Journal of Phonetics 71 (2018) 113–125

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Temporal coordination between focus prosody and pointing gestures in Cantonese ☆

Holly Sze Ho Fung, Peggy Pik Ki Mok *

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong

ARTICLE INFO

Article history:
Received 11 July 2017
Received in revised form 19 July 2018
Accepted 23 July 2018

Keywords:
Co-speech gesture
Focus prosody
Cantonese
Multimodal communication

ABSTRACT

This study investigates the temporal relationship between focus prosody and co-speech pointing gestures in Hong Kong Cantonese. Previous studies have generally shown a close temporal proximity between prosodic and gestural prominence: Gestural prominence tends to be aligned with stressed syllables or words. However, this finding was based solely on studies of stress and pitch-accent languages, and no study has yet tested the phenomenon in a non-stress tone language. Ten native speakers of Hong Kong Cantonese participated in a picture-verification task in which pointing was elicited along with verbal corrections. The acoustic results showed that the corrective focus was marked solely by an on-focus durational increase. The gestural results revealed that there was an alignment between prosodic and gestural prominence, as most of the gesture apices were produced within the focused words. However, in contrast to previous findings, no significant effect of F0 (tone) or focus position was found. Instead, most speakers consistently aligned their apices with the same syllable position in disyllables. Based on the current findings, the prosodic anchor of prosody-gesture alignment is suggested to be the focused word in this language.

1. Introduction

1.1. The phonological synchrony rule

Speech communication is essentially multimodal. Apart from spoken words, information is also conveyed by co-speech gestures, such as bodily movements that one produces as one speaks, which include both manual movements (for example, pointing and hand beats) and non-manual movements (such as head movements).

Despite the formal differences between speech and gesture, McNeill (1992, 2005), among others (Goldin-Meadow, 1998; Kendon, 1972, 2000, 2004; Kelly, Manning, & Rodak, 2008), have argued that gesture and language comprise one system. He gave five reasons for his argument:

(1) Gestures occur almost exclusively during speech;
(2) both speech and gestures convey similar if not the same semantic meanings, in addition to serving identical pragmatic functions;
(3) the two modalities exhibit temporal synchrony;
(4) the development of gestures in children mirrors that of language, both beginning with concrete deixis and ending on the discourse level; and
(5) impaired speech and gestures in aphasic patients show parallel patterns, as both lack coherence but preserve meaning in Broca's aphasics, and exhibit fluency without interpretability in Wernicke's ones.

All the evidence shows not only similarities between speech and gestures on the surface level, but also suggests a shared underlying system processing both of them.

With regard to their regular resemblance in semantic meaning, pragmatic functions and timing (reasons (2) and (3) above), McNeill (1992) proposed three synchrony rules for speech and gesture, namely the semantic, pragmatic and phonological synchrony rules. The first two state that co-occurring speech and gesture present and perform the same meaning and pragmatic functions, whereas the last, phonological synchrony, which is the most relevant to the present study, states that "the stroke of the gesture precedes or ends at, but

☆ This work was supported by the Graduate School of the Chinese University of Hong Kong awarded to the first author (postgraduate studentship, 2013–2015).

* Corresponding author.
E-mail addresses: hollyfung@link.cuhk.edu.hk (H.S.H. Fung), peggymok@cuhk.edu.hk (P.P.K. Mok).

https://doi.org/10.1016/j.wocn.2018.07.006
0095-4470/

does not follow, the phonological peak syllable of speech" (p. 26). The stroke here refers to the only obligatory and the most prominent part of a gesture, preceded and followed optionally by the preparation and recovery phases (Kendon, 1972, 1980; Kita, 1990; McNeill, 1992) (although the apex of the stroke has been considered the unit of gestural prominence instead in many recent studies, as will be reviewed in Section 1.2). In other words, what the rule suggests is, when speech and accompanying gestures unfold, the most prominent parts of both channels are linked temporally.

1.2. Previous studies on the temporal relationship between prosodic and gestural prominence

A number of production studies have been conducted to investigate the occurrence of gestural prominence relative to prosodic prominence in simultaneous speech in different intonation and pitch-accented languages. They can be categorised into four groups according to whether not or an alignment between prominent units in speech and gesture was found and, if it was, if the prominent unit in speech, or the *prosodic anchor*, was (a) a stressed/accented word, (b) a stressed/accented syllable, or (c) an F0 peak. The four groups of studies are reviewed as follows.

A few studies found no effect of the change in lexical/nuclear stress position on the timing of gestural prominence, usually measured by the apex (in other words, the maximal displacement of the gesturing hand/body part). For example, De Ruiter (1998, Experiment 1) found that a change in the metrical structures (stress-initial versus stress-final) of nouns elicited in definite *determiner + noun* responses in Dutch had no significant effect on the apex times of accompanying pointing gestures, although apices did occur before accented syllables as predicted by the phonological synchrony rule. Furthermore, in a picture-naming task, Rusiewicz, Shaiman, Iverson and Szuminsky (2013) elicited pointing gestures co-produced with American English sentences, in which contrastive stress was place on either the first or the second syllables of the target words, which were dimorphemic, trochaic compound nouns. In line with De Ruiter's (1998) finding, the results showed no significant effect of contrastive stress position on the timing of the gesture apex.

Nonetheless, more studies have provided evidence for close temporal alignment or covariation between prosodic and gestural prominences, although with different suggestions regarding the prosodic anchor of alignment. Some have suggested that it is the stressed/focused word. For example, Roustan and Dohen (2010) elicited contrastive focus on either the subject or the object (both being CVCV words) in simple SVO sentences in French, which were accompanied by pointing, beat or control (in other words, button pressing) gestures. They found that the apices of the pointing gestures were consistently aligned with the articulatory target of one of the vowels of the focused word.

There is also evidence for the prosodic anchor being the pitch-accented/stressed syllable rather than the word carrying that syllable. Following the experimental settings of his first experiment, De Ruiter (1998, Experiment 2) elicited pointing gestures co-occurring with Dutch noun phrases in the structure *definite determiner + colour adjective + noun*, in which contrastive stress was placed on four possible positions (two content words × two metrical structures). The results showed a significant positive correlation between the beginning of an apex and the onset of a stressed syllable. By contrast, alignment between the gesture apex and stressed syllable is achieved differently in Brazilian Portuguese. In a study by Rochet-Capellan, Laboissière, Galván and Schwartz (2008), speakers of Brazilian Portuguese were asked to identify and point at pictorial targets, which had either trochaic or iambic labels. The results showed that the stressed syllables of both trochaic and iambic words were in sync with gesture apices (that is, the period of time during which the finger remained pointing at the pictorial target), but in different ways: Stressed syllables of the trochees were aligned with the *beginning* of apices, whereas those of the iambs were aligned with the *end* of them. These experimental results were in line with observations from naturalistic studies. Examining manual gestures produced by American English speakers during natural conversations, Loehr (2012) found that pitch accents were only +17 milliseconds ahead of the nearest gesture apices on average. Also studying spontaneous gestures accompanying English speech, Jannedy and Mendoza-Denton (2005) found that 95.7% of all the observed apices co-occurred with a pitch accent. Evidence of a close alignment is also provided by studies on non-manual gestures. For example, Esteve-Gibert, Borràs-Comes, Swerts and Prieto (2014) investigated head movements produced by Catalan speakers in a semi-spontaneous setting in which target words with different metrical patterns were elicited naturalistically, and found that the apices of the head gestures were aligned with accented syllables. Similarly, Ambrazaitis, Lundmark and House (2015) and Ambrazaitis and House (2017) found that head beats and eyebrow movements were closely associated with focal pitch accents in Swedish broadcast news.

Yet another view is that it is the F0 peak of the pitch-accented/stressed syllable that attracts gestural prominence. Leonard and Cummins (2011) studied elicited beat gestures co-occurring with English sentences and found that, among three different possible speech landmarks of speech-gesture alignment, including the rhythmic pulse (P-centre), the vowel onset and the F0 peak of the accented syllable, the gesture apex was aligned closest to the F0 peak. Similarly, in a controlled setting, in which corrective focus structures in Catalan were elicited simultaneously with pointing gestures, Esteve-Gibert and Prieto (2013) found that the correlation between the gesture apex and the F0 peak was the strongest when compared to other pairs of speech and gestural prominent units, including

(1) the apex and the end of the accented syllable,
(2) the stroke onset and the F0 peak, and
(3) the stroke offset and the F0 peak.

