Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Phrasal stress in Mandarin disyllabic phrases: an investigation using focus

Hao Yi

Department of Linguistics, Cornell University
hy433@cornell.edu

Abstract

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 [1]. Taking NONHEAD STRESS RULE as a point of departure, the current study investigated Mandarin phrasal stress by using focus as a diagnostic tool. 15 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 correlate of phrasal stress—duration—was measured. The hypothesis tested was that the nonheads have phrasal stress. The results showed that at the phrase level, a MODN$_h$ and a homophonous V$_h$OBJ differed significantly in duration ratio, consistent with the interpretation that MODN$_h$ exhibits initial stress and V$_h$OBJ exhibits final stress. Moreover, the duration ratio difference was amplified under NARROWFOCUS. 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.

Index Terms: phrasal stress, focus, duration, Mandarin Chinese

1. Introduction

Mandarin Chinese has been claimed to have phrasal stress. The distribution of stress, according to [1], is governed by NONHEAD STRESS RULE: phrasal stress falls on the nonhead constituent of a phrase, because the nonhead carries more information than its syntactic head. 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$_h$.

This distribution of stress has been addressed by several acoustic studies, all taking [1] as their point of departure. One of the studies, based on a spoken corpus, investigated the rhythmic patterns in Mandarin polysyllabic words [2]. It concluded that there was no difference between disyllabic V$_h$OBJ and MODN$_h$ in terms of stress pattern on the basis of acoustic measurements, and therefore could not confirm NONHEAD STRESS RULE.

Another study used identical MODN$_h$ and V$_h$OBJ disyllabic pairs in a production experiment [3]. During the elicitation, a disyllabic phrase was overtly preceded by its part of speech in the carrier sentence so that a MODN$_h$ and an identical V$_h$OBJ can be differentiated. The results showed that while V$_h$OBJ exhibited final stress, MODN$_h$ showed no initial stress, which, again, did not confirm NONHEAD STRESS RULE.

A third study used homophones that differed in terms of syntactic structures, i.e., each pair of homophonous disyllabic phrases consisted of one V$_h$OBJ and one MODN$_h$ (which differed in orthography) [4]. Target phrases were elicited in isolation. It was concluded that most of the disyllabic phrases in the study exhibited final stress, and that therefore syntactic structure did not govern stress allocation in Mandarin.

While these studies lent great insight into the distribution of Mandarin phrasal stress, none of them confirmed NONHEAD STRESS RULE. Moreover, they raise methodological concerns, such as potential complications due to the unfounded reliance on Mandarin speakers' judgement of parts of speech [3] or due to phrase final lengthening [4].

In the current study, such methodological drawbacks are carefully controlled for. Focus is used as a diagnostic tool to look for prosodic regularities in Mandarin disyllabic phrases. Specifically, this study investigates the phonetic correlates of phrasal stress in Mandarin Chinese, by measuring the duration under both BROADFOCUS and NARROWFOCUS. The effects of focus and syntactic structure on duration are tested in 15 homophonous pairs of one MODN$_h$ and one V$_h$OBJ. If a homophonous pair (MODN$_h$ and V$_h$OBJ) displays contrastive stress patterns, focus-introduced prominence will apply differently: the duration changes induced by NARROWFOCUS for the stressed constituents (the MOD of MODN$_h$ and the OBJ of V$_h$OBJ) will be of greater magnitude than for their unstressed counterparts (the N$_h$ of MODN$_h$ and the V$_h$ of V$_h$OBJ).

2. Methods

2.1. 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. 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.

2.2. Speech materials and data collection

The stimulus set consisted of 15 homophonous pairs of MODN$_h$ and V$_h$OBJ. Homophones were chosen because segmental variations 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. The target stimuli were elicited in two discourse contexts: BROADFOCUS and NARROWFOCUS.

(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.

 doi: 10.21437/SpeechProsody.2016-152

(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.

For a given background sentence (**BACKGROUND**), there were two types of questions: the BROADFOCUS question and the NARROWFOCUS question, which are listed in (**BROADFOCUS**) and (**NARROWFOCUS**), respectively. The speech materials were exemplified in Pinyin, the official phonetic system for transcribing Mandarin Chinese in the Latin alphabet. Tones are omitted. The disyllabic target stimulus is represented as σ_1 σ_2.

