INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Focus Acoustics in Mandarin Nominals

Yu-Yin Hsu, Anqi Xu

Hong Kong Polytechnic University, Hong Kong
yyhsu@polyu.edu.hk, anqi.jy.xu@polyu.edu.hk

Abstract

In addition to deciding what to say, interlocutors have to decide how to say it. One of the important tasks of linguists is then to model how differences in acoustic patterns influence the interpretation of a sentence. In light of previous studies on how prosodic structure convey discourse-level of information in a sentence, this study makes use of a speech production experiment to investigate how expressions related to different information packaging, such as information focus, corrective focus, and old information, are prosodically realized within a complex nominal. Special attention was paid to the sequence of "numeral-classifier-noun" in Mandarin, which consists of closely related sub-syntactic units internally, and provides a phonetically controlled environment comparable to previous phonetic studies on focus prominence at the sentential level. The result shows that a multi-dimensional strategy is used in focus-marking, and that focus prosody is sensitive to the size of focus domain and is observable in various lexical tonal environments in Mandarin.

Index Terms: nominal, information focus, corrective focus, post-focal reduction, Mandarin

1. Introduction

The same sentence can be used to express different information structures, and the way that prosody is used to encode such *information packaging* [1] is unique to each language; some languages use prosody or combine word order variation (e.g., left dislocation in Romance languages) or morphological marking (e.g., different discourse functions carried by Japanese morphemes *–wa* and *–ga*) with prosodic marking to express the full range of possible meanings.

Cross-linguistically, it is acknowledged that the prosodic marking of focus involves interactions between many levels of representation [2] [3]. Tone languages such as Chinese are particularly challenging, since acoustic signals typically associated with the prosodic marking of information structure are at the same time used to distinguish word meanings (e.g., in Mandarin *ma*[high-level] "mother" vs. *ma*[high-falling] "scold"). Characterizing the use of prosody for other purposes therefore requires that such lexical differences be considered. Moreover, terms associated with information structure, such topic/focus, and new/old, are often assumed or defined differently in different studies. Previous research on the prosodic marking of focus in Chinese has mostly emphasized the phonetic prominence of a single focused disyllabic word of Tone 1 (the high-level tone) serving as the subject or object in a Mandarin sentence, but different findings were reported. For example, narrow *wh*-focus may involve longer duration [4], larger f_0 ranges [4] [5], or higher mean f_0 [6], and it is reported that correction is distinguished from old information by longer duration, higher intensity, and larger f_0 range [7]. It remains unclear (a) whether different types of foci (e.g., information focus vs. corrective focus) are prosodically expressed differently, and (b) whether focus representations are acoustically distinguishable from the underlying lexical tones. Attempting to see the whole picture through a controlled and parallel investigation, we took a multi-dimensional approach to study how different information structure roles are realized prosodically (through duration, intensity, and f_0) in the same lexical-phrasal environment, and how they interact with different underlying lexical tones.

In light of previous studies on how prosodic structure convey discourse information, and assuming the framework of alternative semantics of focus [8] [9], we investigated how the size of focus constituents interacts, in terms of prosodic realization, with syntactic position (subject vs. object), and distinct lexical tonal environments. Special attention was paid to a special phrasal environment: the sequence of "numeral-classifier-noun" in Mandarin. Each unit therein expresses a semantic core and syntactic phrase by itself, and this sequence naturally provides a phonetically controlled phrasal environment comparable to previous studies on focus-related phonetic prominence at the sentential level.

2. Method

2.1. Stimuli

The target items were four-syllable complex nominals containing a disyllabic numeral, a monosyllabic measure word, and a monosyllabic noun. Every syllable in the target item bears the same underlying Mandarin tone as follows: tone 1, tone 3, and tone 4. Tone 2 was not included because there is no disyllabic numeral bearing consecutive tone 2. It is known that two adjacent tone 3 syllables in Mandarin often requires the first tone 3 syllable to be pronounced as tone 2 (e.g., *lao3shu3* 'mouse' → *lao2shu3*). Concerning this sandhi phenomenon, we decided to include but distinguish the lexical item *yi* 'one' from other tone 1 items as a separate condition, because *yi* 'one' undergoes obligatorily tone sandhi based on the tone of its following syllable unit (i.e., when its following unit bears tone 1, 2, or 3, *yi* 'one' is pronounced as tone 4; when its following syllable is tone 4, *yi* is pronounced as tone 2). In this study, *yi* 'one' that sandhied to tone 4 and tone 3 words that sandhied to tone 2 were included.