As reviewed above, a number of prosodic units have been proposed as the prosodic anchor of speech-gesture coordination. They include (from larger to smaller)

(1) the focused/accented word in a sentence,
(2) the stressed/accented syllable of that word, and
(3) the F0 peak of that syllable.

Despite the lack of consensus regarding what the prosodic anchor should be, the results of previous production studies on prosody-gesture alignment have generally suggested that gestural prominence, widely accepted to be the apex, is aligned with prosodic prominence, which is manifested primarily by a higher F0.

1.3. The present study

While the temporal relationship between prosodic and gestural prominences is quite well established in the literature, it is based only on studies of non-tone languages with lexical stress such as English (Jannedy & Mendoza-Denton, 2005; Loehr, 2012), Catalan (Esteve-Gibert, Borràs-Comes, et al., 2014; Esteve-Gibert, Pons, et al., 2014; Esteve-Gibert & Prieto, 2013), Dutch (Krahmer & Swerts, 2007), French (Roustan & Dohen, 2010), Brazilian Portuguese (Rochet-Capellan, Laboissière, & Galva´n, A., & Schwartz, J.-L. , 2008) and Swedish (Ambrazaitis & House, 2017; Ambrazaitis, Svensson Lundmark, & House, 2015). If it is the word bearing the nuclear stress/the nuclear stress syllable itself/the F0 peak of that syllable that attracts gestural prominence, one legitimate question would be whether and how prosody and gesture coordinate temporally in non-stress tone languages, in which nuclear stress is non-existent and F0 peak does not necessarily correspond to prominence. Thus, the present study investigates the prosody-gesture coordination in Hong Kong Cantonese, a non-stress language with a complex tone system (see Table 1 below for its tone inventory), which serves as a good testing ground.

Unlike the non-tone languages mentioned above and a number of Chinese tone languages such as Mandarin (Chen, Wang, & Xu, 2009; Xu, Chen, & Wang, 2012; Xu, Xu, & Sun, 2004) and some Northern Wu dialects (Wang, Zhang, Xu, & Ding, 2016), whether Hong Kong Cantonese makes use of F0 variations to signal prosodic prominence is still subject to debate. On the one hand, some studies found on-focus F0 rising (Gu & Lee, 2007; Man, 1999, 2002) and post-focus compression (Man, 2002), arguing that F0 was the primary acoustic correlate of focus prominence. More recent studies, on the other hand, suggested that it was durational lengthening, not on-focus or post-focus F0 changes, that marked the prosodic prominence on narrow focus in Cantonese (Wu & Xu, 2010; Fung & Mok, 2014). While research on the perception of prosodic prominence in Cantonese is scarce, there is evidence that Cantonese speakers interpret a longer duration as a perceptual cue of prominence (Leemann et al., 2016).

Regarding the uncertainty over the role of F0 in marking prosodic prominence in Cantonese and that no study has yet been done on the temporal relationship between prosodic and gestural prominences in the language, the following questions were raised:

(1) Is prosodic prominence marked by F0 variations in addition to durational lengthening in Cantonese?
(2) Is prosodic prominence aligned to gestural prominence in Cantonese?
(3) If there is alignment between prosodic and gestural prominences, what is the prosodic anchor of the alignment?
    (3.1) If F0 is indeed an acoustic correlate of prosodic prominence in Cantonese, does the presence of lexical tones and the absence of lexical stress affect the way the prosodic anchor, whatever it is, attracts gestural prominence?
    (3.2) If F0 is NOT an acoustic correlate of prosodic prominence, what else can the prosodic anchor be?

Assuming that language and gesture are parts of the same communicative system and that the temporal synchrony between the two modalities is universal, it is hypothesised that prosodic and gestural prominences are aligned with one another in Hong Kong Cantonese. If F0 is not an acoustic correlate of prosodic prominence, the prosodic anchor of alignment is predicted to be the syllable with emphatic stress, which is lengthened in duration and does not necessarily carry the F0 peak. In addition, the tone shape of the emphatically stressed syllable or word – whether it is rising, falling or level – is not expected to have an effect on the prosody-gesture alignment.

If F0 is indeed an acoustic correlate of prosodic prominence in addition to durational increase, the alignment pattern is predicted to be more complicated. Several alignment patterns are possible in addition to the F0 peak given the tone contours. It is quite difficult to predict which pattern is more likely at this stage because we do not have enough information. We would leave this prediction open.

To test the above hypotheses, an object-verification task was conducted to elicit corrective focus (as a form of prosodic prominence) and co-speech pointing gestures.

2. Method

The object-verification task was adopted and modified from that in Rusiewicz et al. (2013). It aimed to test how prosodic prominence on corrective focus was realised and, by manipulating properties of the focus, how the alignment of gestural prominence could be affected. Details of the experiment are as follows.

2.1. Participants

Four male and six female students from the Chinese University of Hong Kong (CUHK) aged 23–28 participated in the experiment. All were native speakers of Hong Kong Cantonese with normal or corrected-to-normal vision and no reported speech, hearing or motor impairment, and were rewarded with course credits for their participation.

Table 1
Summary of Cantonese lexical tones.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
Tone shape	high level [55]	high rising [25]	mid level [33]	low falling [21]	low rising [23]	low level [22]

Table 2
Monosyllabic (mono-σ) and disyllabic (di-σ) target words used in the experiment.

	Mono-σ		Di-σ	
Tone 1	刀 /tou$_1$/	“knife”	西瓜 /sɐi$_1$ kʷa$_1$/	“watermelon”
	貓 /mau$_1$/	“cat”	花樽 /fa$_1$ tsɵn$_1$/	“vase”
	遮 /tsɛ$_1$/	“umbrella”	青椒 /tsʰɛŋ$_1$ tsiu$_1$/	“green pepper”
	鐘 /tsʊŋ$_1$/	“clock”		
Tone 2	橙 /tsʰaŋ$_2$/	“orange”	鎖鏈 /sɔ$_2$ lin$_2$/	“chain”
	紙 /tsi$_2$/	“paper”	水手 /sɐi$_2$ sɐu$_2$/	“sailor”
	磅 /pɔŋ$_2$/	“scale”	相架 /sœŋ$_2$ ga$_2$/	“photo frame”
	井 /tsɛŋ$_2$/	“well”		
Tone 4	狼 /lɔŋ$_4$/	“wolf”	羊駝 /jœŋ$_4$ tʰɔ$_4$/	“llama”
	蛇 /sɛ$_4$/	“snake”	皮鞋 /pʰei$_4$ hai$_4$/	“leather shoes”
	床 /tsʰɔŋ$_4$/	“bed”	牛油 /ŋɐu$_4$ jɐu$_4$/	“butter”
	船 /syn$_4$/	“ship”		

2.2. Materials

PowerPoint slides were used to display coloured pictures of common animals and objects. The slides were embedded with Cantonese audio prompts pre-recorded by the first author for the elicitation of 21 Cantonese words in carrier frames. Details of the selection of target words and material design are as follows.

As shown in Table 2, 12 monosyllabic and nine disyllabic Cantonese words were used as target words. The following factors were considered when compiling the word lists:

(1) A word’s lexical tone,
(2) its onset and coda consonants,
(3) the coda of its classifier, and
(4) its picturability.

Firstly, three lexical tones of different shapes, namely Tone 1 (level), Tone 2 (rising) and Tone 4 (falling)[1], were included to investigate whether the tone shape of a prosodically prominent word/syllable affected gesture alignment (for each disyllabic target word, both syllables share the same tone). Secondly, words with oral stop onsets and associated classifiers with unreleased stop codas p˺, t˺ or k˺ were avoided. As will be illustrated further in the following paragraph, the target words preceded by classifiers were elicited in the carrier sentence /hɐi$_6$/m$_4$ hɐi$_6$, jɐu$_5$ CL ___ hɐi$_2$ CL ___ sœŋ$_6$ min$_6$/. “Yes/No, there is a ___ above the ___.”. The immediate precedence of an unreleased stop poses problem to the measurement of duration of stop-initial words, since it is difficult to pinpoint the onset of stop closure spectrographically. Therefore, items like /kʷɐi$_1$/ “turtle” and /kɐu$_2$/ “dog”, which begin with an oral stop and are classified by tsɛk˺$_3$, were not included in the inventory. Thirdly, words ending with unreleased stops were also excluded because they are shorter; hence, they are less favourable for tone measurement. Fourthly, since it was an object-naming task, the production of the target words was elicited (partially) by pictures of them. Therefore, the chosen words were mainly common objects and animals that could easily be represented visually.

