

# Gesture and Speech Coordination: The Influence of the Relationship Between Manual Gesture and Speech

*Benjamin Roustan and Marion Dohen*

Speech and Cognition Department – GIPSA-lab – UMR5216 CNRS, Grenoble University

{benjamin.roustan, marion.dohen}@gipsa-lab.grenoble-inp.fr

## Abstract

Communication is multimodal. In particular, speech is often accompanied by manual gestures. Moreover, their coordination has often been related to prosody. The aim of this study was to further explore the coordination between prosodic focus and different manual gestures (pointing, beat and control gestures) on ten speakers using motion capture. As compared to previous studies, results show that the coordination between gestures and speech is modulated by the relationship between the manual gesture and speech, especially for the pointing gesture. Moreover, this study shows that different strategies might be adopted so as to adapt to the changes in this relationship. **Index terms:** Speech/Gesture coordination, pointing, beats, prosodic focus, multimodal deixis

## 1. Introduction

Manual gestures are often co-produced with speech in natural communication (see *e.g.* 1; 2) and some authors have put forward a link between prosody and manual gestures (*e.g.* 3; 4). An important issue is the study of the coordination between manual gestures and speech.

Two types of communicative gestures appear to be interesting to study relatively to prosody in general and prosodic focus, namely index finger pointing and beats (or batons; small up and down flicks of the hand). Pointing is used to show the object of interest in space. In speech, prosodic focus is used to emphasize a word or a group of words in an utterance in order to designate it as being the main object of communication. Pointing and prosodic focus therefore appear to be strongly linked (see 5, for further discussion). Several studies provide preliminary information on how pointing and focus are coordinated in time. de Ruiter (6) found that the onset of a pointing gesture was influenced by the location of contrastive stress within a noun phrase (adjective + noun). Rochet-Capellan *et al.* (7) also found that the pointing gesture was shifted towards the lexically stressed syllable among two. Concerning beats, several authors have suggested that they were linked to prosody and focus in particular (1; 4).

In a previous study (detailed in 8, hereafter referred to as *Exp1*), we used motion capture to investigate the

temporal coordination between prosodic contrastive focus and different types of manual gestures (pointing, beat and control gesture *i.e.* button press). Simple subject-verb-object sentences were used in which either the subject or the object was focused. Participants were simply instructed to produce the manual gesture while speaking. We found that prosodic focus “attracts” the manual gesture whichever its type (most of the gesture apices are realized within the focused constituent). The tightest coordination was observed for the pointing gesture and was realized between the pointing apex and a speech articulatory target.

When comparing the results of different quantitative studies on speech / manual gesture coordination, one can notice that the results are variable. These discrepancies could be accounted for by the variations in the nature of the communicative relationships between manual gestures and speech. Concerning pointing for example, in *Exp1*, the pointing gesture showed the same object as the one being vocally focused. In de Ruiter’s study, where the pointed object represented the entire noun phrase, coordination was different whether the distinctive feature in speech was a property of the object (its color) or the object itself. This suggests that, even though manual pointing designates a larger constituent than prosodic focus (object + color), it is influenced by the location of prosodic focus (object or color). However this study only dealt with the onset of the pointing gesture which is not the part of the gesture that actually shows.

The aim of this study is to explore the potential modulation of speech / gesture coordination by the communicative relationship between manual gestures and speech. While *Exp1* tested instances in which pointing and focus designated the same element, the present study will examine cases in which the sentence and the pointed object describe the same information but prosodic focus affects only part of that information.

## 2. Methodology

### 2.1. Participants

Ten right-handed adult native speakers of French (8 women, 2 men), aged 30.2 on average (s.d.: 8.94), participated

in the experiment. The participants were the same as in *Exp1* (order of experiments random).

## 2.2. Materials

Six French subject (S) - verb (V) - object (O) sentences were used (ex: Le bonbon est rouge - 'The candy is red'). The syllable structure was the following: S=1+2 syl (article + object name); V=1syl (state verb, present tense); O=1syl (color). All S syllables were CVs and all O syllables were CVCs.

## 2.3. Experimental design

We explored four *gesture conditions*: speech alone, index finger pointing (deictic communicative gesture), beat gesture (non-deictic communicative) and control gesture (button press; non-deictic non-communicative) as well as two narrow *focus conditions*: subject (SF) and object (OF) focus.