Table 1: *Target items of different tones*

Tones			
Tone 1	三千枝花	san qian zhi hua	"three thousand flowers"
Tone 1 (*yi* 'one'-sandhied)	一千只猪	yi qian zhi zhu	"a thousand pigs"

http://dx.doi.org/10.21437/Interspeech.2017-1167

Tone 3 (sandhied)	五百碗酒	wu bai wan jiu	"five hundred bowls of alcohol"
Tone 4	六万对袜	liu wan dui wa	"sixty thousand pairs of socks"

Such complex NPs in Table 1 were embedded in sentences illustrating the following six different information structures: the answer to a *wh*-NP (ANP), the correction of the whole NP (CNP), the answer to a *wh*-numeral (ANUM), the correction of a numeral (CNUM), the answer to a *wh*-question about a new event (NEWS, i.e., the wide focus referred in previous studies), and when the whole NP is part of the background, old information (ODNP). The target items were manipulated as either the subject or object of a sentence. The stimuli consist of 288 target sentences in total (6 items ×4 tonal conditions ×6 information structures × 2 NP positions). Stimuli were all randomized, so that no identical target item was immediately adjacent in the trials while being presented.

Table 2: *Leading questions and target sentences of six types of information structures*

Information structure	**Leading questions**	**Target sentences**
Answering the whole NP (ANP)	Sheme-dongxi zhuangshi-le hunli? "What was used to decorate the wedding?"	San qian zhi hua zhuangshi-le hunli. "Three thousand flowers was used to decorate the wedding."
Correcting the whole NP (CNP)	Wu bai pen lvluo mai-le yi wan yuan. "Five hundred pots of dill were sold for ten thousand dollars."	Bu, san qian zhi hua mai-le yi wan yuan. "No, three thousand flowers were sold for ten thousand dollars."
Answering the disyllabic numeral (ANUM)	Ji zhi hua jie-le huabao? "How many flowers budded?"	San qian zhi hua jie-le huabao. "Three thousand flowers budded."
Correcting the disyllabic numeral (CNUM)	Liang qian zhi hua mai-le yi wan yuan. "Two thousand flowers were sold for ten thousand dollars."	Bu, san qian zhi hua mai-le yi wan yuan. "No, three thousand flowers were sold for ten thousand dollars."
Answering a full-sentence (NEWS)	Zenme yi fu jiangya de biao qing? "Why do you look surprised?"	San qian zhi hua yao yi wan kuai qian! "Three thousand flowers worth ten thousand!"
NP as a part of the old information (ODNP)	Huadian mai lai de san qian zhi hua zenme yang le. "What happened to three thousand flowers that the flower shop bought?"	San qian zhi hua man man ku wei le. "Three thousand flowers gradually wither away."

2.2. Participants

Six native speakers of *Putonghua* Mandarin from Northern China participated in the experiment (3female; 3male), aged between 20 and 28 (mean: 23.5). None of them reported any history of hearing problems. The ethics approval for the data collection and the basic geographic information were obtained before each participant started the experiment. Each participant was paid HK$60 compensation after the experiment.

2.3. Procedure

Each participant first filled out a language background questionnaire and signed an information consent form. During the experiment, all of the stimuli were presented on a computer screen in a sound-attenuated room. Participants were instructed to listen and response to pre-recorded utterances as casual and natural as possible; no instruction was given to emphasize any token. Participants listened to the leading questions through a headphone and read the target sentences on the screen. Following a given trial, the next was presented 2s later. They only repeated the sentence once unless they mispronounced the words or paused in the middle of utterances. Recordings were made in WAV format at a sampling rate of 44.1 kHz and a 16-bit quantization. Every participant had three practice trails before the experiment. The participants were forced to take a 5-minute break after 144 trials. The experiment lasted about 50 minutes.

