Tone2) of Tone1+Tone2 combination. Significant factors are shown in bold.

Figure 11 shows the mean values of $S_{rise}$ across speakers and repetitions on $\sigma_2$ (Tone2) of Tone1+Tone2 for four conditions. The main effect of TYPE was marginally significant: $S_{rise}$ on $\sigma_2$ (Tone2) for $\text{V}_\text{h}\underline{\text{OBJ}}$ was larger than that for $\underline{\text{MOD}}\text{N}_\text{h}$ [$t_{\text{V}_\text{h}\underline{\text{OBJ}}}(46) = 1.843$, $p < 0.1$],

which indicates that the $F_{st}$ rise on $\sigma_2$ (Tone2) for $V_h$OBJ was steeper than that for MODN$_h$, as shown in Figure 7b. The interaction effect between DISCOURSE and TYPE induced a significant decrease of $S_{rise}$ on $\sigma_2$ (Tone2) from BROADFOCUS to NARROWFOCUS for MODN$_h$, and a significant increase of $S_{rise}$ on $\sigma_2$ (Tone2) from BROADFOCUS to NARROWFOCUS for $V_h$OBJ [$t_{TYPE:DISCOURSE}(46) = 3.173$, $p < 0.01$]. The main effect of DISCOURSE induced a marginally significant decrease in $S_{rise}$ on $\sigma_2$ (Tone2) for NARROWFOCUS [$t_{NARROWFOCUS}$ (46) = -1.94, $p < 0.1$]. This can be attributed to the nearly flat $F_{st}$ contours on $\sigma_2$ (Tone2) of MODN$_h$ under NARROWFOCUS.

In sum, TYPE significantly affected $F_{st}$ measurements on both $\sigma_1$ (Tone1) and on $\sigma_2$ (Tone2). However, the effects were in different directions: $F_{st\_mean}$ on $\sigma_1$ (Tone1) for MODN$_h$ was larger than that for $V_h$OBJ, whereas $S_{rise}$ on $\sigma_2$ (Tone2) for MODN$_h$ was smaller than that for $V_h$OBJ. The interaction between TYPE and DISCOURSE was found significant on $S_{rise}$ on $\sigma_2$ (Tone2): from BROADFOCUS to NARROWFOCUS, DISCOURSE induced a significant decrease of $S_{rise}$ for MODN$_h$, but a significant increase of $S_{rise}$ on $\sigma_2$ (Tone2) for $V_h$OBJ.

**Tone4+Tone1**

**Tone4 ($\sigma_1$)**

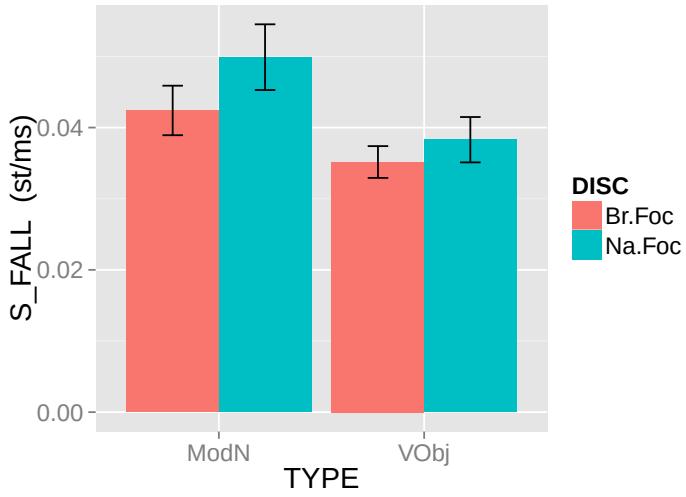



Figure 12: Mean $S_{fall}$ ($\pm$1 standard error) on $\sigma_1$ (Tone4) of Tone4+Tone1 combination by TYPE and by DISCOURSE.