[1] Since the factor concerned is tone *shape*, to simplify the experiment, Tones 1, 2 and 4 were chosen as delegates of the three possible contours in the Cantonese tone inventory. Although using words of Tones 6 (low level) and 5 (low rising) instead of Tones 1 and 2 could have addressed the same question, it would have been more difficult to compile the word list, since there are fewer words with Tones 5 and 6 than there are with Tones 1 and 2 in Cantonese (Kwan et al., 2003).

Fig. 1. The PowerPoint slide that accompanied the question prompt /hɐi6 mɐi6 jɐu5 pa2 tshœŋ1 hɐi2 kɔ3 tshaŋ2 sœŋ6 min6 a3/? “Is there a gun above the orange?”, used for eliciting corrective focus on the word tou1 “knife”.

Pictures of the target items, arranged one above the other on PowerPoint slides (see Fig. 1 for an example), and questions in the form /hɐi$_6$ mɐi$_6$ jɐu$_5$ CL ___ hɐi$_2$ CL ___ sœŋ$_6$ min$_6$ a$_3$/? “Is there a ___ above the ___?” were used as prompts eliciting production of the target words under different focus conditions. Each word was to be produced in each of the empty slots in the carrier sentence mentioned above. Depending on whether a question about the object being verified was correct, participants would respond with either the affirmative (neutral focus) or the negative (corrective focus) version of the carrier sentence. Moreover, for disyllabic items specifically, corrective focus was elicited on

(1) the first syllable,
(2) the second syllable and
(3) both of them.

To elicit corrective focus on the first syllable of a disyllabic target, an incorrect item, which shared only the second syllable with the target, was mentioned in the question prompt. For example, the prompt /hɐi$_6$ mɐi$_6$ jɐu$_5$ tɵi$_3$ pɔ$_1$ hai$_4$ hɐi$_2$ kɔ$_3$ sœŋ$_2$ ga$_2$ sœŋ$_6$ min$_6$ a$_3$/? “Is there a pair of sneakers above the photo frame?” was used to elicit the response /m$_4$ hɐi$_6$, jɐu$_5$ tɵi$_3$ pʰei$_4$ hai$_4$ hɐi$_2$ kɔ$_3$ sœŋ$_2$ ga$_2$ sœŋ$_6$ min$_6$/. “No, there is a pair of leather shoes above the photo frame.” with a focus on the first syllable of the target word /pʰei$_4$ hai$_4$/. Similarly, corrective focus on the second syllable and both syllables was elicited via an incorrect item sharing only the first syllable and

Table 3
Focus conditions elicited for each mono- and disyllabic target word.

Word	Verbal response [focus condition]				
Mono-σ	Yes, No,	there is a	[neutral] [on-focus]	above the	[neutral]. [post-focus].
Di-σ	Yes, No,	there is a	[neutral] [on-focus] (1st σ) [on-focus] (2nd σ) [on-focus] (both σ)	above the	[neutral]. [post-focus].

neither of the syllables with the target word, respectively. As a result, a monosyllabic target word was produced in four focus conditions and a disyllabic one in six conditions (see Table 3), elicited by a total of 60 picture-question pairs.

2.3. Procedures

The experiment was conducted in a quiet room with audio-visual equipment. Each participant was seated at a table, being approximately 2.5 meters away from the screen on which visual stimuli were projected. Two camcorders, recording at 25 frames per second (fps), were set in front of and to the left of the participant respectively. Each camcorder was connected to a digital sound recorder used as an external microphone recording at the sampling rate of 44,100 Hertz (Hz), 16 bit.

Consistent with similar studies of elicited co-speech pointing (e.g., Esteve-Gibert & Prieto, 2013; Rusiewicz et al., 2013), the experiment consisted of three parts, including a familiarisation session, a training session, and a testing session. In the familiarisation session, participants learned the mappings between pictures of the target/filler items and their corresponding labels. The pictures were first shown once, one at a time in a random order and accompanied by their aurally-presented labels. Following this, all the pictures were presented again, this time without the labels and in a different order, and participants were asked to name them. The practice continued until every picture could be named correctly in two successive attempts.

In the training session, participants practised responding to question prompts verbally and, in certain cases, gesturally as well. As mentioned in Section 2.2, all target words were to be produced in the carrier phrase "there is a ___ above the ___.", either beginning with a "Yes" as an affirmative statement, or a "No" as a negative/corrective statement. The choice between the two verbal responses depended on whether the question prompt "Is there a ___ above the ___?" matched the visual stimuli on the screen. If it did, the affirmative statement was to be elicited. On the other hand, if there was a mismatch, participants were trained to provide a negative verbal response, as well as to point to the correct object on the screen with their dominant hand, imaging that the person who asked the question could not see the object clearly. Apart from the fillers (which, as mentioned, always elicited positive answers about the lower objects), there were a total of six target scenarios regarding whether

(1) the word was monosyllabic or disyllabic,
(2) the response was affirmative (neutral focus) or negative (corrective focus), and
(3) for disyllabic words, the corrective focus was on the first or second syllable or both (see Table 4).

Note that corrective focus was only elicited for the sentence-medial targets, but not the sentence-final ones.

To drill participants to respond to all the eight target and filler scenarios, 16 picture-question pairs were used for training, two for each scenario. The picture-question pairs were presented in a random order, and rotated until each question could be responded to correctly in three successive attempts. When not pointing, participants were reminded to rest their forearms on the table in order to have a uniform reference point from which to measure the beginning and end of pointing for all speakers (details about gesture analysis will be provided in Section 2.4.2). A separate set of practice items was used instead of the targets and fillers, since the pilot tests reflected that the training was too long when using the latter.

In the testing session, a total of 75 picture-question pairs were presented to each participant, among which 60 were targets (that is, one repetition for each target word under each focus condition) and 15 were fillers. Two breaks were given during the session. In the event of incorrect responses (such as giving a neutral response when it should have been correc-

Table 4
Summary of all the response scenarios (target and filler) in the experiment (English letters represent target syllables).

	Word	Focus	Visual stimulus	Question prompt ("Is there a/an…")	Elicited response	
					Verbal	Gestural
Target responses						
1.	Mono -σ	Neutral	[Y above Z]	"… Y above the Z?"	√	×
2.		Corrective		"… X above the Z?"	√	√
3.	Di-σ	Neutral	[AB above FG]	"… AB above the FG?"	√	×
4.		Corrective (1st σ)		"… CB above the FG?"	√	√
5.		Corrective (2nd σ)		"… AC above the FG?"	√	√
6.		Corrective (both σ)		"… DE above the FG?"	√	√
Filler responses						
1.	Mono -σ	Neutral	[P above Q]	"… Q below the P?"	√	×
2.	Di-σ	Neutral	[ST above UV]	"… UV below the ST?"	√	×

tive, omission of pointing, and so on) or hesitation (which happened more often with the gestures), the questions concerned were re-asked at the end of the session.

2.4. Data analysis

2.4.1. Acoustic data

After extracting the audio files from the videos (specifically, only the videos taken from the front angle), segmentation and annotation was performed using Praat (Boersma & Weenink, 2013). To corroborate the previous finding that duration, not F0, was the primary acoustic correlate of focus prominence in Cantonese, all target words were measured for duration (in milliseconds), mean F0 and F0 range. For disyllabic target words, the three measures were taken separately for the first and the second syllables.

The reason for including both mean F0 and F0 range was that, while the former can capture changes in the height of level Tone 1, the addition of the latter can reflect the presence/absence of changes in height and/or range of rising Tone 2 and falling Tone 4, as presented graphically in Fig. 2. A change in mean F0 is represented by a parallel shift in the contour (i.e., Case 2); a change in F0 range is respresented by a change in slope of the contour (i.e., Case 3); a change in both mean F0 and F0 range is represented by a contour shift with a change in slope (i.e., Case 4).

