Co-Production of Contrastive Prosodic Focus and Manual Gestures: Temporal Coordination and Effects on the Acoustic and Articulatory Correlates of Focus

Benjamin Roustan and Marion Dohen

Speech and Cognition Department – GIPSA-lab – UMR5216 CNRS, Grenoble University
961, rue de la Houille Blanche, 38402 Saint-Martin-d'Hères, France

Abstract

Speech, and prosody in particular, is tightly linked to manual gestures. This study investigates the coordination of prosodic contrastive focus and different manual gestures (pointing, beat and control gestures). We used motion capture on ten speakers to explore this issue. The results show that prosodic focus "attracts" the manual gesture whichever its type, the temporal alignment being stricter for pointing and mainly realized between the apex of the pointing gesture and articulatory vocalic targets. Moreover, it appears that the production of a gesture, whichever its type, does not affect the acoustic and articulatory correlates of prosodic focus.

1. Introduction

The production of speech and manual gestures are tightly bound (see *e.g.* [1, 2]). In particular, [3–8] underline the link between manual gestures and prosody.

Deixis is the ability to draw attention. In the concrete space, it can be achieved through index finger pointing. In speech, it can be achieved through the use of focus. In particular, prosodic contrastive focus is used to emphasize a word or group of words as opposed to another. In this sense, prosodic focus and pointing are strongly linked (see [9] for discussion). The literature provides several clues on how the different unimodal realizations of deixis are coordinated in time. de Ruiter [10] found that the position of lexical stress within a single word did not affect the temporal realization of the concomitant pointing gesture. However, he also found that the onset of the pointing gesture was influenced by the location of contrastive stress within a noun phrase (adjective + noun). Rochet-Capellan *et al.* [11] used a pointing-naming task (2-syllable non-word target with lexical stress on one or the other syllable) and found that the stressed syllable is always included in the part of the gesture that shows (from the apex to the return of the index finger toward a rest position). There thus seems to be a link between the temporal realization of deixis in speech (expressed through prosody) and manual pointing but it remains unclear. Moreover, these studies analyzed simple speech productions (isolated words, non-words or phrases). [1, 5] also put forward the potential link between prosody and beat gestures (or batons; small up and down flicks of the hand).

Another crucial question is the influence of the production of manual gestures on speech. Krahmer and Swerts [12] found that the production of a "visual beat" (manual beat gesture, head nod or rapid eyebrow movement) has a significant effect on duration and F_2 formant in the same direction as the effect of the production of a pitch accent. This suggests that the production of a manual or facial gesture has an effect on the acoustic and articulatory correlates of the simultaneously produced speech.

The aim of this study is to analyze the coordination between contrastive prosodic focus in French and several types of manual gestures in entire utterances using motion capture. The questions addressed are the following: 1. how are prosodic focus and manual gestures coordinated in time? 2. is this potential coordination influenced by the type of gesture and its functional link to speech? 3. does the production of a manual gesture affect the acoustic and articulatory correlates of prosodic focus?

2. Methodology

2.1. Experimental design

2.1.1. Corpus

The corpus used consisted of four subject (S) - verb (V) - object (O) sentences in French (ex: Mumu tient le bébé. 'Mumu holds the baby.') with the following syllable structure: S=2 syl (first name); V=1syl (action verb, present tense); O=1+2syl (article + common noun). All target word syllables (S and O) were CV syllables and the first consonant of the word was a bilabial consonant.

2.1.2. Experimental design

Two narrow focus conditions were elicited: focus on the subject (SF) and on the object (OF). Four gesture conditions were explored: no gesture (speech alone), index-finger pointing (deictic communicative), beat gesture (non-deictic communicative) and control gesture (button press; non-deictic non-communicative).

A correction task was used to naturally elicit the production of prosodic contrastive focus. Participants heard an audio prompt in which two speakers were talking and were instructed to correct the sentence uttered by the second speaker as in the following example (capital letters signal focus):

Prompt	S1	Mumu tient le bébé.
	S2	S1 a dit : Baba tient le bébé?
		'S1 said: Baba holds the baby?'
Answer	Participant	MUMU tient le bébé.

Two images were then displayed (for the latter example: images of Baba and Mumu). The participants were instructed to gesture at the same time as they performed the correction task. In the pointing condition, they had to point at the image corresponding to the correction performed. In the beat condition they were instructed to produce a rapid up-down flick of the hand (gesture often used in spontaneous speech when focusing). In the control condition, participants had to press a button on the table. No further indication was given on when to gesture.