Figure 12 shows the mean values of $S_{fall}$ across speakers and repetitions on $\sigma_1$ (Tone4) of Tone4+Tone1 for four conditions. The main effect of TYPE was significant: $S_{fall}$ on $\sigma_1$ (Tone4) for MODN$_h$ was larger than that for $V_h$OBJ [$t_{VO}(46) = -2.611$, $p < 0.05$], which indicates that the $F_{st}$ fall on $\sigma_1$ (Tone4) for MODN$_h$ was steeper than that for $V_h$OBJ, as shown in Figure 9a. The main effect of DISCOURSE was significant [$t_{NF}(46) = -1.94$, $p < 0.1$]: it induced a significant increase in $S_{fall}$ on $\sigma_1$ (Tone4) from BROADFOCUS to NARROWFOCUS for MODN$_h$, whereas the increase was marginal for $V_h$OBJ. No significant effect of interaction between TYPE and DISCOURSE was found on $S_{fall}$ on $\sigma_1$ (Tone4) [$t_{TYPE:DISC}(46) = -1.238$, $p > 0.1$].

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 0.05 | 0.01 | 2 | 5.605 | 0.0234 | * |
| **TypeV$_h$Obj** | -0.01 | 0.00 | 46 | -2.611 | 0.0122 | * |
| **DiscourseNarrowFocus** | 0.01 | 0.00 | 46 | 2.237 | 0.0302 | * |
| TypeV$_h$Obj: DiscourseNarrowFocus | -0.00 | 0.00 | 46 | -1.238 | 0.2222 | |
| **Random effects:** | | | | | | |
| Groups | Name | | Variance | | Std. Dev. | |
| Spk | (Intercept) | | 1.815e-04 | | 0.0135 | |
| Residual | | | 5.176e-05 | | 0.0072 | |
| Number of observations: 52, groups: Spk , 3 | | | | | | |

Table 4: Results of the mixed-effects linear regression of $S_{fall}$ on $\sigma_1$ (Tone4) of Tone4+Tone1 combination. Significant factors are shown in bold.

**Tone1 ($\sigma_2$)**

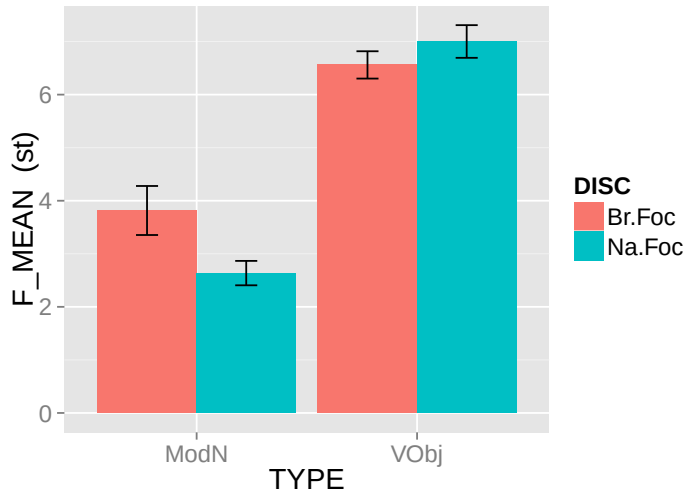



Figure 13: Mean $F_{st\_mean}$ ($\pm$1 standard error) on $\sigma_2$ (Tone1) of Tone4+Tone1 combination by Type and by Discourse.

Figure 13 shows the mean values of $F_{st\_mean}$ across speakers and repetitions on $\sigma_2$ (Tone1) of Tone4+Tone1 for four conditions. The main effect of Type was significant: $F_{st\_mean}$ on $\sigma_2$ (Tone1) for V$_h$Obj was larger than that for ModN$_h$ [$t_{V_hOBJ}$ (48) = 6.038, $p < 0.001$], which indicates that the $F_{st}$ contours on $\sigma_2$ (Tone1) for V$_h$Obj were higher than that for ModN$_h$, as shown in Figure 9a. The interaction effect between Discourse and Type induced a significant decrease of $F_{st\_mean}$ on $\sigma_2$ (Tone1) from BroadFocus to NarrowFocus for ModN$_h$, and a marginal increase of $F_{st\_mean}$ on $\sigma_2$ (Tone1) from BroadFocus to NarrowFocus for V$_h$Obj [$t_{TYPE:DISCOURSE}$(48) = 2.579, $p < 0.05$]. The main effect of Discourse induced a significant decrease in $F_{st\_mean}$ on $\sigma_2$ (Tone1) for NarrowFocus [$t_{NARROWFOCUS}$ (48) = -2.554, $p < 0.05$], primarily contributed by the significant decrease in $F_{st\_mean}$ on $\sigma_2$ (Tone1) of ModN$_h$ under NarrowFocus.