(**BACKGROUND**)	ta juede shuo σ_1 σ_2 shun henduo. *'He thinks it's a lot more fluent to say σ_1 σ_2.'*
(**BROAD FOCUS**)	ta juede shenme? *'What does he think?'*
(**NARROW FOCUS**)	ta juede shuo shenme shun henduo? *'What does he think is more fluent to say?'*

Table 1: Example of elicitations under BROADFOCUS and NARROWFOCUS.

In every block of elicitation, there were 30 (= 15 tone combinations × 2 syntactic types) BROADFOCUS trials and 30 NARROWFOCUS trials. The trials were presented in a random order. The blocks were separated by five-minute breaks. The experimenter would ask the speakers to repeat the answer if the experimenter failed to perceive the intended focus. In total, 833 trials were collected.

2.3. Data analysis

The start and the end of both syllables of the target stimuli were manually labelled in Praat [5]. Durations at the syllable level were obtained in MATLAB™. I took into consideration that durations of the target syllables vary with their syllable structures, therefore deriving the DURATIONRATIO—the ratio between the durations of the first syllable (σ_1) and the second syllable (σ_2) within a disyllabic phrase.

The effects of syntactic structure (TYPE: MODN$_h$ and V$_h$OBJ) and discourse context (DISCOURSE: BROADFOCUS and NARROWFOCUS) on DURATIONRATIO were tested using Linear Mixed Models (lme4 [6] in R version 3.2.0). Other variables of fixed effects included tone types of both syllables (TONE$_1$ and TONE$_2$). Stimuli (STIM) and speakers (SPK) were included in the mixed model as variables of random effects.

3. Hypothesis and predictions

Hypothesis i: The nonheads have phrasal stress.

Prediction i: The nonheads will have greater durations under both focus conditions. Therefore, the DURATIONRATIO of MODN$_h$ is larger than that of V$_h$OBJ. Consequently, MODN$_h$ and V$_h$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 MODN$_h$ and V$_h$OBJ phrases.

Prediction ii: Under NARROWFOCUS, durational increase of the nonheads will be greater than their syntactic heads. Therefore, the DURATIONRATIO of MODN$_h$ will increase from BROADFOCUS to NARROWFOCUS, whereas the DURATIONRATIO of V$_h$OBJ will decrease from BROADFOCUS to NARROWFOCUS.

4. Results

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 1).

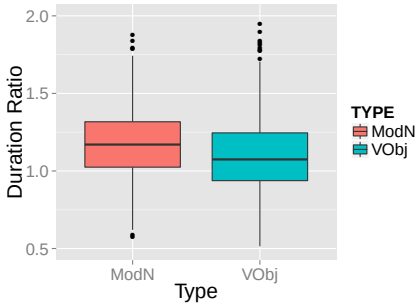

Figure 1: DURATIONRATIO of MODN$_h$ and V$_h$OBJ. Globally, the DURATIONRATIO of MODN$_h$ was larger than that of V$_h$OBJ.

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 2). Moreover, the DURATIONRATIO difference between MODN$_h$ and V$_h$OBJ was more pronounced under NARROWFOCUS (0.097) than under BROADFOCUS (0.057).

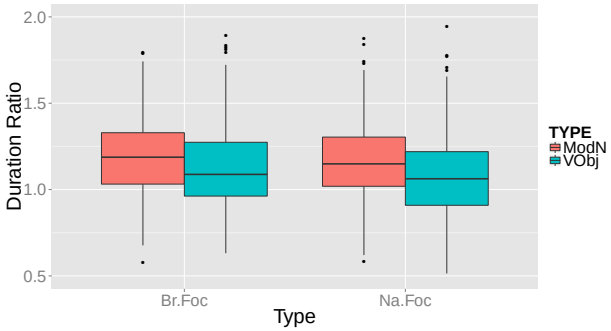

Figure 2: DURATIONRATIO of MODN$_h$ and V$_h$OBJ, grouped by DISCOURSE (BROADFOCUS and NARROWFOCUS). The DURATIONRATIO difference between MODN$_h$ and V$_h$OBJ was more pronounced under NARROWFOCUS than under BROADFOCUS.

