A PREPLIMINARY STUDY OF THE TEMPORAL RELATIONSHIP BETWEEN PROSODY AND GESTURE IN HONG KONG CANTONESE

Holly Sze Ho Fung, Peggy Pik Ki Mok

The Chinese University of Hong Kong
hollyfung_cuhk@yahoo.com.hk, peggymok@cuhk.edu.hk

ABSTRACT

Previous studies of speech and gesture in intonational languages generally suggested that prosodic and gestural prominence are aligned with one another, pitch accented/ stressed syllable or the peak fundamental frequency (F0) of it being the prosodic anchor. A logical question to raise would be whether such alignment exists in tonal languages without lexical stress. To answer the question, this study investigated the timing of pointing gestures relative to their co-occurring corrective foci in Hong Kong Cantonese in a picture-naming task. Results show that prosodic prominence on the focused syllables was solely realized by durational increases. However, the occurrence of gestural apices was found insensitive to such changes. Neither was it found to be affected by different lexical tones of the foci.

Keywords: Gesture and prosody, Hong Kong Cantonese, focus prosody

1. INTRODUCTION

An early proposal of temporal relation between prosody and gesture was from McNeill [1], who stated in his Phonological Synchrony Rule that gestural prominence does not occur randomly but aligns with speech prominence. The rule finds evidence support from a range of empirical studies of different intonational languages in recent years, a number of which looking specifically into the temporal correlation between gestural apex and prosodic stress on focus. By manipulating the position of contrastive focus on English disyllabic compound nouns, Rusiewicz and Shaiman [2] found significant effect of stress position on the duration and onset time of pointing gestures. In similar experiments in Catalan in which stress patterns of targets were controlled, prosodic stress was found to have significant effect on the timing of the apexes as well as the duration of pointing gestures and head nods [3][4]. Other languages reported to show prosody-speech alignment include French [5] and Brazilian Portuguese [6]. While there were different views on what the prosodic anchor to which gestural prominence alignment should be—for instance some suggested it to be the focused syllable, e.g. [2], while some pinpointed it to the F0 peak of that syllable [2][4]—it was generally agreed that prosodic and gestural prominence are temporally correlated.

Nonetheless, such correlation is only based on findings in intonational languages. If the prosodic anchor of prosody-gesture alignment is stressed syllable, or its F0 peak, one question arises: does the correlation exist in tonal, non-stress languages, in which lexical stress is nonexistent? Hong Kong Cantonese (HKC), a tonal language with a complex tone system consisting of both dynamic and static tones (see Table 1) but no lexical stress, provides a good testing ground to see whether such prosody-gesture coordination is language-universal.

Table 1: Summary of Cantonese lexical tones

	T1	T2	T3	T4	T5	T6
Tone shape	high level	high rising	mid level	low falling	low rising	low level
Pitch values	55	25	33	21	23	22

2. METHOD

2.1. Speakers

4 native speakers of HKC (3 male and 1 female, all students of a local university) participated in a picture-naming task. None of them had any reported visual, speech or hearing impairments and all received course credits for their participation.

2.2. Materials

12 monosyllabic and 9 disyllabic Cantonese keywords were used (see Table 2), covering one level tone (T1), one rising tone (T2) and one falling tone (T4). All were elicited in the carrier frame *Hai/M hai, jau CL [X] hai CL [Y] soeng min* "Yes/No, there is a [X] above the [Y].", prompted by the question *Hai mai jau CL [X] hai go [Y] soeng min aa?* "Is there a [X] above the [Y]?" Each monosyllabic keyword was elicited in three

focus conditions: the 1) neutral focus, 2) corrective focus, and 3) post-focus conditions; each disyllabic keyword was elicited in four conditions: the 1) neutral focus, 2) focus-on-first-syllable, 3) focus-on-second-syllable, and 4) post-focus conditions. Note that corrective foci were only elicited in slot [X]. Details of how different focus conditions were elicited are given in section 2.3.

In addition, powerpoint slides were used to present auditory-visual stimuli. Each slide showed pictures of two objects from the keyword lists, one above the other, and was embedded with a pre-recorded prompt question, which was spoken without emphasis on any of the syllables.