The experiment was divided into 4 blocks (one for each gesture condition). Participants briefly trained before each block The order of the blocks was varied across participants. Within each block, the order of the sentences and focus conditions was random and varied across participants. Each block consisted of 16 trials (4 sentences, 2 focus conditions, 2 repetitions).

2.2. Experimental setup

The participants sat in a chair facing a screen on which visual targets appeared. A rest position was marked on a table on their right-hand side. They were instructed to place their index finger on this mark and always depart from this position to gesture and come back to it after the gesture. The movements of their lips and right hand were tracked using a 3D motion capture device (NDI Optotrak). Four markers were positioned on the lips (2 on each lip corner, 1 in the middle of the upper and 1 in the middle of the lower lip) and three on their right hands (index finger: 1 on the tip and 1 on the first phalanx; back of the hand: 1). The acoustic speech productions were recorded using a microphone (synchronized with Optotrak recording).

2.3. Participants

Ten adults (8 women, 2 men; age: mean: 30.2, s.d.: 8.94) participated in the experiment. All were native speakers of French and right-handed.

2.4. Measurements

All the stimuli were acoustically validated by two independent judges The acoustic production errors were excluded from analysis. Utterances in which participants had omitted to produce or had misproduced the manual gesture were also excluded from analysis.

The acoustic boundaries of syllables were labeled using Praat [13]. The fundamental frequency (F_0) and intensity (Int) peaks of the focused constituent (S or O) were detected. The duration of the focused constituent (Dur) was computed. Lip opening (distance between upper and lower lip markers) and protrusion (forward movement of upper lip marker) were extracted from the motion data. The two articulatory vocalic targets (VT_1 and VT_2) corresponding to the two vowels of the focused constituent were also detected (peaks of amplitude of lip opening or lip protrusion depending on the vowel). For the finger movement data, the apex (P_A) and beginning of the return stroke (P_R) of the gesture were labeled. For the pointing gesture, the apex corresponds to the complete extension of the index finger pointed at the target. For the control gesture, it corresponds to the point at which the index finger reaches the button. For the beat gesture, it was identified as the end of the downbeat. The timing corresponding to the realizations of the above mentioned events were normalized against the acoustic duration of the sentence (beginning of the sentence: 0; end: 1). This was done in order to compensate for effects of semantic content or response time.

All dependent variables were tested using two-way ANOVAs with two within subject factors: focus condition (2 levels: SF, OF) and gesture condition (for gesture variables *i.e.* P_A and P_R: 3 levels: pointing, beat and control gestures; for acoustic and articulatory variables *i.e.* Dur, F_0, Int, VT_1 and VT_2: 4 levels: speech only + 3 gesture types).

3. Results

3.1. Timing: Speech/gesture temporal coordination

3.1.1. *General results*

Table 1 provides the results of the statistical analyses on the normalized times of occurence of the different events measured (notation $\mathrm{t}_\mathrm{event}$).

Table 1. *Two-way ANOVAs on all time variables.*

	focus condition	**gesture condition**
$\mathrm{t}_{\mathrm{P}_\mathrm{A}}$	$F(1,9) = 114.4\ \ , p < .001$	$F(2,18) = 13.9, p < .01$
$\mathrm{t}_{\mathrm{P}_\mathrm{R}}$	$F(1,9) = 99.5\ \ \ \ , p < .001$	$F(2,18) = 3.4\ \ , p = .09$
$\mathrm{t}_{\mathrm{F}_0}$	$F(1,9) = 1571.6, p < .001$	$F(3,27) = 3.1\ \ , p = .1$
t_Int	$F(1,9) = 2478.6, p < .001$	$F(3,27) = 7.6\ \ , p = .01$
$\mathrm{t}_{\mathrm{VT}_1}$	$F(1,9) = 3746.1, p < .001$	$F(3,27) = .4\ \ \ \ , p = .7$
$\mathrm{t}_{\mathrm{VT}_2}$	$F(1,9) = 2655.7, p < .001$	$F(3,27) = 1.2\ \ , p = .4$

Manual gesture time variables – We found significant main effects of *focus condition* on both apex ($\mathrm{t}_{\mathrm{P}_\mathrm{A}}$) and return ($\mathrm{t}_{\mathrm{P}_\mathrm{R}}$) times. The gesture tends to occur later within the utterance for OF: focus "attracts" the gesture. We found a significant main effect of *gesture type* on apex but not on return time. This suggests that the different gestures are not produced in the same manner.