To compute the two F0 measures, F0 values were measured at 10 equally distant points in the sonorant part of the word. To allow across-gender comparison, the values were then normalised to semitones 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 55 Hz for male speakers and 100 Hz for female speakers (Zou, Wang, & He, 2012).

Subsequently, after aggregating and averaging the data for each speaker, repeated measure (RM) ANOVAs were performed, two-way (Tone × Focus) for monosyllabic target words and three-way (Tone × Focus × Syllable) for disyllabic ones. Sentence-medial and -final words were analysed separately. Independent variables for the analyses of monosyllabic tokens were Tone (three levels: Tone 1, Tone 2 and Tone 4) and Focus (two levels: neutral and on-focus (for sentence-medial tokens)/post-focus (for sentence-final ones). Note again that prominence of corrective focus was elicited only on the sentence-medial tokens, but not on their sentence-final counterparts.

With regard to the disyllabic tokens in the sentence-medial position, independent variables in the RM ANOVAs included Tone (three levels), Focus (four levels: neutral focus, single-syllable focus, off-focus and whole-word focus (see Table 5 for an illustration using an English example), and Syllable (two levels: first and second). With regard to their

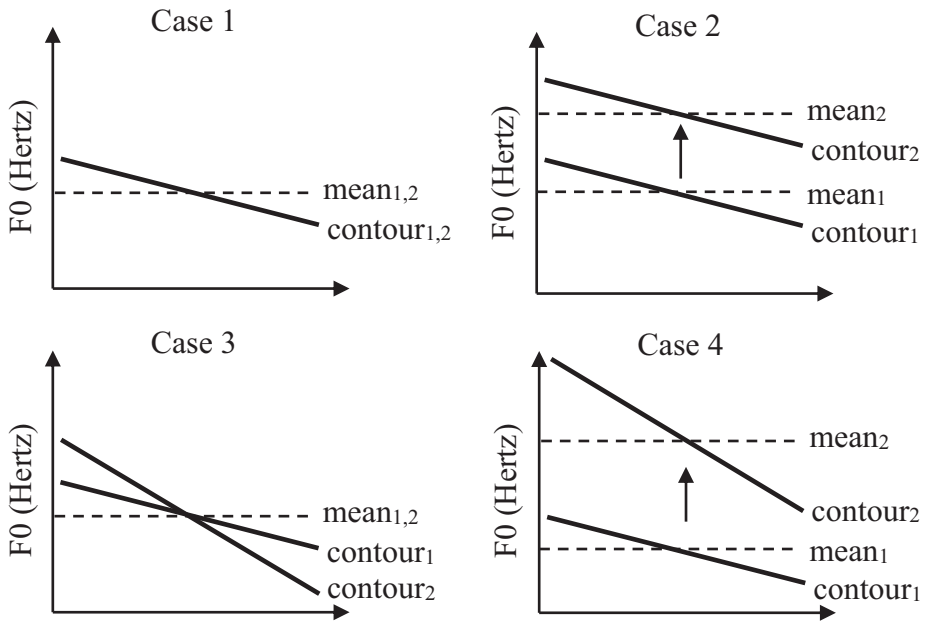

Fig. 2. Illustrations of how the measurement of mean F0 and range F0 could reflect different changes in the contour tones (number subscripts: 1 = neutral focus condition, 2 = corrective focus condition).

Table 5
Illustration of the four focus conditions manipulated for each syllable in a disyllabic target item (focus indicated by capitalisation).

Scenario	Example	Focus condition on 1st σ	Focus condition on 2nd σ
No correction	raincoat	neutral focus	
Correction of 1st σ	RAINcoat (contrasting with trenchcoat)	single-syllable focus	off-focus
Correction of 2nd σ	rainCOAT (contrasting with rainboots)	off-focus	single-syllable focus
Correction of entire word	RAINCOAT (contrasting with umbrella)	whole-word focus	

sentence-final counterparts, the variables were the same except that Focus had only two levels, namely neutral and post-focus. In the event of a violation of the sphericity assumption, the degree of freedom was corrected using Greenhouse-Geisser estimates of sphericity. For the post-hoc pairwise comparisons, the Bonferroni correction was adopted. Note that, since two female speakers (Speakers F1 and F3) were very creaky in their Tones 2 and 4, leading to much missing data, both were excluded from mean F0 and F0 range analyses.

2.4.2. Gestural data

Videos taken from the front and the side of each speaker were synchronised using Adobe Premiere CC Pro before being imported to ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) for gesture annotation. To allow for the examination of temporal coordination between prosodic and gestural prominences, Praat tiers bearing annotations of the target words were also added to the annotation interface (see Fig. 3).

Each instance of pointing was measured to determine its total gesture duration, the onset being the last frame before the pointing arm was lifted from the table, and the offset being the first frame in which the arm touched the table again. Its apex was identified as the moment when the gesturing hand was maximally displaced from the rest position; in other words, when both the pointing arm and the finger reached maximal extension.

In addition, four other measures—three intervals (in milliseconds) and one ratio—were calculated for assessment of the timing of pointing relative to the co-occurring target word in focus. The first was the interval between gesture onset and word onset (henceforth the 'GO-WO interval'), which was adopted from Rusiewicz et al. (2013). A positive number means that pointing started after the focused word began, vice versa. Similarly, the distance between gesture and word offsets (the 'GF-WF interval') was also measured; a positive number indicated that the pointing ended after the word, and vice versa. The third measure was the duration of gesture launch; that is, the distance between the onset and the apex of a gesture. The fourth measure, or the alignment-with-word ratio ('AW ratio'), was calculated by dividing the interval between the apex and the word onset by the length of the focused word. This indicates the closeness of the alignment between prosodic and gestural prominence. If the value is between 0 and 1, it means that the apex occurs within the span of the focused word. If it is smaller than 0 (or larger than 1), the apex occurs before (or after) it. Figs. 4 and 5 below are graphical presentations of the calculations of these measures.

The reasons for including these five measures were as follows. The AW ratio was to assess whether or how closely the apex of a gesture coincided with the focused word, and the interval measures reveal how exactly the alignment (if any)

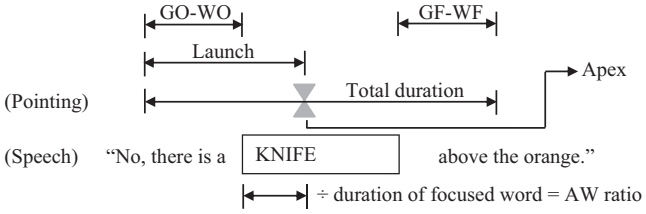

Fig. 4. Illustration of the gestural measures.

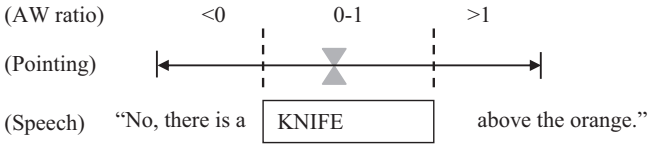

Fig. 5. Interpretation of the AW ratio.