2.4. Analysis

The target items were segmented using a custom-written script ProsodyPro [10] for Praat [11]. Syllable boundaries were determined by using both visual (the waveform and spectrogram) and auditory information. The vocal pulses detected by Praat [11] were manually checked and corrected when there were missing pulses, increased pitch on stops, or creaky voice. The following acoustic measurements of each target syllable were generated by ProsodyPro [10] automatically across speakers: duration, mean intensity, and normalized f_0. The normalization of f_0 was realized by dividing each syllable into 10 intervals equal in time and calculating the trimmed f_0 values [12]. The f_0 value was converted from Hz to semitone scale, relative to 1 Hz by the following formula: $12 \ln (x / 1) / \ln 2$.

We conducted Linear Mixed-Effects model on the duration and mean intensity using *lmer()* function [13] in R [14]. The fixed effects were 'information structure', 'tonal condition', and 'NP position'. The fixed effects were only incorporated in the model if they led to a better fit, which was tested with the *anova()* function in R [14]. We also included 'listeners' and 'repetition' as random intercepts. Random slopes for fixed effects were not introduced because it resulted in a model that did not converge. The Satterthwaite approximation for degrees of freedom was used to estimate *p*-values. We encoded NP with old information as the baseline condition. To observe f_0 contour patterns of different foci, Smoothing Spline Analysis of Variance (SSANOVA [15]) was applied to compare the normalized f_0 (in semitone) by using *ssanova()* function from the gss package [16] in R [14] to generate the contour plots. This analysis estimates 95% Bayesian confidence intervals and they were plotted by package ggplot2 [17]. Two conditions are considered significantly different, if the confidence intervals shown in the plot do not overlap.

3. Results

In the following sections, we report results about duration, intensity, and f_0 for each syllable. The attention will be paid to the difference between old information and different foci on the one hand, and acoustic cues related to different foci and their post-focal reduction on the other.

3.1. Duration and intensity

3.1.1. *The first numeral in NP*

The analysis revealed a significant main effect of information structure (F=44.65, p<.001), and tonal condition (F=388.37, p<.001) on the duration of the first numeral. Likewise, the information structure (F=4.755, p<.001) tonal conditions (F=45.692, p<.001), and syntactic condition (F=74.607, p<.001) had significant effect on intensity. All focus conditions showed duration longer than the old information (Table 3). However, only when the numeral was corrected, its intensity was stronger (Table 3).

Table 3: *The effect of information structure on* the *duration and mean intensity of first numeral.*

Fixed effect: Information structure (intercept: ODNP)	**Variable**	***β***	**95% CI**	***p***
ANP	duration	28.44	9.29 – 47.59	**.004**
	intensity	0.57	-1.13 – 2.26	.512
ANUM	duration	44.49	25.33 – 63.64	**<.001**
	intensity	0.88	-0.81 – 2.57	.308
CNP	duration	26.84	7.69 – 46.00	**.006**
	intensity	-0.18	-1.87 – 1.51	.835
CNUM	duration	51.26	32.11 – 70.42	**<.001**
	intensity	2.00	0.31 – 3.70	**.020**
NEWS	duration	32.07	12.92 – 51.23	**.001**
	intensity	0.36	-1.34 – 2.05	.680

3.1.2. *The second numeral in NP*

The information structure (F=26.85, p<.001), tonal condition (F=864.81, p<.001), and syntactic condition (F=13.20, p<.001) showed significant main effects on the duration of the second numeral. The duration of second numeral in NP was significantly lengthened when the numeral was corrected (β = 12.96, t(1706) = 2.188, p = .029). Likewise, there is a significant effect of information structure (F=4.297, p<.001), tonal condition (F=252.335, p<.001), and syntactic condition (F=164.937, p<.001) on the intensity. Yet, the intensity of the second numeral across information structures did not show significant differences from the baseline condition.