Table 2: Keywords (transcribed in Jyutping) used in the experiment

Tone	Monosyllabic	Disyllabic
T1	dou "Knife"	ceng ziu "Green pepper"
	maau "Cat"	faa zeon "Vase"
	ze "Umbrella"	sai gwaa "Watermelon"
	zung "Clock"	
T2	bong "Scale"	seoi sau "Sailor"
	caang "Orange"	soeng gaa "Photo frame"
	zeng "Well"	so lin "Chain"
	zi "Paper"	
T4	cong "Bed"	joeng to "Llama"
	long "Wolf"	ngau jau "Butter"
	se "Snake"	pei haai "Leather shoe"
	syun "Ship"	

2.3. Procedures

The task, the method of focus and gesture elicitation of which was adopted from [2], was conducted in a quiet room, in which visual materials were presented on a screen approximately 1 meter away from where the speakers were seated. Two camcorders were positioned in front of and next to the speakers to record their gestures at the rate of 25 frames per second. Each of the camcorders were connected to an external recorder, which was placed near the speakers to record their speech at the sampling rate of 44100Hz and 16 bits.

Before the task started, speakers were first taught the mappings between pictures of the keyword objects and their corresponding labels (aurally presented), tested on their memory of them, and then trained to respond differently to two kinds of prompt questions. When what the prompt question asked matched with what was presented on the screen, they were instructed to respond verbally by repeating the prompt in a statement. On the other hand, when what the prompt asked differed from what they saw on the screen, apart from correcting the mistaken object using the carrier frame, they were instructed to point at the correct object on the screen, imagining that the person who asked the question could not see it clearly.

2.4. Data analysis

2.4.1. Acoustic data

Syllables of the keywords were annotated using Praat, each measured for its: 1) duration, 2) mean F0, and 3) F0 range. F0 values, obtained from 10 equal-distant points of the sonorant part of each syllable, were semitone-transformed using the formula $ST=12*log(f/f_{ref})/ log2$, f being the frequency to be transformed and f_{ref} being the reference frequency, which was 55Hz for the male speakers and 100Hz for the female speaker.

2.4.2. Gestural data

Videos of the front and side views of each speaker were synchronized using Adobe Premiere Pro CC, and annotated using ELAN. Considering the initial (final) movement of the pointing arm as the onset (offset), and the maximal extension of the arm and the index finger as the apex of a pointing gesture, 2 intervals and 2 ratios were measured and calculated for each gesture, namely the 1) gesture onset-word onset (GO-WO) and 2) gesture offset-word offset (GF-WF) intervals, and the 3) apex-word and 4) apex-focus ratios. The interval measures show the general picture of the gesture's timing relative to the word with focus (note: the *whole word* instead of only the focused syllable for a disyllabic keyword). The apex-word ratio, i.e., the division of the keyword's duration by the interval between the apex and word onset, shows how close the apex is to the word. Analogously, the apex-focus ratio measures the relative distance of the apex from the focused syllable. A value below 0 (above 1) indicates that the apex occurred before (after) the keyword/ focused syllable.

3. RESULTS

3.1. Acoustic results

3.1.1. Duration

Mean durations (averaged across 4 speakers, in milliseconds) of mono- and disyllabic keywords are shown in Figures 1 and 2 (error bars showing

95% confidence intervals). Repeated-measure ANOVAs showed a significant effect of focus on both monosyllabic targets, F(1,3)=14.683, p=0.031, and disyllabic ones, F(2,6)=169.252, p=0.000. Bonferroni post hoc tests found that both types of keywords were significantly longer in the corrective focus condition than in the other condition(s). (Results of the post-focus condition, which are less relevant to the evaluation of gestural timing, are not discussed here (the same in sections 3.1.2 and 3.1.3)).

Figure 1: Durations of monosyllabic keywords

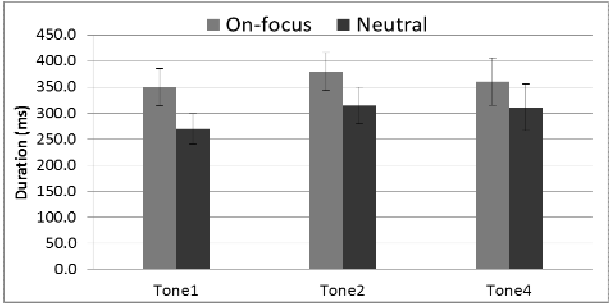
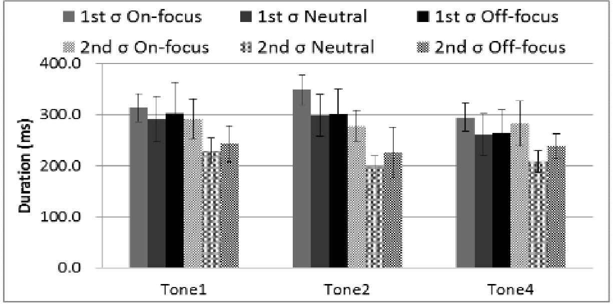


Figure 2: Durations of disyllabic keywords



3.1.2. *Mean F0*

Averaged mean F0s of individual syllables (in semitones) of mono- and disyllabic keywords are shown in Figure 3 and 4. As expected, keywords of the high level Tone 1 had the highest mean F0s, followed by those of the high rising Tone 2, and finally by the low falling Tone 4. The effect of focus was found insignificant on both types of keywords.