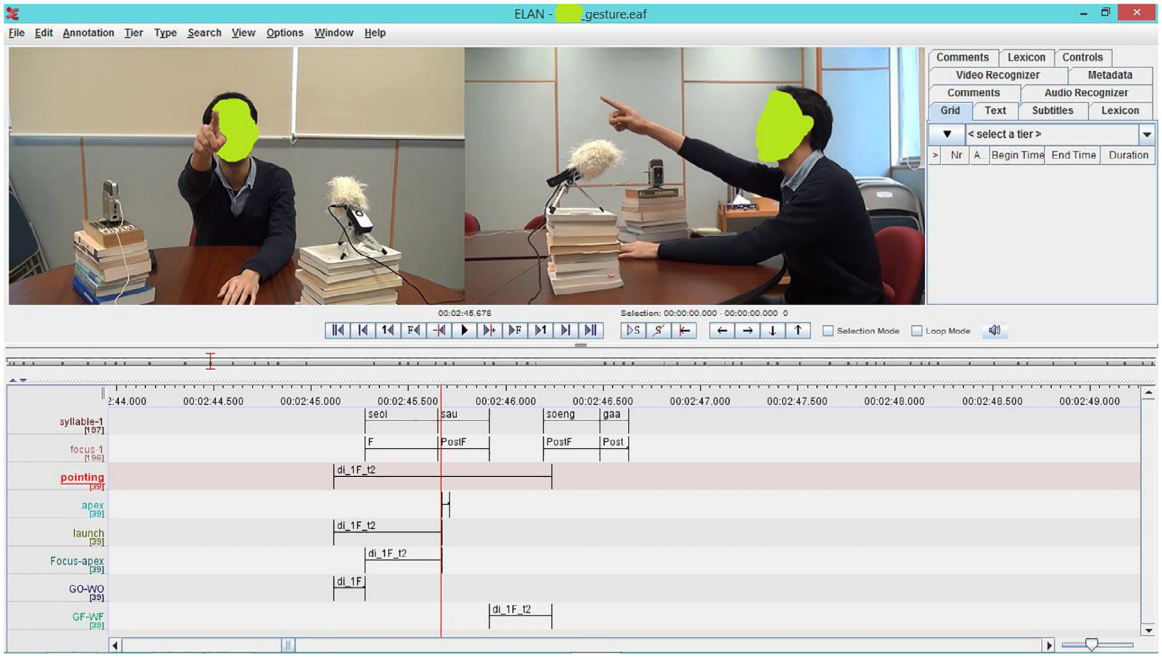

Fig. 3. Example of the Elan interface of gesture annotation.

was achieved. There are a few logical possibilities that a gesture may be adjusted according to manipulations of prosodic prominence on focus, one of which is that the entire gesture is initiated earlier/later for an earlier/later prosodic prominence. In this case, the GO-WO and GF-WF intervals change, while the total gesture duration and duration of the launch remain relatively constant. Another possibility is to change the length of the launch in order to have an earlier or later apex, which may or may not be accompanied by changes in the total gesture duration, GO-WO and GF-WF intervals.

Data for all gestural measures were analysed statistically using RM ANOVAs, one-way (Tone) for gestures accompanying foci on monosyllabic words (henceforth 'monosyllabic pointing gestures') and two-way (Tone × Focus) for those accompanying foci on disyllabic words ('disyllabic pointing gestures'). Other statistical corrections/post-hoc procedures followed those in the acoustic analyses. A total of 120 monosyllabic pointing gestures (four words × one focus condition × three tones × 10 speakers) and 270 disyllabic pointing gestures (three words × three focus conditions × three tones × 10 speakers) were examined.

3. Results

The results of our experiment will be presented in two parts. The first, results of the acoustic analyses, which aimed to determine the acoustic correlates of prosodic prominence on corrective focus, will be presented in Section 3.1. The results of the gesture analyses, which aimed to address the main research questions regarding prosody-gesture alignment in Cantonese, will follow in Section 3.2.

3.1. Acoustic analyses

Consistent with previous studies by Wu and Xu (2007) and Fung and Mok (2014), our results showed that focus in Cantonese was marked primarily by durational lengthening (see Section 3.1.1) and was not accompanied by any post-focus F0 lowering or compression (see Section 3.1.2).

3.1.1. Neutral versus corrective focus

Not surprisingly, for the monosyllabic target words, Tone had a significant effect on the mean F0, $F(2,14) = 123.15$; $p < 0.001$. Post hoc Bonferroni-adjusted pairwise comparisons showed that words with Tone 1 had a significantly higher mean F0 than did those of Tones 2 and 4, which were both $p < 0.001$. The effect of Focus, on the other hand, was non-significant.

With regard to the F0 range, there was significant Tone-Focus interaction, $F(2,14) = 5.24$; $p = 0.02$, in addition to significant main effects of both Tone, $F(2,14) = 19.32$; $p < 0.001$, and Focus, $F(1,7) = 18.52$; $p = 0.004$. While the magnitude of on-focus F0 range expansion was greater for Tone 2 than it was for Tones 1 and 4, it was found to be non-significant in post-hoc Bonferroni-adjusted pairwise comparisons. In other words, no significant on-focus change in the F0 range was found in any of the tones.

With regard to word duration, target words were significantly longer under corrective focus than they were under neutral focus, $F(2,18) = 17.08$, $p < 0.001$. The effect of Tone was also significant, $F(1,9) = 44.10$, $p < 0.001$. Post-hoc Bonferroni-adjusted pairwise comparisons showed that words in Tone 1 were significantly shorter than were those in Tone 2 ($p = 0.001$) and Tone 4 ($p = 0.03$). Tone-Focus interaction was non-significant.

With regard to the disyllabic words, there was significant interaction between Tone and Syllable on mean F0, $F(2,14) = 54.09$, $p < 0.001$, in addition to a significant main effect of Tone, $F(2,14) = 137.8$, $p < 0.001$. Post-hoc Bonferroni-adjusted pairwise comparisons showed significant between-syllable difference in mean F0 only for words in Tones 2 and 4, both $p < 0.001$, but not for Tone 1. The effect of Focus was non-significant.

With regard to the F0 range, there were significant interactions between Tone and Focus, $F(6,42) = 3.00$, $p = 0.17$, and between Tone and Syllable, $F(2,14) = 12.19$, $p < 0.001$, along with significant main effects of Tone, $F(2,14) = 8.18$, $p = 0.004$, Focus, $F(3,21) = 9.15$, $p < 0.001$, and Syllable, $F(1,7) = 28.80$, $p = 0.001$. Post-hoc Bonferroni-adjusted pairwise comparisons revealed that only Tone 2 syllables under single-syllable focus underwent significant F0 range expansion ($p = 0.048$), and that only target words in the contour tones exhibited significant between-syllable difference in the F0 range (both $p < 0.001$).

With regard to duration, Focus was found to have a significant main effect on duration, $F(1.52,13.71) = 26.31$, $p < 0.001$. Post-hoc pairwise comparisons showed that syllables were significantly shorter under neutral focus than they were for single-syllable focus ($p = 0.002$), whole-word focus ($p = 0.003$) and off-focus conditions ($p = 0.001$). They were also significantly shorter under off-focus and whole-word focus conditions than they were under the single-syllable focus condition ($p = 0.012$ and 0.039, respectively). In other words, from the greatest to the smallest, the order of syllable durations under different focus conditions was as follows: single-syllable focus > whole-word focus ≈ off-focus > neutral focus.

There were also significant main effects of Tone, $F(2,18) = 18.45$, $p < 0.001$, and Syllable, $F(1,9) = 42.97$, $p < 0.001$, as well as a significant interaction between the two variables, $F(2,18) = 17.66$, $p < 0.001$.

3.1.2. Neutral versus post-focus

For the monosyllabic words, Tone again had a significant main effect on mean F0, $F(2,14) = 82.44$; $p < 0.001$, F0 range, $F(2,14) = 24.59$; $p < 0.001$ and duration, $F(2,18) = 14.61$; $p < 0.001$. Post-hoc Bonferroni-adjusted pairwise comparisons showed that (1) the mean F0 was significantly higher in Tone 1 targets than it was in their dynamic tone counterparts (both $p < 0.001$), and was also significantly higher in Tone 2 targets than it was in Tone 4 ones ($p = 0.045$); (2) the F0 range was significantly smaller in Tone 1 targets than in those of the dynamic tones (again $p < 0.001$) and smaller in Tone 4 than in Tone 2 ones ($p = 0.033$); (3) Tone 1 targets were significantly longer than were both Tone 2 and Tone 4 ones ($p = 0.003$ and 0.015, respectively). The effects of Focus and Focus-Tone interaction, however, were found to be non-significant in all three measures.

Similarly, for the disyllabic words, no significant effect of Focus or significant interaction involving Focus was found for the three acoustic measures. The only exception was the significant Tone-Focus interaction in the F0 range, $F(2,14) = 4.61$,

$p = 0.03$. However, post-hoc Bonferroni-adjusted pairwise comparisons showed no significant difference in the F0 range between neutral and post-focus conditions for any of the tones.

3.2. Gesture analyses

3.2.1. Gesture onset -to- word onset interval (GO-WO interval) and gesture offset -to- word offset interval (GF-WF interval)

All of the 120 monosyllabic pointing gestures were found to begin before the onsets and end after the offsets of their co-occurring words in focus, with mean GO-WO and GF-WF intervals of −507 ms (SD = 204 ms) and 940 ms (SD = 505 ms), respectively. The effect of Tone was non-significant in both measures, although there was a trend towards significance for the GO-WO interval, $F(2,18) = 3.35$, $p = 0.058$.