3.1.3. *The measure word in NP*

The result showed that information structure (F=2.611, p=0.02) and tonal conditions (F=48.53, p<.001) had significant main effects. The duration of measure word in focus conditions was significantly or marginally significantly longer than the duration of old information (Table 4). The information structure (F=17.462, p<.001), tonal condition (F=9.995, p<.001), and NP position (F= 199.258, p<.001) also had a significant main effect on intensity. When the preceding numeral was corrected, the decrease in intensity of the measure word was significant (Table 4).

Table 4: *The effect of information structure on* the *duration and mean intensity of measure word.*

Fixed effect: Information structure (intercept: ODNP)	**Variable**	***β***	**95% CI**	***p***
ANP	duration	23.47	8.49 – 38.46	**.002**
	intensity	-0.12	-1.72 – 1.47	.878
ANUM	duration	17.49	2.51 – 32.48	**.022**
	intensity	-1.26	-2.85 – 0.34	.124
CNP	duration	14.66	-0.33 – 29.64	.055
	intensity	-0.86	-2.46 – 0.74	.293
CNUM	duration	22.46	7.48 – 37.45	**.003**
	intensity	-1.85	-3.45 – -0.25	**.024**
NEWS	duration	15.99	1.00 – 30.97	**.037**
	intensity	-0.07	-1.67 – 1.53	.932

3.1.4. *The noun in NP*

The result showed significant main effects of information structure (F= 22.959, p<.001), tonal condition (F=42.118, p<.001) and syntactic condition (F=33.091, p<.001) on the duration of the noun. Only when the whole noun phrase was focused (ANP and CNP conditions), the duration was significantly longer than when it was in the old information (Table 5). The information structure (F=33.14, p<.001), tonal condition (F=66.19, p<.001) and syntactic condition (F= 373.24, p<.001) also showed significant main effects on intensity. Remarkably, when the preceding numeral was focused, the reduction in intensity was significant compared with the old information (Table 5).

Table 5: *The effect of information structure on* the *duration and mean intensity of noun.*

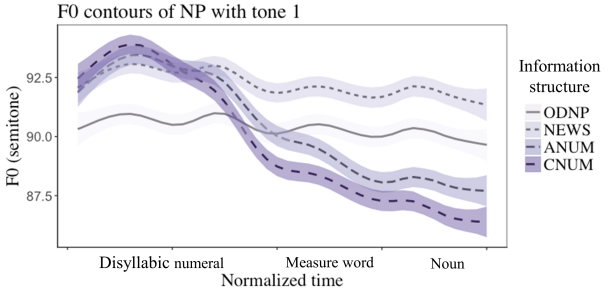
Fixed effect: Information structure (intercept: ODNP)	**Variable**	***β***	**95% CI**	***p***
ANP	duration	33.20	9.85 – 56.55	**.005**
	intensity	-0.14	-1.77 – 1.49	.865
ANUM	duration	9.12	-14.23 – 32.47	.444
	intensity	-1.88	-3.51 – -0.25	**.024**
CNP	duration	31.45	8.10 – 54.80	**.008**
	intensity	-0.64	-2.27 – 0.99	.441
CNUM	duration	7.20	-16.15 – 30.55	.546
	intensity	-2.73	-4.36 – -1.10	**.001**
NEWS	duration	20.58	-2.77 – 43.93	.084
	intensity	-0.65	-2.28 – 0.98	.432

3.2. f_0 contours

3.2.1. *Numeral as Focus*

The plots from SSANOVA indicate that the f_0 contours of sentences with the numeral expressing information focus (ANUM) and corrective focus (CNUM) exhibited on-focus f_0 rise and post-focal compression. As Figure 1 shows, the f_0 contours coincided with the focused numeral position. Both types of numeral foci displayed clear post-focus compression. The effect of compression was significantly stronger in corrective focus (CNUM) than that in ANUM condition.