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 3.82 | 0.34 | 48 | 11.222 | 0.0000 | *** |
| **TypeV$_h$Obj** | 2.74 | 0.45 | 48 | 6.038 | 0.0000 | *** |
| **DiscourseNarrowFocus** | -1.18 | 0.46 | 48 | -2.554 | 0.0139 | * |
| **TypeV$_h$Obj:** **DiscourseNarrowFocus** | 1.62 | 0.63 | 48 | 2.579 | 0.0130 | * |
| **Random effects:** | | | | | | |
| Groups | Name | Variance | Std. Dev. | | | |
| Spk | (Intercept) | 0.000 | 0.000 | | | |
| Residual | | 1.272 | 1.128 | | | |
| Number of observations: 52, groups: Spk , 3 | | | | | | |

Table 5: Results of the mixed-effects linear regression of $F_{st\_mean}$ on $\sigma_2$ (Tone1) of Tone4+Tone1 combination. Significant factors are shown in bold.

To summarize, Type significantly affected $F_{st}$ measurements on both $\sigma_1$ (Tone1) and on $\sigma_2$ (Tone1). However, the effects were in different directions: $S_{fall}$ on $\sigma_1$ (Tone4) for ModN$_h$ was larger than that for V$_h$Obj, whereas $F_{st\_mean}$ on $\sigma_2$ (Tone1) for ModN$_h$ was smaller than that for V$_h$Obj. The effect of interaction between Type and Discourse was found to be significant on $F_{st\_mean}$ on $\sigma_2$ (Tone1): from BroadFocus to NarrowFocus, Discourse induced a significant decrease of $F_{st\_mean}$ on $\sigma_2$ (Tone1) for ModN$_h$, but a marginally significant increase of $F_{st\_mean}$ on $\sigma_2$ (Tone1) for V$_h$Obj.

**Tone4+Tone4**

**Tone4 ($\sigma_1$)**

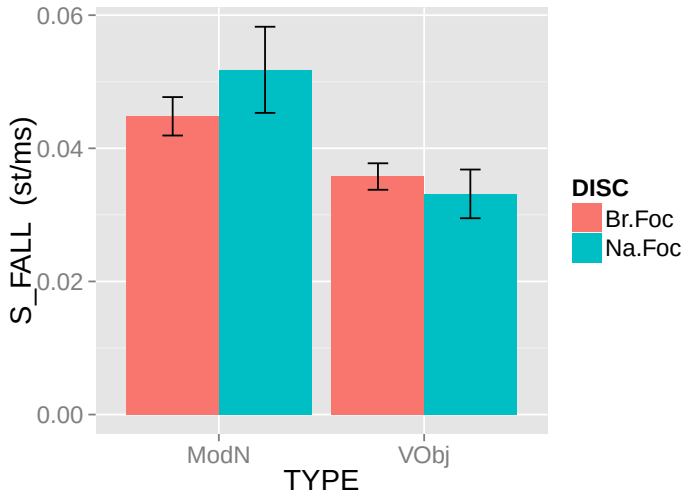



Figure 14: Mean $S_{fall}$ ($\pm$1 standard error) on $\sigma_1$ (Tone4) of Tone4+Tone4 combination by Type and by Discourse.