As with their monosyllabic counterparts, all disyllabic pointing gestures were initiated before their associated words in focus, and nearly all ended afterwards (except for four instances in which the gestures ended 2 ms to 128 ms prior to word offsets), the mean GO-WO and GF-WF intervals being −500 ms (SD = 205 ms) and 957 ms (SD = 623 ms), respectively. The results showed only a significant main effect of Tone on GO-WO, $F(2,18) = 4.20$, $p = 0.032$. Post-hoc pairwise comparisons revealed that gestures associated with the foci of Tone 2, the mean GO-WO interval of which was −490 ms (SD = 61 ms), were initiated significantly later than were those associated with the foci of Tone 4 ($p = 0.034$), which had a mean GO-WO interval of −532 ms (SD = 69 ms). Other effects on the two measures were all non-significant. The averages of the two measures of the disyllabic words are shown in Fig. 6.

3.2.2. Alignment-with-word ratio (AW ratio)

As explained in Section 2.4, the AW ratio is calculated by dividing the distance between the apex and the onset of the focused word by the duration of that word. A number between 0 and 1 indicates that an apex occurred within the span of the word in focus, whereas a number smaller than 0 (larger than 1) means that an apex preceded (followed) it.

Among the 120 monosyllabic pointing gestures, 92 (76.67%) were produced with their apices coinciding with their associated words in focus, with a mean AW ratio of .440 (SD = 0.246). Of the remainder, 12 (10%) had their apices preceding the foci, with a mean AW ratio of −0.559 (SD = .489), and 16 (13.33%) had them afterwards, the mean AW ratio being 1.512 (SD = 0.344). The RM ANOVA revealed no significant effect of Tone on the ratio, $F(2,18) = 1.42$, $p = 0.268$.

With regard to the 270 disyllabic pointing gestures, 238 (88.15%) were produced with their apices within the spans of their associated words in focus, with a mean AW ratio of 0.349 (SD = 0.236). Of the remaining 32, half (5.9%) had their apices occurring before the focused words, with a mean AW ratio of −0.277 (SD = 0.235), and half afterwards, the mean AW ratio being 1.406 (SD = 0.305). There was no significant effect of Focus, $F(2,18) = 0.22$, $p = 0.804$ or Tone, $F(1.07,9.64) = 0.89$, $p = 0.377$. Neither was their interaction significant, $F(1.58,14.25) = 0.16$, $p = 0.805$.

Table 6 shows the by-speaker distribution of the 60 pointing gestures in which the apices were produced before or after their associated words in focus. While most of the speakers produced their pointing gestures with apices reached either before or during the course of the focused words, M4 stood out by having 23 of his 39 gestures produced with late apices.

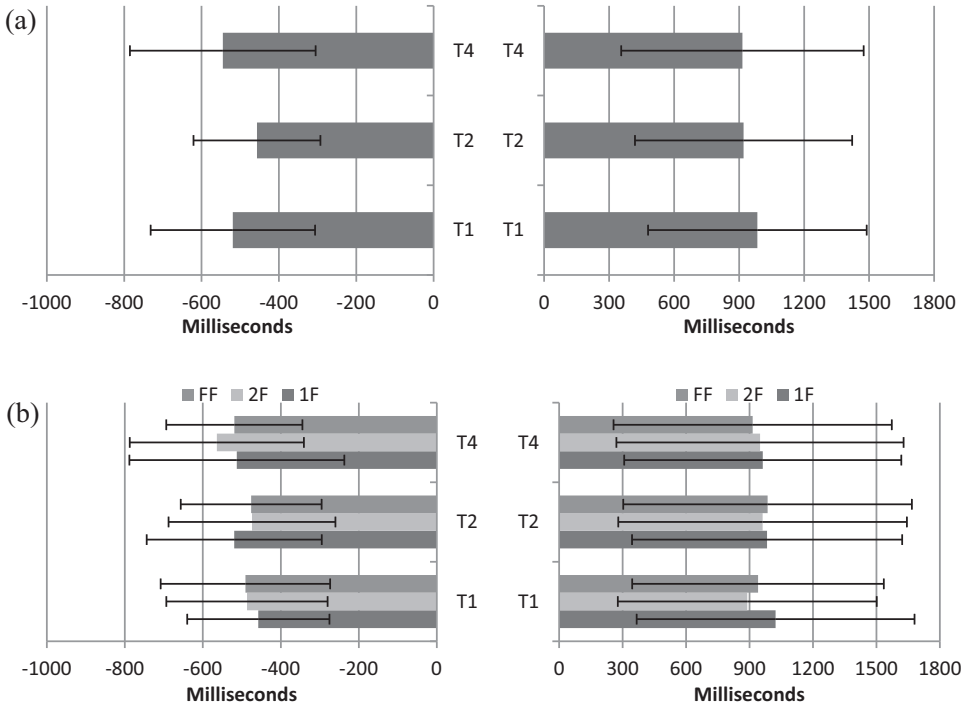

Fig. 6. Average durations of GO-WO (left panels) and GF-WF (right panels) intervals (in ms) of monosyllabic gestures (panel (a)) and disyllabic gestures (panel (b)), error bars showing ± 1 SD (abbreviations: T = Tone; 1F, 2F, and FF = first-syllable, second-syllable, and whole-word foci).

Table 6
Distribution of gestures with apices occurring before/after associated words in focus.

Speaker	Before word	Before: mono-σ	Before: di-σ	After word	After: mono-σ	After: di-σ
F1	1	1	0	1	1	0
F2	19	6	13	0	0	0
F3	6	4	2	0	0	0
F4	0	0	0	0	0	0
F5	0	0	0	0	0	0
F6	0	0	0	7	6	1
M1	2	1	1	0	0	0
M2	0	0	0	0	0	0
M3	0	0	0	1	1	0
M4	0	0	0	23	8	15

Table 7
By-speaker analysis of the alignment pattern of disyllabic pointing gestures with apices occurring within spans of associated words in focus.

Speaker	Focus on 1st σ		Focus on 2nd σ		Focus on both σ		Total
	1st σ	2nd σ	1st σ	2nd σ	1st σ	2nd σ	
F1	9	0	9	0	9	0	27
F2	5	0	3	2	3	1	14
F3	8	1	9	0	6	1	25
F4	9	0	9	0	9	0	27
F5	9	0	9	0	9	0	27
F6	1	7	0	9	2	7	26
M1	9	0	8	0	9	0	26
M2	9	0	8	1	9	0	27
M3	6	3	8	1	7	2	27
M4	2	2	2	3	0	3	12

Individual differences were also observed with regard to which syllables (the first versus the second) of the disyllabic words with which the speakers aligned their gesture apices. Table 7 shows the alignment pattern of each speaker under each of the three focus conditions, with the figure in each cell indicating the number of gestures in which the apex occurred within the referred syllable. Most speakers aligned their apices with the first syllables of the disyllabic words in focus, regardless of the actual positions of the foci, except for three of them: (1) Speaker F2, the "early pointer" who produced half of her disyllabic pointing gestures with apices occurring ahead of the words, (2) Speaker M4, the "late pointer" whose gesture apices were mainly produced after the associated foci, and (3) Speaker F6, who mainly aligned hers with the second syllables instead of the first.

To summarise, for both mono- and disyllabic pointing gestures, apices were mainly produced within the spans of their associated foci. Neither Tone (for both mono- and disyllabic pointing gestures) nor Focus (for disyllabic ones only) was found to have a significant effect on the alignment pattern. Despite changes in focus conditions, most speakers exhibited a consistent alignment of apices with either the first or the second syllables of the disyllabic words carrying focus, with the first preferred more frequently to the second.

3.2.3. Summary of the gesture analyses

Firstly, pointing gestures co-occurring with foci of different tone shapes and focus conditions were found to be largely homogeneous. For monosyllabic pointing gestures, the effect of Tone was found to be insignificant in all gestural measures (including *total gesture duration* and *launch duration*, statistical results of which were all non-significant and hence not reported separately). Similarly, for disyllabic ones, no significant effect of Tone or Focus was found on any of the measures except for a significant effect of Tone on the length of the GO-WO interval.

