Co-production of speech and pointing gestures in clear and perturbed interactive tasks: multimodal designation strategies

Marion Dohen, Benjamin Roustan

▶ **To cite this version:**

Marion Dohen, Benjamin Roustan. Co-production of speech and pointing gestures in clear and perturbed interactive tasks: multimodal designation strategies. Interspeech 2017 - 18th Annual Conference of the International Speech Communication Association, Aug 2017, Stockholm, Sweden. ⟨hal-02367749⟩

HAL Id: hal-02367749

https://hal.science/hal-02367749v1

Submitted on 18 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Co-production of speech and pointing gestures in clear and perturbed interactive tasks: multimodal designation strategies

Marion Dohen[1], *Benjamin Roustan*[2]

[1] Univ. Grenoble Alpes, CNRS, GIPSA-lab, F-38000 Grenoble, France
[2] UroMems SAS, F-38000 Grenoble, France
marion.dohen@gipsa-lab.grenoble-inp.fr

Abstract

Designation consists in attracting an interlocutor's attention on a specific object and/or location. It is most often achieved using both speech (e.g., demonstratives) and gestures (e.g., manual pointing). This study aims at analyzing how speech and pointing gestures are co-produced in a semi-directed interactive task involving designation. 20 native speakers of French were involved in a cooperative task in which they provided instructions to a partner for her to reproduce a model she could not see on a grid both of them saw. They had to use only sentences of the form 'The [target word] goes there.'. They did this in two conditions: silence and noise. Their speech and articulatory/hand movements (motion capture) were recorded. The analyses show that the participants' speech features were modified in noise (Lombard effect). They also spoke slower and made more pauses and errors. Their pointing gestures lasted longer and started later showing an adaptation of gesture production to speech. The condition did not influence speech/gesture coordination. The apex (part of the gesture that shows) mainly occurred at the same time as the target word and not as the demonstrative showing that speakers group speech and gesture carrying complementary rather than redundant information.

Index Terms: multimodality, pointing, speech/gesture coordination, perturbed communication, noise, interaction

1. Introduction

Designation is a communicative function aiming at putting forward specific information for the interlocutor (e.g., location, object or both). It is most often achieved using multimodality especially speech combined with manual gestures. Index finger manual pointing is very efficient to convey designation [1]. In speech, it can be achieved in various ways. Demonstratives such as 'here' or 'there' can for example be used to convey location information. According to Diessel [2], demonstratives play an important role in joint attention between interlocutors. Designation can also be conveyed in speech using prosodic focus, which puts forward a word or a group of words within an utterance (e.g., [3]).

In communication, speech and manual gestures are tightly coupled and coordinated [4-6]. Even though several studies shed light on how this coordination actually occurs in designation tasks (e.g., [7-10]), there still needs further work to understand the precise underlying mechanisms. In two previous studies we examined multimodal designation using manual pointing and prosodic focus. In particular, we analyzed how hand and mouth were coordinated in such tasks. Both studies [11,12] showed that prosodic focus attracts manual pointing whether the image pointed at exactly corresponds to the focused word ([11]) or not ([12]). Most of the time the apex of the pointing gesture (the part of the gesture that shows) co-occurs with prosodic focus. Also note that these studies showed that the apex of the pointing gesture is most often aligned with an articulatory rather than an acoustic target. This is in line with [9] but not with [10] who found alignment of manual gestural features with the tone rather than the articulatory gesture. An important issue is however to further understand how speech and manual gestures are coordinated in terms of information transmission. Are they grouped based on function (the pointing gesture co-occurs with the part of speech achieving designation)? Or are they grouped in terms of information transmission (e.g., the pointing gesture carrying information on the location co-occurs with the part of speech that describes what is placed at that location)?

Speech/gesture coordination may also be modulated by constraints imposed by the communicative environment. For example, we know that speech production is affected by the presence of background noise: it is the Lombard effect [13]. People speak louder, slower, the acoustic and articulatory properties of their speech are modified… (e.g., [14]). In noise the acoustic channel is altered, so it is easily understandable that speech is modulated since it is conveyed acoustically. The visual channel, through which gesture is conveyed, is however not. Manual gestures do not need to be produced differently in noise, but are they? And since they are coordinated with speech, which is modified by the presence of noise, how is speech / gesture coordination affected?

Also note that most of the studies cited above used controlled laboratory tasks and no actual inter-personal interaction. But designation is directed towards an interlocutor. Speech / gesture co-production could therefore be different in a true interactive setting in which the speaker transmits actual information to an interlocutor.