We also observed that items with tone 3 showed a distinct pattern from other tonal conditions. Specifically, when the numeral was focused, only the f_0 of the following numeral was significantly higher than that of old information. The post-focus reduction of f_0 was significant in the region of measure word but not significantly different in the region of noun.



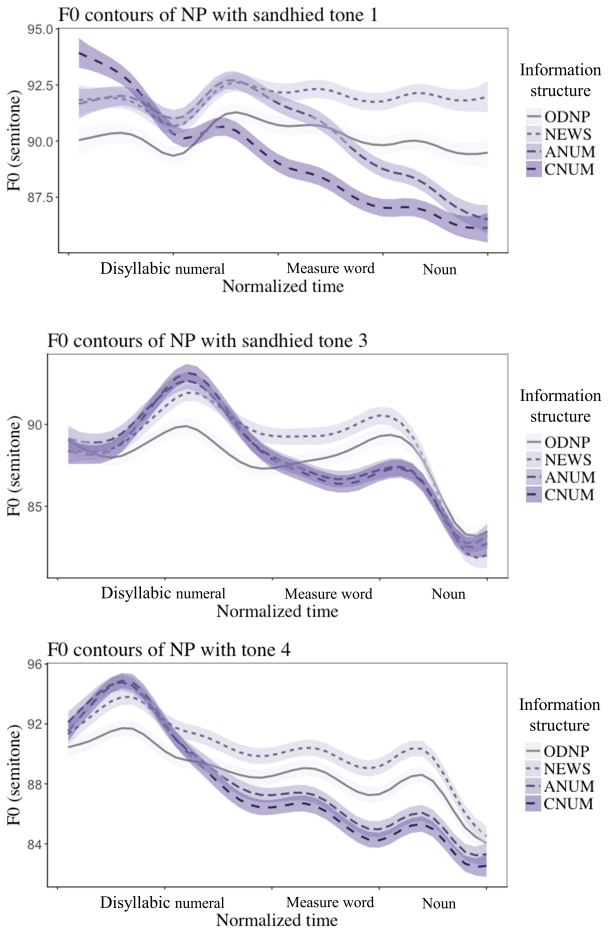


Figure 1: *Normalized f_o contours of NPs in four tonal conditions across four information structures calculated by SSANOVA.*

3.2.2. Whole NP as Focus

As shown in Figure 2, when the whole NP was (part of a) focus (i.e., NEWS, ANP, and CNP conditions), the overall f_o was higher than NPs expressing old information. Interestingly, the f_o contour of NP extracted from the wide-focus (NEWS) did not significantly differ from that of narrow-NP foci (ANP, CNP). Yet, the f_o of every word in these three focus conditions was higher than words in the old information.

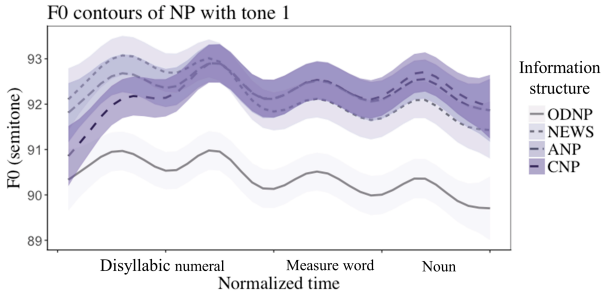


Figure 2: *Normalized f_o contours of whole NP with tone 1 across four different information structures calculated by SSANOVA.*

4. Discussion

The result showed that even within the nominal domain, focus constituents in general are different from constituents expressing old information. The former shows longer duration, higher intensity, and higher overall f_o. With respect to narrow foci (i.e., answering the numeral (ANUM) and correcting the numeral (CNUM) in this study), the analysis of duration and intensity showed that the focused phrase was significantly lengthened, the duration and intensity was mostly only marginally different on its following measure word, the significant reduction was found in intensity and duration of the post-focal region, and significantly lower post-focal f_o contours. The reduction of f_o varied with tonal conditions, namely, different from sandhied tone 1, sandhied tone 3 did not show clear post-focal compression. We suppose that this may be due to the intrinsic contour nature of tone 3. However, we did observe some cross-speaker variations in the tone 3 sandhi patterns. Due to the limit of current study, we cannot address this issue fully. We leave it for future study.