Secondly, gesture apices were produced within the spans of the associated words in focus for most speakers. In addition, for apices of disyllabic pointing gestures specifically, alignment was made consistently with the same syllables (either the first or the second) of the foci by each speaker despite changes in the focus position, with the first syllables being preferred by most speakers.

4. Discussion

The goal of the present study is two-fold: (1) to revisit the role of F0 in marking prosodic prominence in Hong Kong Cantonese, and (2) to find out whether or not prosodic and gestural prominences are aligned to one another in the language, and how. As shown in the last section, we confirmed that prosodic prominence on corrective focus in Cantonese is marked not by F0 variations, but by durational lengthening.

We also confirmed our hypotheses that there was a temporal alignment between prosodic and gestural prominences in Hong Kong Cantonese, and that the alignment was not affected by the tone shape of the prosodically prominent unit. However, contrary to our prediction, the manipulation of the focus position (the focus on the first versus the second syllables of the disyllabic target words) had no effect on the timing of the gesture apices.

With regard to the temporal coordination between prosodic prominence in corrective focus and gestural prominence in manual pointing in Hong Kong Cantonese, our results revealed some interesting patterns that have not been reported in previous studies. While there were certainly some close temporal relationships between prosody and gesture, as evidenced by the fact that nine of the 10 participants had gesture apices that

either preceded or coincided with the focused words as predicted by the phonological synchrony rule, the way in which they coordinated was somewhat unexpected. In contrast to stress languages in which a high F0 corresponds to prosodic prominence and attracts gestural prominence, F0 does not seem to play a role in prosody-gesture alignment in Hong Kong Cantonese. Instead, the word carrying the prosodic stress manifested by durational increase is more likely to be the prosodic anchor, as supported by current findings.

It is not surprising that tone had no effect on the timing of the apex or on any other temporal parameters, since prosodic prominence is not manifested by F0 change in Hong Kong Cantonese, as confirmed in the present study. On the other hand, it was surprising to find that focus position (whether the focus was on the first, second or both syllables of a disyllabic target word) had no effect on any of the gestural measures despite the fact that focused syllables were produced more prominently with a significant durational increase. Such homogeneity of gestural alignment is uncommon in the literature. Even in Rusiewicz et al. (2013), in which no alignment between the apex and the focused syllable was found, other temporal parameters such as gesture duration, gesture launch and the GO-WO interval did vary significantly across focus positions. One might attribute the absence of effect of focus position to the design of the pointing task. The fact that the contrasted objects (i.e., the wrong object mentioned in the audio prompt and the correct object shown on the screen) never appeared on the same PowerPoint slide or contrasted in the same sentence might have encouraged the speakers to produce the verbal corrections and the gestures with a different pragmatic purpose than the one intended. However, since changes in focus position were found to affect gestural parameters in some previous studies adopting the same experiment protocol (e.g., Esteve-Gilbert & Prieto, 2013), the insensitivity of gesture timing to changes in focus position exhibited by our Cantonese speakers was unlikely to be the result of our experimental design.

Even more interestingly, each of the eight participants whose gesture apices coincided with the target words aligned them with the same syllable of the disyllabic words consistently. This suggested that the focused syllable was not the speech unit that attracted gestural prominence. Instead, the focused word seemed more likely to be the prosodic anchor of prosody-gesture alignment in Hong Kong Cantonese. In this case, the homogeneity of the alignment pattern across the focus conditions could be explained: Since the changes in the focus position took place within the word, there was no change in the position of the most prominent word itself in the sentence; hence, no change in the interval measures such as the GO-WO/GF-WF intervals, total gesture duration, and so on was necessary.

However, this still leaves one question unanswered: Why did most of these eight speakers — seven of them, to be precise — choose to align the apices with the first syllable rather than with the second? If the focused word was the anchor, there should have been a more random distribution of gesture apices between the two syllables. The fact that the majority of the speakers exhibited a preference for the first syllable seems to suggest additional factors determining the alignment pattern.

One possibility is that the psycholinguistic prominence of the word-initial syllables might have been a contributing factor. From studies of lexical retrieval errors and the effect of syllable cues on resolving them (Browman, 1978; Hofferberth-Sauer & Abrams, 2014), evidence has shown that the word-initial syllable is crucial in lexical retrieval and word recognition. Lexical retrieval has also been found to be facilitated by co-speech lexical gestures, such as iconic and metaphorical gestures that "have shapes or dynamics related to the content of the accompanying speech" (Krauss & Hadar, 1999, p. 104). Although studies have not yet examined whether pointing facilitates word search in the same way that its lexical counterparts do, there might nonetheless be an interaction between the courses of gesture planning and lexical retrieval, particularly when the pointing gesture shares the same referent with — and thus performs the same deictic function as — the corresponding word label (the focused word in the case of this study). Such interaction might have been attributable to the alignment of the psycholinguistic prominence of the first syllables of the disyllabic target words with the apices of the pointing gestures.

Another interesting finding was the presence of the consistent 'late pointer', Speaker M4. In previous experimental studies, the average pattern of alignment was usually reported. There has barely been any mention of individual differences in gesture alignment patterns. To the best of our knowledge, there have not been any reports of any cases of consistent apex lags by the same speaker, as was the case with Speaker M4. Here it should be emphasised that there were no hesitations in his verbal or gestural responses in those instances of apex lag, and that his "lagged" gestures actually appeared to be as natural as did his "in-sync" ones and those of other participants. It could have been that he was attempting to align the gestural *strokes* (the defining and the only obligatory phase of a gesture) with the focused word rather than with the apices. Because of the partial overlapping between the strokes and the focused words, the gestures did not appear asynchronic at all. Although further tests are required to confirm this, it is possible that speakers of the same language could have different gesture alignment strategies.

Despite many questions remaining open with regard to the temporal relationship between prosody and gesture in Hong Kong Cantonese, this study contributes to the study of prosody-gesture alignment by introducing a new perspective from a language of a different prosodic typology. Thus far, many empirical studies on gesture alignment seem to have acknowledged pitch prominence, whether carried by a word, a syllable or simply as the F0 peak, as the anchor in speech that attracts gestural prominence, although only certain types of languages (intonation and pitch-accent languages with lexical stress) have been investigated. Nonetheless, just as languages can differ greatly in their prosodic features, it should not be surprising to see variability in the temporal properties of their accompanying gestures as well. The current finding that co-speech pointing gestures do seem to unfold with a different alignment pattern in Hong Kong Cantonese than they do in previously studied languages is a good demonstration of such cross-linguistic variability. Our hypothesis is that, while gesture alignment generally follows the rule of prosodic-gestural prominence synchrony, that unit of prosodic prominence is variable depending on the prosodic characteristics

of individual languages. This would explain the difference between previously studied languages and Hong Kong Cantonese in terms of their gestural alignment patterns. The apex is generally aligned with the F0 peak (or the stressed word/syllable bearing it) in stress languages because that is the acoustic manifestation of prominence in those languages. Similarly, the apex co-occurs with the first syllable of the focused word regardless of tone in Hong Kong Cantonese because it is the most salient unit in an utterance. In brief, it is hypothesised that patterns of temporal coordination between prosody and gesture can be both universal and language-specific: universal in the sense that gestural prominence occurs close to prosodic prominence in general, and specific in the sense that the actual unit of prosodic prominence depends on how it is realised in the language concerned. Our discovery of a different gesture alignment pattern in Hong Kong Cantonese highlights the need for more empirical studies of languages with different prosodic typologies before conclusions regarding the nature of prosody-gesture interaction can be drawn confidently.

Nonetheless, there is one limitation in the present study, which is that the location of focus was only manipulated within target words in the sentence-medial position. To verify whether or not the homogenous temporal patterns of the pointing gestures found in this study were indeed due to the prosodic anchor being the focused word, which was fixed in one position in the carrier sentence in the experiment, further studies varying the position of the focused word are necessary.

Another possible direction for further studies could be to examine different kinds of (manual) gestures. The hand beat, which is the up-and-down or back-and-forth movement of the gesturing hand (McNeill, 2005), could be a good place to start. Unlike the iconic and metaphorical gestures, the beat gesture does not bear the image of a lexical affiliate, nor does it have a referent like the pointing gesture does. In other words, while the timing of other gestures could be bound by factors other than speech prosody, such as semanticity, the beat is more closely related to speech prosody itself. Future studies could be directed towards both hand beats produced in an experimental setting co-elicited with, for example, contrastive foci as in Roustan and Dohen (2010), and those accompanying spontaneous speech (Loehr, 2012). It would be interesting to compare the alignment patterns of (1) the beat and the pointing gestures to see whether gesture type is important in the prosody-gesture synchrony, and (2) the elicited and the spontaneous beats to explore whether the latter show looser alignment with prosodically prominent words, and whether they necessarily co-occur with prosodically prominent words in the first place.