The purpose of the present study is to examine the co-production of speech and pointing gestures in a semi-directed interactive task involving designation. The task naturally elicits the production of manual pointing gestures and the semi-directed speech material contains demonstratives. It was achieved in two conditions: clear vs. noisy environment.

2. Methods

2.1. Participants

A total of 20 native speakers of French participated in the study (10 females – age: mean=26.9 sd=7.7). All were right-handed (Edinburgh handedness inventory, [15]) and reported normal or corrected to normal vision and no auditory impairments.

2.2. Experimental material

Participants were required to use the following carrier sentence and only that sentence: '*Le X va là.*' (The X goes there). This was used in order to ensure that the target words (TW) X were always produced in a similar context. A total of six bisyllable CVCV target words were used: *pompon* (pompom), *bonbon* (candy), *chameau* (camel), *chapeau* (hat), *bambou* (bamboo) and *lama* (llama). Each TW was associated to a drawing (see Figure 1).

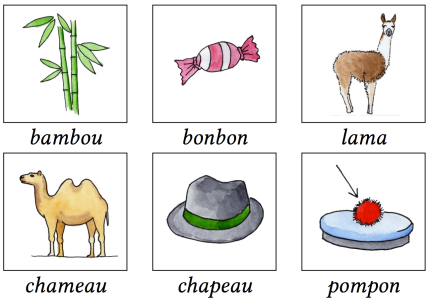

Figure 1: *Drawings for each target word.*

2.3. Experimental procedure

2.3.1. Task

The participants were informed that they would be playing a cooperative game. Its aim was to provide a partner with the necessary information for her to put down cards with pictures (Figure 1) at the right location on a random grid (Figure 2). The partner was an accomplice, the same for all participants.

Both partners started with an empty grid, located in between them, marked with 12 dots at random locations. They were informed that the positions marked by the dots corresponded to the locations where the cards should be placed. The accomplice had the cards with the drawings and the participant saw a model with 12 images (each image appeared twice – see Figure 2 for an example) displayed on a screen the accomplice could not see. The participant then had to give instructions to the accomplice using only sentences of the type described in section 2.2. This naturally elicited the production of a pointing gesture (the demonstrative '*là*' is not understandable without the gesture). The accomplice interacted only by putting down the cards (no speech).

Participants first filled in 3 grids in silence. Then they were informed that they would be doing this in noise (3 grids). Silence always preceded noise in order to record the most natural productions possible in silence. If participants had had the noisy condition first their subsequent productions in silence could have been influenced. The order of the models presented was randomized across participants. The participant was instructed that she was free to move as long as she stayed seated on the chair. Once the partners managed to fill in a grid correctly, a new model was displayed to the participant.

2.3.2. Experimental conditions

Participants were tested on the task in two different experimental conditions: *silence* and *noise*. In the noisy condition both interlocutors wore headphones in which they heard a cocktail party noise (BDBRUIT database, [16]). Noise level was set to perturb spoken communication but not to make it impossible (the spoken message was perceptible but difficult to understand, 85dB(C)).

Figure 2: *Example of a model participants saw.*

2.3.3. Experimental setup

The experiment took place in a sound-proof room at GIPSA-Lab. The game zone (70cm long, 53cm wide) was set on a table. The participants sat on a chair in front of the table facing a motion capture device (Optotrack 3020, Northern Digital). They wore a headband with three optoelectronic markers (to correct for head motion). Four markers were fixed on their mouth (upper and lower lip in the middle and both corners). Both their index fingers were equipped with two markers (tip and middle). Finally, one marker was placed on the back of both hands. The 3D coordinates of all markers were recorded with a sampling frequency of 150 Hz. The table was located at a sufficient distance for the motion capture device to cover the entire surface of the game zone as well as the participant.

The accomplice was seated in front of the participant (a bit on the right). Apart from the participant-accomplice axis, at her right, was a screen on which the model was displayed. The screen was also used for displaying instructions. When relevant, participant and accomplice wore headphones (AKG-K77) in which the exact same noise was heard. A video camera located beneath the Optotrak also filmed the participant and the game zone. The participant's speech was recorded using a microphone (AKG C1000S – Fs = 44100 Hz). Motion capture and audio recordings were synchronized by simultaneously sending a bip signal on the audio track and on one of the Optotrak leds.

2.4. Analyses

The acoustic data were labeled using Praat [17]. The following landmarks were extracted for the 1st and 2nd syllables of TW and for the demonstrative: acoustic boundaries, intensity (Int) and fundamental frequency (f0) peaks. Int and f0 peaks were also extracted over all the utterance.