Speech time variables – We found significant main effects of *focus condition* on all variables ($\mathrm{t}_{\mathrm{F}_0}$, t_Int, $\mathrm{t}_{\mathrm{VT}_1}$

and $\mathrm{t_{VT_2}}$). This corresponds to the fact that the acoustic and articulatory correlates were measured on S for SF and O for OF. We found no significant effects of *gesture condition*. The production of a gesture (whichever its type) does not affect the internal temporal organization of the utterance (no difference with speech only condition).

3.1.2. Temporal alignments

Two points of interest are said to be aligned in time if the difference of their times of occurence is close to zero. In order to examine the potential alignment of manual gestures and prosodic focus we computed the differences between the times of occurence of gesture variables ($\mathrm{t_{P_A}}$ and $\mathrm{t_{P_R}}$) and those of acoustic ($\mathrm{t_{F_0}}$ and $\mathrm{t_{Int}}$) and articulatory ($\mathrm{t_{VT_1}}$ and $\mathrm{t_{VT_2}}$) correlates for each utterance. A mean was computed over all utterances for each participant. Figure 1 shows the means and standard deviations of the time differences over all participants (if a box is close to zero, the variables considered are close in time).

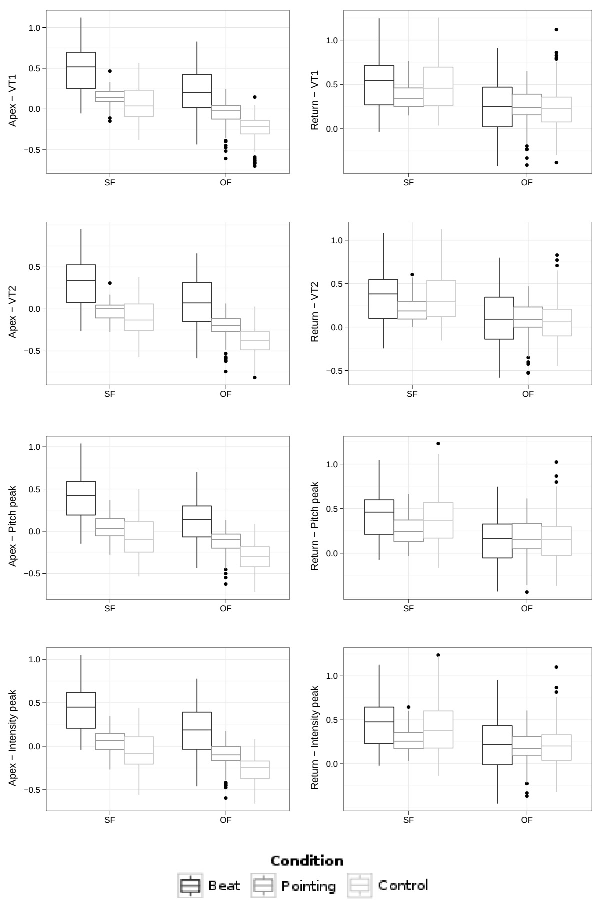

Figure 1: *Differences in timing between gesture apices (t_{P_A}) and returns (t_{P_R}) and acoustic (t_{F_0}, $t_{P_{Int}}$) and articulatory correlates (t_{VT_1}, t_{VT_2}) for all focus conditions (SF, OF) and all gesture types (pointing, beat, control).*

Table 2 provides the results of the ANOVAs on all time differences. We found significant main effects of *focus condition* on all time differences. Whichever the gesture type, temporal coordination between speech and manual gestures within the focused constituent differs from SF to OF. We also found significant main effects of *gesture type* on all time differences for $\mathrm{P_A}$ but not for $\mathrm{P_R}$. Different manual gestures are thus coordinated with prosodic focus in different ways: $\mathrm{P_R}$ is constant for all gestures relative to acoustic and articulatory correlates (consistent with the results presented in section 3.1.1) but not $\mathrm{P_A}$.