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 0.05 | 0.01 | 1 | 5.679 | 0.0807 | . |
| **TypeV$_h$Obj** | -0.01 | 0.00 | 31 | -2.158 | 0.0388 | * |
| **DiscourseNarrowFocus** | 0.01 | 0.00 | 31 | 1.947 | 0.0607 | . |
| **TypeV$_h$Obj:** **DiscourseNarrowFocus** | -0.01 | 0.01 | 31 | -1.990 | 0.0555 | . |
| **Random effects:** | | | | | | |
| Groups | Name | | Variance | | Std. Dev. | |
| Spk | (Intercept) | | 1.132e-04 | | 0.0106 | |
| Residual | | | 7.156e-05 | | 0.0084 | |
| Number of observations: 36, groups: Spk , 2 | | | | | | |

Table 6: Results of the mixed-effects linear regression of $S_{fall}$ on $\sigma_1$ (Tone4) of Tone4+Tone4 combination. Significant factors are shown in bold.

Figure 14 shows the mean values of $S_{fall}$ across speakers and repetitions on $\sigma_1$ (Tone4) of Tone4+Tone4 for four conditions. The main effect of Type was significant: $S_{fall}$ on $\sigma_1$ (Tone4) for ModN$_h$ was larger than that for V$_h$Obj [$t_{V_hOBJ}(31) = -2.158$, $p < 0.05$], which indicates that the $F_{st}$ fall on $\sigma_1$ (Tone4) for ModN$_h$ was steeper than that for V$_h$Obj, as shown in Figure 9c. The interaction effect between Discourse and Type was marginally significant [$t_{TYPE:DISCOURSE}(31) = -1.990$, $p < 0.1$]: it induced a significant increase of $S_{fall}$ on $\sigma_1$ (Tone4) from BroadFocus to NarrowFocus for ModN$_h$, and a marginal decrease of $S_{fall}$ on $\sigma_1$ (Tone4) from BroadFocus to NarrowFocus for V$_h$Obj. The main effect of Discourse induced a marginally significant increase in $S_{fall}$ on $\sigma_1$ (Tone4) for NarrowFocus [$t_{NARROWFOCUS}(31) = 1.947$, $p < 0.1$], primarily contributed by the significant increase of $S_{fall}$ on $\sigma_1$ (Tone4) of ModN$_h$ under NarrowFocus.

**Tone4 ($\sigma_2$)**

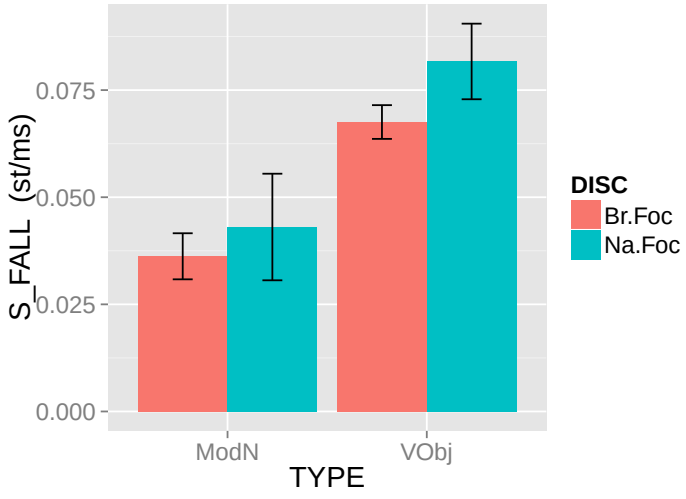



Figure 15: Mean $S_{fall}$ ($\pm$1 standard error) on $\sigma_2$ (Tone4) of Tone4+Tone4 by Type and by Discourse.

| **Fixed effects:** | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
| (Intercept) | 0.04 | 0.02 | 1 | 2.315 | 0.2229 | |
| **TypeV$_h$Obj** | 0.03 | 0.01 | 31 | 4.111 | 0.0003 | *** |
| DiscourseNarrowFocus | 0.01 | 0.01 | 31 | 1.061 | 0.2969 | |
| TypeV$_h$Obj: DiscourseNarrowFocus | 0.00 | 0.01 | 31 | 0.335 | 0.7402 | |
| **Random effects:** | | | | | | |
| Groups | Name | | Variance | | Std. Dev. | |
| Spk | (Intercept) | | 4.690e-04 | | 0.0217 | |
| Residual | | | 2.989e-04 | | 0.0173 | |
| Number of observations: 36, groups: Spk , 2 | | | | | | |

Table 7: Results of the mixed-effects linear regression of $S_{fall}$ on $\sigma_2$ (Tone4) of Tone4+Tone4 combination. Significant factors are shown in bold.