Although the design of the current study does not allow us to study post-focal patterns of full NP foci, results from narrow foci still showed clear patterns of post-focal compression. Despite of a small number of speakers, the results of different acoustic dimensions are nevertheless highly similar to what was reported for *wh* and corrective foci at the sentential level. In addition, our results show how types of foci can be differentiated. Through the analysis of narrow foci, it was shown that the higher initial intensity and the greater post-focally compressed f_o are the most prominent acoustic indication that distinguishes corrective foci from *wh*-foci.

5. Conclusions

This study attempted to investigate how information structure is realized prosodically within a complex Mandarin nominal expression. The syntactic environment adopted in this study allowed us to investigate prosodic organization of information structures from various levels. Although patterns varied depending on the size of foci, the lexical tonal conditions, or the type of information structure involved, the results in this study showed that a multi-dimensional strategy was used in marking foci in Mandarin. Furthermore, our study showed that although acoustic cues of f_o, intensity, and duration are important in expressing lexical information in Chinese, different information structural roles can still be expressed distinctly in Mandarin through this prosodic system, and that the acoustic realization of focus is not simply syllable-based, but is sensitive to the size of the constituent that expresses specific information structure.

6. Acknowledgements

We would like to express our gratitude to James Sneed German for his insightful comments and suggestions during the discussion of the earlier stage of this study. We would also like to thank Sui Li, and Ka Keung Lee for their technical support. Mistakes remaining are exclusively our own.

7. References

[1] W. L. Chafe, "Givenness, contrastiveness, definiteness, subjects, topics and point of view," in *Subject and Topic*, New York, Academic Press. , 1976, pp. 27-55.

[2] C. Gussenhoven, "Focus, mode, and nucleus.," *Journal of Linguistics. ,* pp. 377 - 417, 1983.

[3] J. Pierrehumbert & J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in communication*, MIT Press, 1990, pp. 271-311.

[4] S. Jin, An acoustic study of sentence stress in Mandarin Chinese, Columbus, OH: Ohio State University., 1996.

[5] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics,* vol. 27, pp. 55-105, 1999.

[6] S. W. Chen, B. Wang and Y. Xu, "Closely related languages, different ways of realizing focus.," in *Proceedings of Interspeech*, Brighton, UK., 2009.

[7] I. Ouyang and E. Kaiser, "Prosody and information structure in a tone language: an investigation of Mandarin Chinese," *Language, Cognition and Neuroscience,* vol. 30, pp. 57-72, 2015.

[8] M. Rooth, "A theory of focus interpretation," *Natural Language Semantics,* pp. 75-116, 1992.

[9] M. Krifka, "Basic notions of information structure," in *Interdisciplinary Studies of Information Structure 6*, Potsdam, 2007.

[10] Y. Xu, "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis," in *TRASP 2013*, Aix-en-Provence, France, 2013.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," [Online]. Available: http://www.praat.org. [Accessed 05 March 2017].

[12] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics,* vol. 25, pp. 61-83, 1997.

[13] A. Kuznetsova, P. B. Brockhoff and R. H. Bojesen Christensen, "lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) R package," [Online]. Available: http://CRAN.R-project.org/package=lmerTest. [Accessed 25 February 2017].

[14] R. C. Team, "R: A language and environment for statistical computing R Foundation for Statistical Computing," [Online]. Available: http://www.R-project.org/. [Accessed 25 February 2017].

[15] C. Gu, Smoothing Spline ANOVA Models, New York: Springer, 2002.

[16] C. Gu, "gss: General Smoothing Splines," 24 February 2017. [Online]. Available: https://cran.r-project.org/web/packages/gss/index.html. [Accessed 06 March 2017].

[17] H. Wickham, W. Chang and RStudio, "ggplot2," 30 December 2016. [Online]. Available: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. [Accessed 6 March 2017].

[18] R. C. Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria (2014) (Version 3.1.0).