Figure 3: Mean F0s of monosyllabic keywords

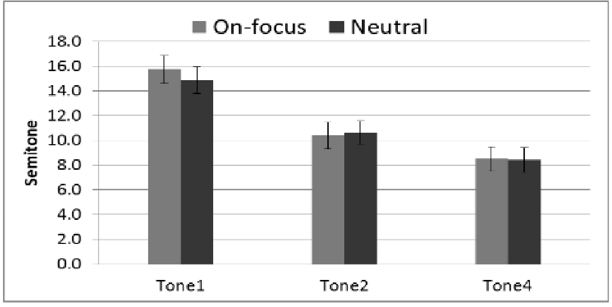
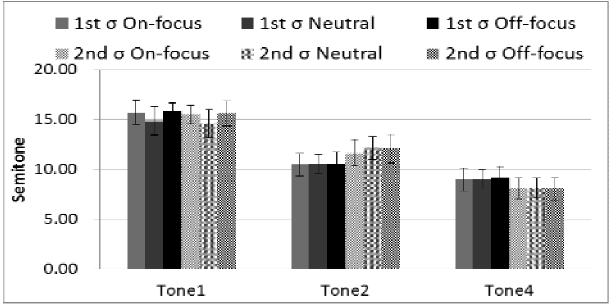


Figure 4: Mean F0s of disyllabic keywords



3.1.3. *F0 range*

Figures 5 and 6 show the averaged F0 ranges of mono- and disyllabic keywords (in semitones). Again as expected, static Tone 1 had a remarkably smaller averaged F0 ranges than the other two contour tones in both types of keywords. Since the distribution of the individual F0 range values was found to be positively skewed, square root transformation was done before repeated-measure ANOVAs were performed. Results showed that tone had a significant effect on both F0 ranges of monosyllabic keywords (F(2,6)=22.852, p=0.002) and disyllabic ones (F(2,6)=14.666, p=0.005), the pitch range of Tone 1 being significantly smaller than the other two.

Figure 5: F0 ranges of monosyllabic keywords

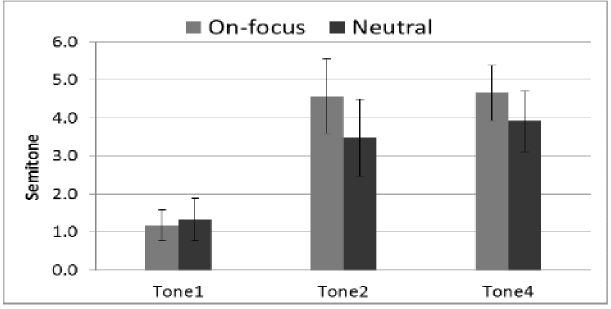
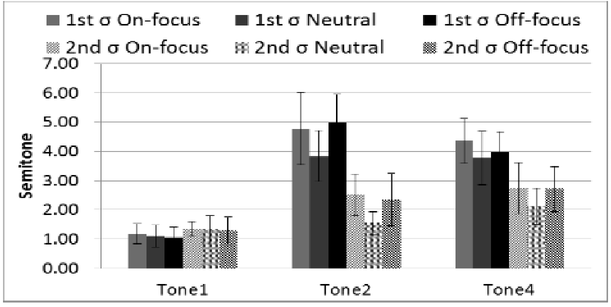


Figure 6: F0 ranges of disyllabic keywords



3.2. Gestural results

Mean values (standard deviations) of three gestural measures are summarized in Table 3. First, all the mean GO-WO intervals are positive, suggesting that pointing gestures generally started ahead of the keywords, regardless of tone, number of syllable, and position of the focused syllable (for disyllabic keywords). Second, all WO-apex/ word length ratios are between 0 and 1, indicating that gestural apices generally occurred within the span of their corresponding keywords, again regardless of all the manipulated factors. In fact, among all the 156 instances of pointing, none started after the onset of the keyword, and only 7 had the apex occur out of the span of the word, among which only one preceded it (by 99 ms). Third, all mean GF-WF intervals are negative, suggesting that pointing generally ended after the keywords (with only 3 exceptional instances where the gestures ended ahead of the word offsets by 12 to 128 ms).