Figure 15 shows the mean values of $S_{fall}$ across speakers and repetitions on $\sigma_2$ (Tone4) of Tone4+Tone4 for four conditions. $S_{fall}$ on $\sigma_2$ (Tone4) for V$_h$Obj was significantly larger than that for Mod N$_h$ [$t_{VO}(31) = 4.111$, $p < 0.001$], which indicates that the $F_{st}$ fall on $\sigma_2$ (Tone4) for V$_h$Obj was steeper than that for V$_h$Obj, as shown in Figure 9c. However, no significant effect of either Discourse or interaction between Discourse and Type was found on $S_{fall}$ on $\sigma_2$ (Tone4).

In sum, Type significantly affected $F_{st}$ measurements on both $\sigma_1$ (Tone4) and $\sigma_2$ (Tone4). However, the effects were in different directions: $S_{fall}$ on $\sigma_1$ (Tone4) for Mod N$_h$ was larger than that for V$_h$Obj, whereas $S_{fall}$ on $\sigma_2$ (Tone4) for Mod N$_h$ was smaller than that for V$_h$Obj. Unlike Tone1+Tone2 and Tone4+Tone4, the effect interaction between Type and Discourse was found significant on $S_{fall}$ on $\sigma_1$ (Tone4) rather than on $\sigma_2$ (Tone4): from BroadFocus to NarrowFocus, Discourse induced a significant increase of $S_{fall}$ on $\sigma_1$ (Tone4) for Mod N$_h$, but a marginally significant decrease of $S_{fall}$ on $\sigma_1$ (Tone4) for V$_h$Obj.

| Tone Combination | Stressed | Unstressed | Unstressed | Stressed |
|---|---|---|---|---|
| | Mod | N$_h$ | V$_h$ | Obj |
| Tone1+Tone2 | – | ⇓ | – | ⇑ |
| Tone4+Tone1 | ⇑ | ⇓ | – | ⇑ (.) |
| Tone4+Tone4 | ⇑ | – | ⇓ (.) | ⇑ |

Table 8: Summary of $F_{st}$ measurements by syllable position in three tone combinations (Tone1+Tone2, Tone4+Tone1, and Tone4+Tone4). The arrows indicate the direction of change in $F_{st}$ measurement from BroadFocus to NarrowFocus. The dots after the arrows indicate the differences in $F_{st}$ measurement between BroadFocus and NarrowFocus were marginally significant. Otherwise, the differences were statistically significant.

For the three tone combinations that exhibited differences in $F_{st}$ contours between V$_h$Obj and Mod N$_h$, the results confirm **Prediction i** in that the Mod and the Obj respectively had larger $F_{st}$ measurements than the V$_h$ and the N$_h$; the results are partly in line with **Prediction ii** in that under NarrowFocus, the $F_{st}$ measurements on the nonheads (the Mod of Mod N$_h$ and the Obj of V$_h$Obj) increased significantly from BroadFocus to

NARROWFOCUSin five out of six instances (except for $\sigma_1$ (Tone1) of Tone1+Tone2), whereas the $F_{st}$ measurements on the heads (the $N_h$ of MODN$_h$ and the $V_h$ of $V_h$OBJ) decreased significantly from BROADFOCUS to NARROWFOCUSin three out of six instances (Table 8).

### 5.3 Summary of results

The main findings can be summarized as follows:

① The DURATIONRATIO of MODN$_h$ was larger than that of $V_h$OBJ.

② The DURATIONRATIO change (decrease) from BROADFOCUS to NARROWFOCUS was significant for $V_h$OBJ, whereas such change was not significant for MODN$_h$.

③ The DURATIONRATIO difference between MODN$_h$ and $V_h$OBJ was more pronounced under NARROWFOCUS than under BROADFOCUS.