Participants performed a correction task which naturally elicited the production of prosodic contrastive focus. They heard an audio prompt consisting of a declarative sentence. Two images appeared on a screen which naturally induced the correction of the sentence heard. The following example gives an idea of how the experiment went (small capitals signal focus):

**Audio prompt** – Le bonbon est vert. ('The candy is green.')

**Images** – red candy and yellow balloon

**Participant** – Le bonbon est ROUGE. ('The candy is RED.')

The instruction was to gesture at the same time as speaking. In the pointing condition, participants had to point at the corresponding image. They therefore pointed at the entire object (in this case a red candy) whereas they vocally focused only one property of the object (in this case the fact that it is red). In the beat condition, they were instructed to produce a rapid up-down flick of the hand. In the control condition, they had to press a button on a table. No further indication was given on when to gesture.

The experiment consisted of 4 blocks (one for each gesture condition) each starting with a brief training session. One block consisted of 24 trials (6 sentences, 2 focus conditions, 2 repetitions). The order of the blocks was varied across participants as well as the order of the sentences and focus conditions within each block.

## 2.4. Experimental setup

Participants sat in front of a screen on which the images were projected. A table was located on their right-hand side (rest position mark and button for control gesture). Participants were instructed to place their right forefinger on the rest position before and after gesturing. An infrared motion-capture device (NDI Optotrak) was used to track their manual and articulatory movements: four markers were placed on the lips (2 on each lip corner, 1

in the middle of the upper lip and 1 in the middle of the lower lip). Audio was recorded with a microphone.

## 2.5. Measurements

Production errors were discarded from analysis (error in gesture type, gesture omission or speech error). Two independent judges assessed the acoustical productions of participants to check for correct focus production.

Praat (9) was used to label syllable boundaries. Acoustic cues (pitch and intensity) and articulatory movements were also analyzed but the results are not presented here. Brachiomanual movements were characterized using apex ( $P_A$ ) and beginning of the return stroke ( $P_R$ ). The segment between  $P_A$  and  $P_R$  corresponds to the gestural hold. The apex of the pointing gesture corresponds to the farthest point reached by the index finger. For the control gesture, the apex is the point at which the button is pressed. The apex of the beat gesture was labelled as the end of the downbeat. A time normalization against the acoustic duration of the utterance was performed so as to overcome effects of semantic content or response time (beginning of the utterance: 0; end of the utterance: 1).

## 3. Predictions

Several predictions can be made on the basis of the results from *Exp1*. The main difference between the two experiments concerns the relationship between manual pointing and vocal pointing: the focused element in speech (ex: red) only corresponds to a property of the element pointed at (ex: a red candy). As far as temporal coordination between manual pointing and speech is concerned, several predictions can be made. Temporal coordination should be different in this experiment. We expected no difference in coordination from subject to object focus (the element pointed at is the same in both cases, ex: a red candy). Concerning what this coordination could be, several possibilities can be put forward: 1. the pointing gesture covers the entire utterance; 2. the pointing gesture is located in the middle (partly on the S and partly on the O); 3. the pointing gesture is synchronized either with S or with O. Concerning beat gestures, if they are considered as being coordinated to prosodic events, we expected no difference with the findings from *Exp1*. The same prediction can be made for the control gesture which is supposedly unrelated to speech in terms of communicative intention.

## 4. Results

All dependent normalized time variables ( $t_{P_A}$  and  $t_{P_R}$ ) were tested using two-way ANOVAs with two within subject factors: focus condition (2 levels: SF, OF) and gesture condition (3 levels: pointing, beat and control gestures). The results are presented in Table 1. The main ef-

Table 1: Results of the two-way ANOVAs on gestural time variables.

	P <sub>A</sub>	P <sub>R</sub>
Focus	$F(1, 9) = 11.49, p < .01$	$F(1, 9) = 14.78, p < .01$
Gesture	$F(2, 18) = 25.77, p < .001$	$F(2, 18) = 1.49, p = .25$
Focus×Gesture	$F(2, 18) = 0.34, p = .71$	$F(2, 18) = 2.13, p = .15$

fect of *focus condition* is significant for both apex ( $t_{P_A}$ ) and return ( $t_{P_R}$ ) times. Similarly as in *Exp1*, the manual gesture tends to occur later for OF. The main effect of *gesture condition* is also significant for  $t_{P_A}$  but not for  $t_{P_R}$ . This suggests that the different gestures are not produced in the same manner. Finally, the interaction is never significant which suggests that the different gestures are equally impacted by focus condition. This general analysis suggests that the results are not different from *Exp1* which corresponds to our predictions for beat and control gestures but not for pointing. However, a closer look at individual data reveals that participants use two very different strategies concerning temporal coordination between manual gestures and speech especially for pointing.