Articulatory features were extracted from the motion capture data: lip opening (distance between upper and lower lip markers) and protrusion (forward movement of upper lip marker). From this, we extracted the vocalic articulatory targets (lip opening or lip protrusion peak amplitude depending on the vowel) for the 1st and 2nd syllables of TW and for the demonstrative.

Manual gestural features were also extracted from the motion capture data and four landmarks were labeled for each pointing gesture: 1. Onset (hand starts moving); 2. apex

(farthest point reached by the index finger); 3. return (finger departs apex position); 4. Offset (hand stops moving). A fifth feature was also analyzed: gestural hold (segment between 2 and 3). Note that the phase between events 1 and 2 (resp. 3 and 4) corresponds to the onset (resp. offset) stroke.

The times of occurrence of the all the above events were normalized against utterance duration so as to compensate for duration differences (0: beginning of utterance; 1: end).

The video data was labeled using ELAN ([18]) in order to extract several features. Each utterance was qualified as being an initial production or an utterance resulting from misunderstanding (error reparation). Participant gaze direction was labeled from the videos and classified into 4 categories: towards the model, towards the interlocutor, towards the game zone, elsewhere. Even though gaze direction labeling is not precise using video recordings, the zones categorized are sufficiently large to obtain reliable annotations.

Task completion time was computed as the difference between the time the model was displayed on the participant's screen and the time the last card was placed on the game zone resulting in a totally correct reproduction of the model.

Statistical analyses reported below are Welch t-tests or repeated measure ANOVAs (significance threshold: $p<.05$).

3. Results

3.1. General description of the interaction – effect of noise

Speech acoustics and articulation – We found the traditional correlates of Lombard speech in noise (e.g. [14]): higher intensity and f0, lower speech rate, hyper-articulation. The analyses are not reported in detail for the sake of space.

Durations – Task completion took significantly more time in *noise* than in *silence* (*silence*: 33.3s, *noise*: 46s – $t(19)=-10.6$, $p<.001$). Utterances were also significantly longer in noise (*silence*: 0.9s, *noise*: 1.2s – $t(19)=-8.9$, $p<.001$) even though this lengthening cannot alone explain larger task completion times since the time spent speaking over the overall duration of the task did not vary between conditions ($t(19)=-1.7$, $p=0.1$). In *noise* there were significantly more pauses within utterances (+15.7% – $t(19)=-3.8$, $p<.001$) but hesitations were as frequent in both conditions ($t(19)=-1.6$, $p=0.1$).

Error reparations – An error corresponds to a case in which a valid utterance was produced by the participant but did not result in correct task completion leading to the production of a new utterance by the participant. Errors could correspond to three cases and result in 3 types of error reparations: 1. the participant made a mistake in her first production leading to an auto-correction; 2. the accomplice made a mistake (e.g., putting the wrong card at the wrong place) leading to a correction by the participant (allo-correction); 3. the accomplice did not understand the participant's first utterance leading to a repetition. As expected there were more error reparations in *noise* (about 10 times more) and the great majority of them were allo-corrections (accomplice misunderstanding what the participant said). Note that most corrections are linked to a misunderstanding of the target since the location is clearly shown by the pointing gesture.

Gaze direction – Table 1 provides the mean proportions of time the participant spent looking into one of the four zones relative to the total task completion duration. In *silence*, participants mainly looked at the game zone and the model (~45% of the time for both directions). They almost never looked at their interlocutor. In *noise,* participants spent significantly more time looking at their interlocutor than in silence ($t(19)=-11.1$, $p<0.001$) and significantly less time looking at the model ($t(19)=6.9$, $p<.001$) even though the difference is also significant for game zone ($t(19)=2.2$, $p<.05$).

Table 1: *Proportions of time spent looking at the four zones as a function of condition (Game = game zone).*

Condition	**Gaze direction**			
	Game	**Model**	**Interlocutor**	**Elsewhere**
silence	46.9%	45%	6%	0%
noise	42.4%	29%	27.4%	0.3%

Gesture production – All participants produced manual pointing gestures during the experiment, which is trivial since the task indirectly compelled them to do so. They produced very little or no other manual gestures except for one participant who extensively used iconic representational gestures in *noise* (but not in *silence*). Although participants were all right-handed there were high inter-individual differences in hand use for pointing during the task: some systematically pointed with their right hand while others systematically used their left hand or both hands indifferently. Hand preference for pointing did not vary in *noise*. Pointing gestures were significantly longer in *noise* (+451 ms – $t(18)=-3.7$, $p<.001$), which was solely due to a longer gestural hold (+440 ms – $t(18)=-3.7$, $p<.001$) the durations of all other gesture phases (onset and offset strokes) being the same in *silence* and in *noise*. Gestures also started later in noise ($t(18)=-2.4$, $p<.05$) and apex and onset times were less variable (onset: $t(17)=2.4$, $p<.05$ – apex: $t(17)$, $p<.01$).