Future studies may also probe into the nature of prosody-gesture alignment in Hong Kong Cantonese or other non-stress tone languages from the perception perspective. For example, asynchrony can be manipulated between the gesture apex and different possible prosodic anchors, such as a focused word versus a focused syllable, to determine whether speakers of these languages are aware of asynchrony between the two modalities in general and, if so, whether they are more sensitive to asynchrony of the apex with certain prosodically prominent units than with others, which may provide additional evidence to determine the prosodic anchor.

In addition, to verify our hypothesis that cross-linguistic differences in prosodic characteristics predict language-specific gesture alignment patterns, perception studies may also be conducted to compare perceptual sensitivity to speech-gesture misalignment of speakers of different languages (for example, stress versus non-stress languages). Auditory stimuli could be manipulated to test whether speakers show a bias to prominence cued by certain acoustic parameters depending on the prosodic features of their first language. Conceivably, stress-language speakers might be more sensitive to gestural lag from a pitch-accented syllable, while non-stress (tone) language speakers might rely on durational cues in the anchoring word to tell the misalignment. No matter whether there is a difference or not, the result would surely add to our current knowledge of prosody-gesture interaction.

References

Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication, 95*, 100–113. https://doi.org/10.1016/j.specom.2017.08.008.

Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2015). Head Movements, Eyebrows, and Phonological Prosodic Prominence Levels in Stockholm Swedish News Broadcasts. *Paper presented at FAAVSP – The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processings, Vienna, Austria* (pp. 42–42).

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer. Retrieved from http://www.praat.org.

Browman, C. P. (1978). Tip of the tongue and slip of the ear. Implications for language processing. *UCLA Working Papers in Phonetics, 42*.

Chen, S., Wang, B., & Xu, Y. (2009). Closely related languages, different ways of realizing focus. In: *Proceedings of Interpseech 2009* (pp. 1007–1010). Brighton, UK. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_1007.pdf.

De Ruiter, J. P. (1998). *Gesture and speech production* Unpublished doctoral dissertation. Nijmegen: Radboud University.

Esteve-Gibert, N., Borràs-Comes, J., Swerts, M., & Prieto, P. (2014). Head gesture timing is constrained by prosodic structure. In: *Proceedings of Speech Prosody 2014* (pp. 356–360). Dublin, Ireland.

Esteve-Gibert, N., Pons, F., Bosch, L., & Prieto, P. (2014). Are gesture and prosodic prominences always coordinated? Evidence from perception and production. In: *Proceedings of Speech Prosody 2014* (pp. 222–226). Dublin, Ireland.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech Language and Hearing Research, 56*, 850–865. https://doi.org/10.1044/1092-4388(2012/12-0049).

Fung, H., & Mok, P. (2014). Realization of narrow focus in Hong Kong english declaratives—a pilot study. In: *Proceedings of Speech Prosody 2014* (pp. 964–968). Dublin, Ireland.

Goldin-Meadow, S. (1998). The development of gesture and speech as an integrated system. *New Directions for Child Development, 79*, 29–42.

Gu, W., & Lee, T. (2007). Effects of tonal context and focus on Cantonese F0. In: *Proceedings of 16th International Congress of Phonetic Sciences* (pp. 1033–1036). Saarbrücken, Germany. Retrieved from http://icphs2007.de/conference/Papers/1689/1689.pdf.

Hofferberth-Sauer, N. J., & Abrams, L. (2014). Resolving Tip-of-the-tongue states with syllable cues. In V. Torrens & L. Escobar (Eds.), *The Processing of Lexicon and Morphosyntax* (pp. 43–68). Newcastle: Cambridge Scholars Publishing.

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. *Interdisciplinary Studies on Information Structure, 3*, 199–244.

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass, 2*(4), 569–588. https://doi.org/10.1111/j.1749-818x.2008.00067.x.

Kendon, A. (1972). Some relationships between body motion and speech. An analysis of an example. In A. W. Siegman & B. Pope (Eds.), *Studies in Dyadic Communication* (pp. 177–210). Elmsford, N.Y.: Pergman Press.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In: M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–227). https://doi.org/10.1515/9783110813098.207.

Kendon, A. (2000). Language and gesture. In D. McNeill (Ed.), *Language and gesture* (pp. 47–63). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.004.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, S. (1990). *The temporal relationship between gesture and speech: A study of Japanese-English bilinguals*. Master's Thesis. University of Chicago.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005.

Krauss, R. M., & Hadar, U. (1999). The Role of Speech-Related Arm/Hand Gestures in Word Retrieval. In R. Campbell & L. Messing (Eds.), *Gesture, Speech and Sign* (pp. 93–116). Oxford: Oxford University Press.

Kwan, T.-W., Tang, W. S., Chiu, T. M., Wong, L. L. Y., Wong, D., & Zhong, L. (2003). 粵語審音配詞字庫 [Chinese Character Database: With Word-formations, Phonologically Disambiguated According to the Cantonese Dialect]. Retrieved from http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/.

Leemann, A., Kolly, M.J., Li, Y., Chan, R., Kwek, G., & Jespersen, A. (2016). Towards a typology of prominence perception: the role of duration. In: *Proceedings of Speech Prosody 2016* (pp. 445–449). Boston, USA. https://doi.org/10.21437/SpeechProsody.2016-91.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218.

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71–89. https://doi.org/10.1515/lp-2012-0006.

Man, C. H. V. (1999). *An acoustic study of the effects of sentential focus on Cantonese tones*. Unpublished Master's Thesis, University of Victoria.

Man, C. H. V. (2002). Focus effects on Cantonese tones: An acoustic study. In: *Proceedings of Speech Prosody 2002* (pp. 467–470). Aix-en-Provence, France. Retrieved from https://www.isca-speech.org/archive_open/sp2002/sp02_467.pdf.

McNeill, D. (1992). *Hand and mind: What gestures reveals about thought*. Chicago: University of Chicago Press.

McNeill, D. (2005). In *Gesture and thought*. https://doi.org/10.7208/chicago/9780226514642.001.0001.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language and Hearing Research, 51*(6), 1507–1521. https://doi.org/10.1044/1092-4388(2008/07-0173).

Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In: *Proceedings of Speech Prosody 2010*. Chicago, USA. Retrieved from http://speechprosody2010.illinois.edu/papers/100110.pdf.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of prosody and position on the timing of deictic gestures. *Journal of Speech, Language and Hearing Research, 56*(2), 458–470. https://doi.org/10.1044/1092-4388(2012/11-0283).

Wang, B., Zhang, Y., Xu, Y., & Ding, H. (2016). Prosodic focus in three northern Wu dialects: Wuxi, Suzhou and Ningbo. In: *Proceedings of the 8th Experimental Linguistic Conference* (pp. 117–120). Crete, Greece.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* (pp. 1556–1559). Genoa, Italy. Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf.

Wu, W., & Xu, Y. (2010). Prosodic focus in Hong Kong Cantonese without post-focus compression. In: Speech Prosody 2010 (pp. 1–4). Chicago, USA.

Xu, Y., Chen, S., & Wang, B. (2012). Prosodic focus with and without post-focus compression: A typological divide within the same language family? *Linguistic Review, 29*, 131–147. https://doi.org/10.1515/tlr-2012-0006.

Xu, Y., Xu, C., & Sun, X. (2004). On the temporal domain of focus. In: *Proceedings of Speech Prosody 2004* (pp. 81–84). Nara, Japan. Retrieved from https://www.isca-speech.org/archive/sp2004/papers/sp04_081.pdf.

Zou, Y., Wang, Y., & He, W. (2012). Diachronic contrastive analysis on read speech in broadcast news: Evidence from pitch and duration. In *Proceedings of the 8th International Symposium on Chinese Spoken Language Processing* (pp. 291–295). https://doi.org/10.1109/ISCSLP.2012.6423498.