As opposed to the findings from *Exp1* suggesting that the pointing apex and hold are attracted by the focused constituent, in this experiment only 4 participants out of 10 maintain this strategy. The participants can therefore be divided into two groups: those using the same strategy as in *Exp1* (Group 1: 6,8,9,10) and those using a different strategy (Group 2: 1,2,3,4,5,7). If the same ANOVA presented in Table 1 is performed independently for both groups (see Table 2), the effect of focus condition on both  $t_{P_A}$  and  $t_{P_R}$  is significant for group 1 but not for group 2. This suggests that, for group 2, the manual gestures are realized in the same manner for both focus conditions. In order to explore this into more details, the locations of the pointing gestures relative to the focused constituent were plotted for each participant and both experiments (see Figure 1). Figure 2 shows the difference in timing of the apices and acoustic beginning of focus from *Exp1* to the present experiment (a negative value corresponds to an event occurring earlier in this experiment than in *Exp1*).

Table 2: Results of the two-way ANOVAs on gestural time variables for group 1 (same strategy as in *Exp1*) and group 2 (different strategy as in *Exp1*).

	P <sub>A</sub>	P <sub>R</sub>
Focus	$F(1, 3) = 53.36, p < .01$	$F(1, 3) = 94.43, p < .01$
Gesture	$F(2, 6) = 13.83, p < .001$	$F(2, 6) = 2.23, p = .19$
Focus×Gesture	$F(2, 6) = 0.73, p = .52$	$F(2, 6) = 6.7, p < .05$
Focus	$F(1, 5) = 2.25, p = .19$	$F(1, 5) = 3.85, p = .11$
Gesture	$F(2, 10) = 12.09, p < .01$	$F(2, 10) = 0.19, p = .83$
Focus×Gesture	$F(2, 10) = 0.48, p = .63$	$F(2, 10) = .07, p = .93$

It appears (see Figure 2) that, for SF, there is a systematic and equal shift of the focused constituent and the apex for all gesture types: they occur later than in *Exp1*. This is simply due to the fact that, in this experiment, there was a determiner before the noun. This suggests that coordination is similar for both experiments for SF. For OF, however, there are differences between gestures.

**Pointing gesture** – For group 1 (except S5), Figure 1 shows that, for SF, the temporal coordination between speech and manual pointing is the same as in *Exp1*. For OF, however, manual pointing is drastically shifted from the object (*Exp1*) to the subject (current experiment). Figure 2 shows that, for OF, the apex indeed occurs much earlier in this experiment than in *Exp1*. This shift cannot be explained by a mere shift in the object location (shift of object from *Exp1* to current experiment:  $-0.061$ ; shift of apex:  $-0.155$  which is twice as big as the acoustic shift).

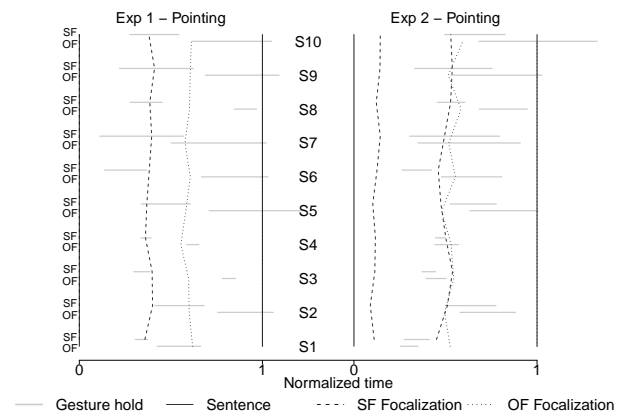


Figure 1: Temporal organization of speech and manual pointing for each participant. Time values are normalized against the acoustic timing of the utterance (0: beginning of the utterance, 1: end of the utterance)

**Beat gesture** – In *Exp1* we did not find beats to be precisely synchronized with prosodic focus especially for SF even if there was a general tendency for the gesture to occur later for OF (8). This was accounted for by the fact that participants appeared to find this manual gesture to be difficult to produce “on-demand” which could explain the lack of coordination between acoustic prosodic cues and the beat flick. In the current experiment and for most participants (8 out of 10), there was a focus condition in which the coordination was similar to that in *Exp1*. In the other focus condition, the beat tended to shift and showed a coordination close to the latter focus condition. This was contrary to our predictions that there would be no difference between the two experiments for beats. Figure 2 however shows that the shift of the apex for the beat gesture in OF was much smaller than that observed for pointing and not much greater than the mere shift of the

object.