Figure 3 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 MODN$_h$ and V$_h$OBJ with DURATIONRATIO

under BROADFOCUS (t(173) = 3.4725, $p < 0.01$).

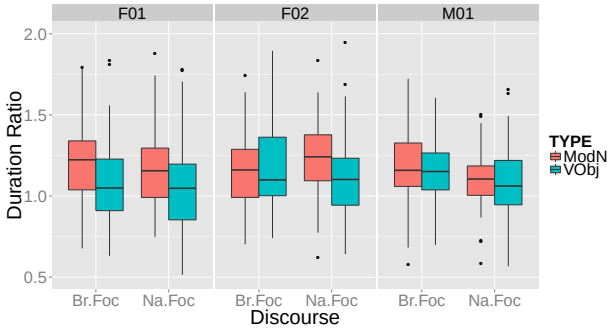

Figure 3: DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ grouped by SPK. Global TYPE effect was observed across speakers; with DURATIONRATIO, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS than under BROADFOCUS, though there existed cross-speaker variations.

Figure 4 shows the DURATIONRATIO grouped by tone combination (TONE_1 + 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).

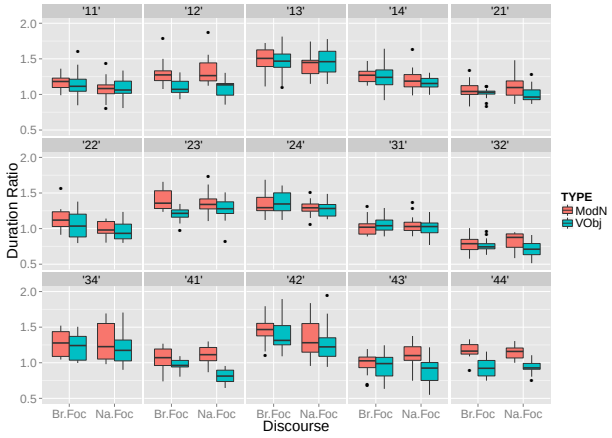

Figure 4: DURATIONRATIO of $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ grouped by tone combination (TONE_1 + TONE_2). Global TYPE effect was observed for the majority of the tone combinations; with DURATIONRATIO, $\underline{\text{MOD}}\text{N}_\text{h}$ and $\text{V}_\text{h}\underline{\text{OBJ}}$ were better differentiated under NARROWFOCUS than under BROADFOCUS, though there existed variations across tone combinations.

The above results shown in Figures 1-4 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 5, 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 2-4, $\underline{\text{MOD}}\text{N}_\text{h}$ 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.

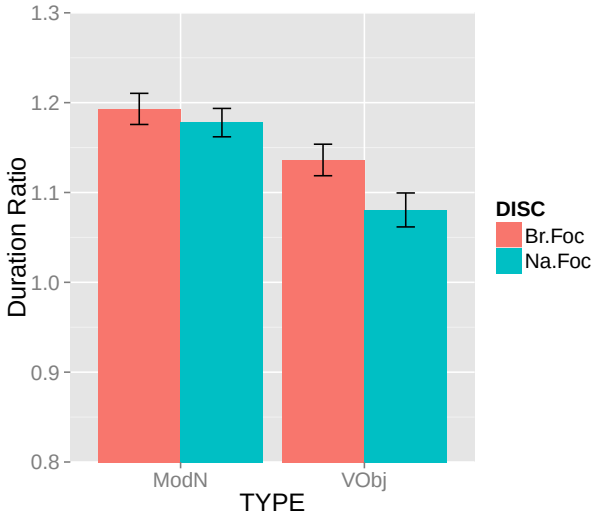

Figure 5: Mean DURATIONRATIO $\pm$ 1 standard error by TYPE and by DISCOURSE. The DURATIONRATIO of $\text{V}_\text{h}\underline{\text{OBJ}}$ significantly decreased from BROADFOCUS to NARROWFOCUS, whereas no significant DURATIONRATIO change was found for $\underline{\text{MOD}}\text{N}_\text{h}$.