④ There existed cross-speaker and cross-stimulus variation.

⑤ For the three tone combinations (Tone1+Tone2, Tone4+Tone1, and Tone4+Tone4), the MOD and the OBJ respectively had larger $F_{st}$ measurements than the $V_h$ and the $N_h$.

⑥ For the three tone combinations, the $F_{st}$ change from BROADFOCUS to NARROWFOCUS was more pronounced on the nonheads than on the heads.

⑦ The majority of the homophonous pairs did not exhibit differences in $F_{st}$ contours between $V_h$OBJ and MODN$_h$.

## 6 Discussion & Conclusion

The DURATIONRATIO difference between MODN$_h$ and $V_h$OBJ (Finding ①) suggested there was a global TYPE effect. Such a difference may arise from one of the following three scenarios: (A) MODN$_h$ stresses the MOD and $V_h$OBJ stresses the OBJ; (B) MODN$_h$ stresses the MOD and $V_h$OBJ has equal stress for both $V_h$ and OBJ; (C) MODN$_h$ has equal stress for both MOD and $N_h$ and $V_h$OBJ stresses the OBJ.

Focus comes in as a handy diagnostic tool: Finding ② suggested Scenario (C) was the likely answer. That is, the different behaviors of DURATIONRATIO change in MODN$_h$ and $V_h$OBJ should be mainly attributed to the final stress of $V_h$OBJ. This agrees with the observations in Shen et al. (2013) that $V_h$OBJ exhibited final stress whereas MODN$_h$ exhibited no initial stress. However, such a claim would essentially undermine the validity of Duanmu's (2007) NONHEAD STRESS RULE.

However, rejecting NONHEAD STRESS RULE as a whole in turn weakens the argument that $V_h$OBJ has final stress, leaving it with no concrete theoretical foundation. Moreover, recall that NONHEAD STRESS RULE is motivated by the assumption that the information load a constituent carries determines its stress status. This assumption is in line with Finding ③, which shows that under NARROWFOCUS, the communicative efficiency is facilitated by means of loading more information into the stressed form, i.e., the OBJ of $V_h$OBJ.

Moreover, the $F_{st}$ differences between the MOD and the $V_h$, and those between the OBJ and the $N_h$ further suggested that the MOD was a stronger position than the $V_h$, and that the OBJ was a stronger position than the $N_h$ (Finding ⑤). Similarly, Finding ⑥ was also

in line with the underlying assumption that motivates the NONHEAD STRESS RULE. Therefore, the discrepancy between Finding ② (that $V_h$OBJ has final stress and MOD$N_h$ has no initial stress) and the information-motivated assumption of NONHEAD STRESS RULE must be reconciled.

Specifically, the DURATIONRATIO change from BROADFOCUS to NARROWFOCUS for MOD$N_h$ needs to be accounted for. One possible reason is that Mandarin disyllabic phrases have trochaic foot structures in that they show a strong–weak alternating pattern Duanmu (2007). Because the first syllable ($\sigma_1$) is already a strong position, NARROWFOCUS does not induce any pronounced change in DURATIONRATIO (ceiling effect). In this case, the focus-introduced metrical prominence is still associated with the MOD of MOD$N_h$, but is disguised by the underlying strong–weak pattern. Note that the underlying trochaic foot structures do not refute NONHEAD STRESS RULE. It can be understood as that the underlying strong-weak pattern sets the baseline for all disyllabic phrases, and that the real comparison should be made between the syllables occupying the same positions, i.e., between the MOD of MOD$N_h$ and the $V_h$ of $V_h$OBJ, and between the $N_h$ of MOD$N_h$ and the OBJ of $V_h$OBJ. Taking that as a departure, the DURATIONRATIO and the $F_{st}$ results show that the MOD is a stronger position than the $V_h$, and the OBJ is a stronger position than the $N_h$. A second possible interpretation is that there exist stimulus-dependent stress patterns that contribute to the overall non-significant DURATIONRATIO change for MOD$N_h$. It is possible that the majority of MOD$N_h$ stimuli in the current study did not exhibit initial stress, therefore disguising the DISCOURSE effect. This interpretation is strongly supported by the $F_{st}$ results, which showed the majority of the homophonous pairs did not exhibit different $F_{st}$ contours between MOD$N_h$ and $V_h$OBJ (Finding ⑦).