Table 2. *Two-way ANOVAs on all time differences.*

	focus condition	**gesture condition**
$\mathrm{t_{P_A}}$ - $\mathrm{t_{F_0}}$	$F(1,9) = 47.8, p < .001$	$F(2,18) = 14.1, p < .01$
$\mathrm{t_{P_A}}$ - $\mathrm{t_{Int}}$	$F(1,9) = 55.3, p < .001$	$F(2,18) = 14.4, p < .01$
$\mathrm{t_{P_A}}$ - $\mathrm{t_{CV_1}}$	$F(1,9) = 57.5, p < .001$	$F(2,18) = 14\ \ , p < .01$
$\mathrm{t_{P_A}}$ - $\mathrm{t_{CV_2}}$	$F(1,9) = 55.6, p < .001$	$F(2,18) = 14\ \ , p < .01$
$\mathrm{t_{P_R}}$ - $\mathrm{t_{F_0}}$	$F(1,9) = 32.8, p < .001$	$F(2,18) = 3.1\ \ , p = .1$
$\mathrm{t_{P_R}}$ - $\mathrm{t_{Int}}$	$F(1,9) = 32.9, p < .001$	$F(2,18) = 4.5\ \ , p = .05$
$\mathrm{t_{P_R}}$ - $\mathrm{t_{CV_1}}$	$F(1,9) = 36.3, p < .001$	$F(2,18) = 3.4\ \ , p = .08$
$\mathrm{t_{P_R}}$ - $\mathrm{t_{CV_2}}$	$F(1,9) = 36.3, p < .001$	$F(2,18) = 3.1\ \ , p = .1$

Welch t-tests were used to compare the normalized time variables corresponding to the different gestures and acoustic and articulatory variables.

Pointing gesture – For SF, $\mathrm{t_{P_A}}$ is not significantly different from $\mathrm{t_{F_0}}$ ($\mathrm{t(9) = 1,\ p = .3}$) and $\mathrm{t_{Int}}$ ($t(9) = 1.5,\ p = .2$) but this is not the case for OF ($\mathrm{t_{F_0}}$: $\mathrm{t(9) = -3.11, p = .008; t_{Int}\text{: } t(9) = -2.7, p = .02}$). For SF, $\mathrm{t_{P_A}}$ is not significantly different from $\mathrm{t_{VT_2}}$ ($\mathrm{t(9) = -0.8, p = .4}$) and for OF, $\mathrm{t_{P_A}}$ is not significantly different from $\mathrm{t_{VT_1}}$ ($\mathrm{t(9) = -1.5,\ p = .2}$). It thus appears that, for pointing, there is a clear tendency towards alignment of the apex with one of the articulatory targets of the focused constituent.

Beat gesture – For SF, there is no tendency towards any alignment between $\mathrm{P_A}$ and any of the acoustic and articulatory correlates of prosodic focus. For OF, $\mathrm{P_A}$ appears to be aligned with the $\mathrm{F_0}$ and intensity peaks ($\mathrm{F_0}$: $\mathrm{t(9) = 1.4,\ p = .2}$; Int: $\mathrm{t(9) = 2,\ p = .08}$) and $\mathrm{VT_2}$ ($\mathrm{t(9) = .7, p = .5}$).

Control gesture – For SF, $\mathrm{P_A}$ is aligned with $\mathrm{F_0}$ and intensity peaks ($\mathrm{F_0}$: $\mathrm{t(9) = -1.2,\ p = .3}$; Int: $\mathrm{t(9) = -0.8,\ p = .4}$). However, this is not the case for OF. For SF, $\mathrm{P_A}$ is aligned with $\mathrm{VT_1}$ ($\mathrm{t(9) = .8, p = .4}$). For OF, $\mathrm{P_R}$ is aligned with $\mathrm{VT_2}$ ($\mathrm{t(9) = 1.2, p = .3}$).

Very interestingly, Figure 1 also shows that alignments are more precise and consistent across speakers for the pointing gesture (see standard deviations). This is especially true for the apex.

3.2. Amplitude: Effect of gesture production on acoustic and articulatory correlates

We analyzed the mean amplitudes of acoustic (duration of the focused element (Dur) and Int and $\mathrm{F_0}$ peaks) and articulatory ($\mathrm{VT_1}$ and $\mathrm{VT_2}$) correlates. Table 3 provides the results of the ANOVAs on these variables.

We found significant main effects of *focus condition* on all five variables. Post-hoc analyses showed that the amplitudes of all acoustic and articulatory correlates of prosodic focus are unsurprisingly significantly smaller for OF than for SF (see [14]). More interestingly, we found no significant main effect of *gesture condition* on any of the variables. It therefore appears that the production of a gesture does not affect the acoustic and articulatory correlates of prosodic focus (no difference between speech + gesture and speech only conditions).

Table 3. *Two-way ANOVAs on all amplitude variables.*

	focus condition	**gesture condition**
Dur	$F(1,9) = 13.5, p = .05$	$F(2,18) = .4 \ , p = .7$
F_0	$F(1,9) = 17 \ \ , p < .01$	$F(2,18) = 2.7, p = .1$
Int	$F(1,9) = 76.2, p < .001$	$F(2,18) = .5 \ , p = .2$
VT_1	$F(1,9) = 13.5, p < .01$	$F(2,18) = 2.3, p = .2$
VT_2	$F(1,9) = 59.6, p < .001$	$F(2,18) = 3.4, p = .08$

4. Conclusions and discussion

The productions of ten speakers were recorded (motion capture on hand and mouth and acoustic recording) under two focus conditions (subject *vs.* object focus) and four gesture conditions (speech only *vs.* pointing *vs.* beat *vs.* control gestures). The aim was to address several questions: determine whether the productions of prosodic focus and manual gestures are coordinated and how this coordination can differ depending on the type of gesture produced; examine the effect of the production of a manual gesture on acoustic and articulatory correlates of speech.