Linear mixed model analysis (Table 2) 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 (σ_2) 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_hOBJ	-0.06	22.7	0.0180	**
DISCOURSENARROWFOCUS	-0.02	798	0.3062	
TYPEV_hOBJ :	-0.04	798	0.1115	'
DISCOURSENARROWFOCUS				
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 2: Results of the mixed model analysis on DURATIONRATIO. Significant factors are shown in bold. Interaction between tones were not shown due to space limit.

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 variations.

5. Discussion & conclusion

The DURATIONRATIO difference between MODN$_h$ and V$_h$OBJ 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 [3] that V$_h$OBJ exhibited final stress whereas MODN$_h$ exhibited no initial stress. Furthermore, such a claim would essentially undermine the validity of 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, because the latter 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. Therefore, the discrepancy between Scenario (C) (that V$_h$OBJ has final stress and MODN$_h$ 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 MODN$_h$ 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 [1]. Because the first syllable (σ_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 MODN$_h$, 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 MODN$_h$ and the V$_h$ of V$_h$OBJ, and between the N$_h$ of MODN$_h$ and the OBJ of V$_h$OBJ. A second possible interpretation is that there might exist stimulus-dependent stress patterns that contribute to the overall non-significant DURATIONRATIO change for MODN$_h$. It is possible that the majority of MODN$_h$ stimuli in the current study did not exhibit initial stress, therefore disguising the DISCOURSE effect. Lastly, another possible reason might lie in the choice of acoustic metric in the current analysis. It was found in [2] that, F0, rather than duration, was the phonetic correlate that better reflected the alteration of prosodic strength in both disyllabic and polysyllabic words. More analyses with F0 measurements are under way to look into the distribution of Mandarin phrasal stress.

For these reasons, I will tentatively argue that in line with Scenario (A) (as well as NONHEAD STRESS RULE), the DURATIONRATIO differences between MODN$_h$ and V$_h$OBJ reflect the difference between initial stress and final stress, which is further indicative of two different syntactic structures.

Last but not least, the cross-speaker and cross-stimulus variation needs to be accounted for. While there existed variations, no speakers or stimuli showed patterns that went in the opposite direction of Findings ①–③. For F01 and F02, the DURATIONRATIO of MODN$_h$ was larger than that of V$_h$OBJ; for M01, the DURATIONRATIO of MODN$_h$ and V$_h$OBJ were not differentiable under either BROADFOCUS or NARROWFOCUS. For some homophonous pairs, the DURATIONRATIO of MODN$_h$ was larger than DURATIONRATIO of V$_h$OBJ; for others, the DURATIONRATIO of MODN$_h$ was no different from the DURATIONRATIO of V$_h$OBJ. Therefore, I suggest that such variations are more of idiosyncrasies than randomness, 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, 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 MODN$_h$ and V$_h$OBJ depends heavily on the idiosyncrasies of particular lexical items or individual speakers.

The study strongly suggests that the tendency of contrasting MODN$_h$ and V$_h$OBJ results from NONHEAD STRESS RULE. It is argued that NONHEAD STRESS RULE, despite being weak, exists in Mandarin Chinese, because it helps to facilitate communicative efficiency when needed. Future studies should look into other acoustic correlates such as F0 measurements. 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.

6. References

[1] S. Duanmu, *The phonology of Standard Chinese*, 2nd ed. Oxford University Press, 2007.

[2] C. Lai, Y. Sui, and J. Yuan, "A corpus study of the prosody of polysyllabic words in Mandarin Chinese," in *Proceedings of Speech Prosody 2010*, 2010.

[3] W. Shen, J. Vaissière, and F. Isel, "Acoustic correlates of contrastive stress in compound words versus verbal phrase in Mandarin Chinese," *Computational Linguistics and Chinese Language Processing*, vol. 18, no. 3, pp. 45–58, September 2013.

[4] Y. Jia, "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)]*, vol. 51, no. 9, pp. 1307–1312, 2011.

[5] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]," 2015.

[6] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.