Table 3: Summary of gestural measures, averaged across 4 speakers (for the disyllabic section, results of 1st-syllable foci are shown in grey cells and those of 2nd-syllable foci in white)

Syllable no. and tone		Apex-word ratio	GO-WO (ms)	GF-WF (ms)
Mono	T1	.42 (.21)	355 (129)	-953 (468)
	T2	.50 (.31)	314 (96)	-861 (375)
	T4	.94 (.39)	381(92)	-971 (576)
Di	T1	.49 (.26)	316 (147)	-908 (626)
		.43 (.21)	369 (123)	-910 (760)
	T2	.46 (.27)	377 (141)	-887 (604)
		.42 (.27)	317 (99)	-865 (688)
	T4	.37 (.24)	367 (151)	-830 (550)
		.38 (.31)	381 (25)	-844 (714)

Table 4: Apex-focus ratios of disyllabic pointing gestures of individual speakers, each averaged across 3 trials (abbr.: M—male speaker; F—female speaker)

	1st-syllable focus			2nd-syllable focus		
	T1	T2	T4	T1	T2	T4
M1	0.59	0.43	0.57	**-0.60**	**-1.93**	**-0.86**
M2	0.78	0.99	0.38	**-0.30**	0.02	0.55
M3	0.55	0.38	0.39	**-0.51**	**-1.05**	**-1.12**
F1	**1.56**	**1.29**	**1.29**	0.41	0.54	0.43

Repeated-measure ANOVAs showed no significant effect of tone on the three measures for monosyllabic keywords, and no significant effect of tone, focus position or their interaction for disyllabic ones, suggesting that the gestures were homogeneous in all combinations of factors.

The apex-focus ratios (see Table 4) of individual speakers show more clearly that gestural apices generally co-occurred consistently with the same syllable regardless of tone and focus position. Ratios indicating out-of-sync apices are highlighted in bold. For speakers M1 and M3, apices co-occurred with the first syllables for both 1st- and 2nd-syllable focus conditions, and the reverse patterned was exhibited by speaker F1.

4. DISCUSSION

In contrast to previous studies on intonational languages, our data seem to provide little support to the existence of alignment between prosodic and gestural prominence in HKC. While it is not entirely surprisingly that tone had minimal effect on the timing of gestural apices given that F0 is not a reliable acoustic correlate of prosodic prominence in the language, as shown by F0 results of corrective foci in sections 3.1.2 and 3.1.3 and some previous studies [7], it is interesting to find that duration seems not to have an effect on the timing of gestural apices either. As suggested by our results and those of other studies of focus prosody of HKC [8][9], duration is a consistent acoustic correlate of prosodic prominence in the language. If gestural prominence were to be temporally related to prosodic prominence in HKC, gestural apices should have occurred earlier when focus was placed on the first syllables of the disyllabic keywords than when on second syllables, reflected by significant difference between the two groups of apex-word ratios as well as all-positive apex-focus ratios. Neither did gestures of the two types of focus differ with respect to their onset and offset time, as shown by the insignificant effect of focus on GO-WO and GF-WF intervals.

One possibility could be that gestural apices were not aligned with *syllables* on which prosodic prominence was assigned, but the *words* containing them. This would explain why speakers M1, M2 and F3 had their gestural apices consistently “misaligned” to one syllable when the other was focused in disyllabic keywords. Further research on the hypothesis is currently underway with data collected from more speakers being analyzed.

5. REFERENCES

[1] D. McNeill, *Hand and Mind: What gesture reveals about thought*. Chicago: University of Chicago Press, 1992.

[2] H. L. Rusiewicz and S. Shaiman, “Effects of prosody and position on the timing of deictic gestures,” *J. Speech, Lang. Hear. Res.*, vol. 56, no. April, pp. 458–471, 2013.

[3] N. Esteve-gibert and P. Prieto, “Prosodic Structure Shapes the Temporal Realization of

Intonation and Manual Gesture Movements," vol. 56, no. 850, pp. 850–865, 2013.

[4] N. Esteve-gibert, J. Borràs-comes, M. Swerts, P. Prieto, and U. P. Fabra, "Head gesture timing is constrained by prosodic structure," in *Proceedings of Speech Prosody 2014*, 2014, pp. 356–360.

[5] B. Roustan and M. Dohen, "Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus," in *Speech Prosody 2010*, 2010.

[6] A. Rochet-Capellan and R. Laboissière, "The speech focus position effect on jaw–finger coordination in a pointing task," *J. Speech*, …, vol. 51, no. December 2008, pp. 1507–1521, 2008.

[7] H. Fung and P. Mok, "Realization of Narrow Focus in Hong Kong English declaratives—a Pilot Study," in *Speech Prosody 2014*, 2014, pp. 964–968.

[8] W. Wu and Y. Xu, "Prosodic focus in Hong Kong Cantonese without post-focus compression," in *Speech Prosody 2010*, 2010, pp. 1–4.

[9] V. Man, "Focus effects on Cantonese tones: An acoustic study," in *Speech Prosody 2002, International Conference*, 2002, pp. 2–5.