While there existed variation, no speakers or stimuli showed patterns that went in the opposite direction of Findings ①–③ or Findings ⑤–⑥. For F01 and F02, the DURATIONRATIO of MOD$N_h$ was larger than that of $V_h$OBJ; for M01, the DURATIONRATIO of MOD$N_h$ and $V_h$OBJ were not differentiable under either BROADFOCUS or NARROWFOCUS. The DURATIONRATIO of MOD$N_h$ was either larger than or was not significantly different from that of $V_h$OBJ; the $F_{st}$ measurements of the MOD and the OBJ were either larger than or were not significantly different from those of the $V_h$ and the $N_h$, respectively.

For these reasons, I will tentatively argue that in line with Scenario (A) (NONHEAD STRESS RULE), the differences in DURATIONRATIO and in $F_{st}$ measurements between MOD$N_h$ and $V_h$OBJ reflect the difference between initial stress and final stress, which is further indicative of two different syntactic structures, and that the information-motivated NONHEAD STRESS RULE is an important component to the prosodic process in Mandarin as it facilitates communicative efficiency by loading more stress into forms with more information. However, I also argue that such variation are more of idiosyncrasies than randomness. It is also acknowledged that NONHEAD STRESS RULE is a weak universal in that whether the phrasal stress patterns will surface to differentiate between a homophonous pair of MOD$N_h$ and $V_h$OBJ depends heavily on the idiosyncrasies of particular lexical items or individual speakers. This is because Mandarin, above all, is a tone language that uses pitch variation to contrast lexical meanings. The outstanding differences in surface prominence will inevitably result in the repressing (even loss) of the tonal realization, which is not ideal for words with a relatively low lexical frequency. On the other hand, differences in prominence are more likely to surface in highly frequent words (especially those with frequent affixes) without resulting in communicative inefficiency. As a matter of fact, such differences in prominence that comply with NONHEAD STRESS RULE facilitate the recognition as they render a homophonous pair of MOD$N_h$ and $V_h$OBJ maximally differentiable from each other.

The study strongly suggests that the tendency of contrasting $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ results from NONHEAD STRESS RULE. It is argued that NONHEAD STRESS RULE, despite being weak, is an important component of the prosodic process in Mandarin Chinese, because it facilitates communicative efficiency by loading more stress into forms with more information. Perception studies are further needed in order to show whether such knowledge of contrast does exist for those homophonous pairs that do not exhibit overt contrastive phrasal stress patterns in acoustics.

## References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67:1–48.

Boersma, Paul, and David Weenink. 2015. Praat: Doing phonetics by computer [Computer program].

Brookes, Mike. 2005. VOICEBOX: Speech Processing Toolbox for MATLAB.

Chen, Yiya, and Carlos Gussenhoven. 2008. Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics* 36:724–746.

Duanmu, San. 2007. *The phonology of Standard Chinese*. Oxford University Press, 2 edition.

Jia, Yuan. 2011. Putonghua tonyinyigou liangyinzu zhongyin leixing bianxi [Stress patterns of disyllabic terms with identical pronunciation and different morph-syntactic structures in Standard Chinese]. *Qinghua Daxue Xuebao (Ziran Kexue ban) [Journal of Tsinghua University (Science & Technology)]* 51:1307–1312.

Lai, Catherin, Yanyan Sui, and Jiahong Yuan. 2010. A corpus study of the prosody of polysyllabic words in Mandarin Chinese. In *Proceedings of Speech Prosody 2010*.

Shen, Weilin, Jacqueline Vaissière, and Frédéric Isel. 2013. Acoustic correlates of contrastive stress in compound words versus verbal phrase in Mandarin Chinese. *Computational Linguistics and Chinese Language Processing* 18:45–58.

Hao Yi
Department of Linguistics
Cornell University
Ithaca, NY 14853
*hy433@cornell.edu*