The results show that focus and manual gestures are coordinated in the sense that focus "attracts" the manual gesture. Gestures' apices occur within or close to the focused element. Concerning pointing, in line with [11], we found that the part of speech that shows (prosodic focus) and the part of the gesture that shows always overlap. Temporal coordination between focus and manual gestures is the most consistent for pointing. The apex of the pointing gesture is aligned with articulatory targets rather than acoustic correlates. It appears that temporal coordination between prosodic focus and manual gestures depends on the functional link between gesture and speech. Coordination is the most consistent for pointing which realizes the same function as prosodic focus *i.e.* deixis. Coordination is the least precise for the control gesture which is a non-communicative gesture even if their appears to be a coordination (we expected there would be none at all). This may be due to the fact that button pressing in this case was close to pointing (participants had to extend their index finger to press the button).

Concerning question 3, the results show that the production of a manual gesture (whichever its type) does not affect the acoustic and articulatory correlates of prosodic focus. There is indeed no significant difference in the amplitudes of these correlates from speech alone to combined speech and gesture conditions. These results are not in line with the findings of [12] who found that producing a beat gesture (whether manual or facial) is likely to increase muscular activity related to articulation. We believe that the effects found by these authors are artefacts of their experimental paradigm. In some conditions, participants were asked to produce a beat gesture on a different word than the one bearing the pitch accent. This type of incongruent productions are quite unnatural and it is likely that the speakers tended to produce pitch accents on the word they were asked to produce a beat gesture on.

This study enabled a more precise characterization of the coordination between speech prosody and manual gestures. Gesture apices mainly seem to be aligned with articulatory targets especially for pointing which is functionally close to prosodic focus. During the recordings, it however appeared that it was quite difficult for speakers to produce manual beat gestures when instructed to do so. It would therefore be crucial to conduct motion capture recordings in natural conversation to confirm these findings.

5. References

[1] D. Mcneill, *Hand and Mind: What Gestures Reveal about Thought.* University Of Chicago Press, 1992.

[2] A. Kendon, "Gesture," *Annual Review of Anthropology*, vol. 26, no. 1, pp. 109–128, 1997.

[3] D. Bolinger, "Intonation and gesture," *American Speech*, vol. 58, pp. 156–174, 1983.

[4] S. Nobe, "Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production," Ph.D. dissertation, The Faculty of the Division of the Social Sciences, 1996.

[5] E. Mcclave, "Pitch and manual gestures," *Journal of Psycholinguistic Research*, vol. 27, no. 1, pp. 69–89, 1998.

[6] J. Boyer, A. Di Cristo, and I. Guaïtella, "Rôle de la voix et des gestes dans la focalisation," in *Oralité et gestualité. Interaction et comportements multimodaux dans la communication*, C. Cavé, I. Guaïtella, and S. Santi, Eds. L'Harmattan, 2001, pp. 459–463.

[7] D. P. Loehr, "Gesture and intonation," Ph.D. dissertation, Faculty of the Graduate School of Arts and Sciences of Georgetown University, 2004.

[8] S. Duncan, "Gesture and speech prosody in relation to structural and affective dimensions of natural discourse," in *GESPIN - Gesture & Speech in Interaction*, 2009.

[9] H. Lœvenbruck, M. Dohen, and C. Vilain, *Pointing is special.* Peter Lang Verlag, 2009, pp. 211–258.

[10] J. P. de Ruiter, "Gesture and speech production," Ph.D. dissertation, Catholic University of Nijmegen, Netherlands, 1998.

[11] A. Rochet-Capellan, R. Laboissière, A. Galvan, and J.-L. Schwartz, "The speech focus position effect on jaw-finger coordination in a pointing task," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 1507–1521, December 2008.

[12] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, 2007.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 1995-2009. [Online]. Available: www.praat.org

[14] M. Dohen and H. Lœvenbruck, "Pre-focal rephrasing, focal enhancement and post-focal deaccentuation in french," in *Interspeech-ICSLP 2004*, 2004, pp. 1313–1316.