**Control gesture** – In *Exp1*, the coordinative pattern observed for the control gesture was quite similar to that observed for manual pointing (see 8). In this experiment, it appears that coordination is different than that observed in *Exp1*. Just like the pointing gesture, the control gesture tends to occur earlier in OF. The shift is however smaller than for pointing and not much bigger than the mere shift of the object.

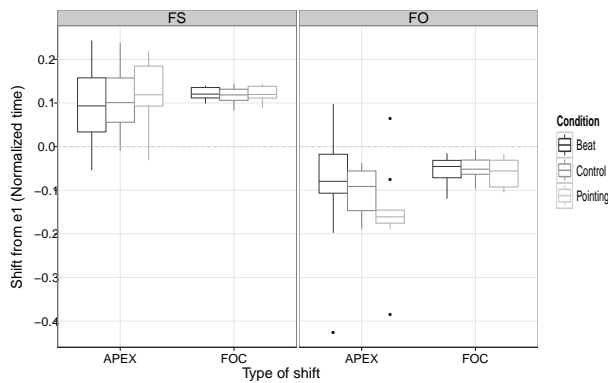


Figure 2: Temporal shifts for apex and focused component from *Exp1* to the present experiment

## 5. Conclusions and discussion

The aim of this study was to test the influence of the communicative link between manual gesture and speech on their coordination. In particular, the manual pointing targets included information on the object's nature and its color whereas, in speech, either the object's nature (its name) or its color was contrastively focused though both information were present (ex: The candy is RED). The methodology used was the same as in a previous study (8, *Exp1*) in which the manual pointing target corresponded to the focused constituent in speech. The purpose was to compare coordinative strategies in both experiments. The productions of ten speakers were recorded (motion capture and acoustic recording) under two focus conditions (subject vs. object) and four gesture conditions (speech alone vs. speech + pointing vs. beat vs. control gestures).

A preliminary general analysis suggested that the global coordinative patterns did not differ from those observed in *Exp1*: focus attracts manual gestures. This corresponded to our expectations for beat and control gestures but not for manual pointing (see section 3). However, a more in detail examination of the results revealed that the attraction of the gesture by the focused constituent is weaker than in *Exp1*. Participants can actually be divided into two groups: some use the same coordinative pattern as in *Exp1* but not the others especially for pointing. For the latter group, for SF in the pointing gesture condition, the coordinative pattern is indeed approximately the same as for *Exp1*. For OF, however, even if

there is still a tendency towards shifting the manual gesture towards the object, the shift is much smaller than in *Exp1*. The manual gesture is actually temporally stretched over the two vocal constituents. Most often, there is a greater overlap of the gestural hold with the focused constituent which results in the weak difference observed between focus conditions. The difference between SF and OF is thus clearly less marked than in *Exp1*. As in *Exp1*, the control gesture seems to yield similar coordinative patterns as manual pointing which can be explained by the fact that it may be too close to pointing (see 8, for further discussion). However, the results are closer to those of *Exp1* than for pointing. Finally, even though the difference between experiments exists, beats do have a tendency to be coordinated to speech in a quite similar way as in *Exp1*.

It therefore appears that coordinative patterns are modulated by the relationship between manual gestures and speech. In *Exp1* pointing was coordinated to prosodic focus but this coordination changes when the relationship between speech and manual pointing changes. Beats are rather simply coordinated to focus production which is illustrated by the fact that the results are approximately the same as in *Exp1* except for a small temporal shift which is not much greater than the shift of the focused part of speech. Moreover, as in *Exp1*, it appears that the variation in the coordinative patterns used by participants is much smaller for pointing than for other manual gestures.

## 6. Acknowledgments

We are grateful to Coriandre Vilain for his technical help as well as to Jean-Luc Schwartz for his advice.

## 7. References

- [1] D. Meneill, *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, 1992.
- [2] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, October 2004.
- [3] D. Bolinger, "Intonation and gesture," *American Speech*, vol. 58, pp. 156–174, 1983.
- [4] E. Z. McClave, "Pitch and manual gestures," *Journal of Psycholinguistic Research*, vol. 27, no. 1, pp. 69–89, 1998.
- [5] H. Løevenbruck, M. Dohen, and C. Vilain, *Pointing is special*. Peter Lang Verlag, 2009, pp. 211–258.
- [6] J. P. de Ruiter, "Gesture and speech production," Ph.D. dissertation, Catholic University of Nijmegen, Netherlands, 1998.
- [7] 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.
- [8] 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 conference*, Chicago, USA, May 2010.
- [9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 1995-2010. [Online]. Available: [www.praat.org](http://www.praat.org)