3.2. Speech / gesture coordination

The aim of the analyses described hereafter is to understand how speech and pointing gestures were coordinated in time and how this coordination was affected by noise. Table 2 shows that, regardless of condition, pointing gestures were almost always initiated before speech onset. The return most often occurred after speech offset. The gestural hold systematically at least partly overlapped speech. The apex almost always occurred during speech and most of the time during TW pronunciation and especially its first syllable (or just before it).

Table 2: *Proportions of: gestures initiated before speech onset, apex produced during speech and more particularly during target word and demonstrative pronunciation, return produced after speech offset and gestural hold at least partially overlapping speech.*

			silence	***noise***
Gesture starts before speech			96.5%	97.1%
Apex	**In speech**		85.1%	92.9%
	Before target word		12.8%	26%
	In target word	**Syllable 1**	37%	46.2%
		Syllable 2	24.6%	12.4%
	In demonstrative		2.3%	3.2%
Return after speech offset			76.9%	82.9%
Gestural hold and speech overlap			99.3%	100%

We then analyzed whether there were any alignments between gestural and speech (acoustic or articulatory) features. Figure 3 provides the distributions of normalized times of occurrence of the gesture apex and several acoustic and articulatory features. Statistical analyses show that the time of occurrence of the gesture apex does not appear to be significantly different from that of the vocalic articulatory target ($t(19)=0.02$, $p=.98$) and the intensity ($t(19)=0{,}66$, $p=.5$) and f0 ($t(19)=0.4$, $p=.7$) peaks of TW syllable 1. The smallest differences in times of occurrence were observed for apex and articulatory target.

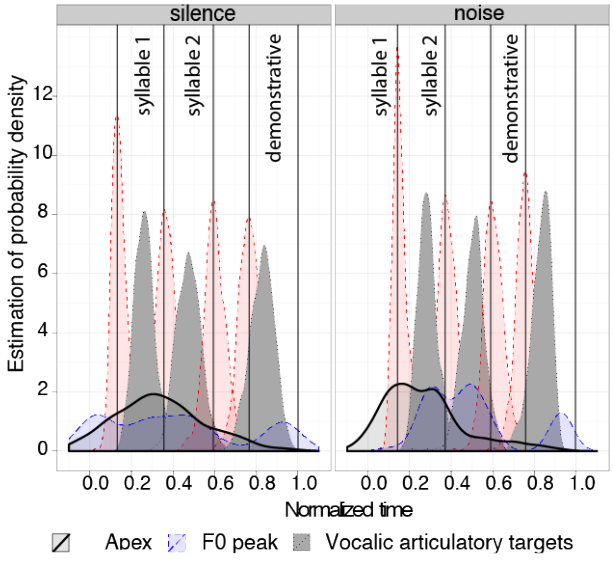

Figure 3: *Distributions of times of occurrence (over all productions) of: gesture apex (black), acoustic boundaries of 1st and 2nd syllables of target word and demonstrative (red), f0 peak over the entire utterance (blue) and vocalic articulatory targets (grey).*

4. Discussion and conclusions

This study examined the multimodal behaviors of 20 participants involved in a semi-directed interactive task involving designation. The task was performed in two conditions: silence and noise. Participants were asked to provide instructions to an interlocutor (accomplice) for her to place cards at correct locations on a grid following a model only the participant could see. Acoustic, articulatory and manual productions were recorded during the entire duration of the experiment (microphone, video, motion capture).

We found that the participants displayed all the acoustic and articulatory correlates of the Lombard effect when speaking in noise (e.g. [14]). It also took longer for the communicative partners to complete the task in noise. Participants did not speak more (same relative duration compared to overall task duration), but they spoke slower and made more pauses. As expected, the accomplice made more errors requiring more reparations by the participant. More surprisingly the participants also made more errors in noise. This could be explained by the fact that the presence of noise added cognitive load to the participant resulting in speech production errors. Another interesting result is that, in silence, participants mainly direct their gaze toward the game zone or the model and barely ever towards their interlocutor. In noise, they spend more time looking at their interlocutor and less time looking at the model. They could be compensating for the lack of acoustic intelligibility by providing more visual speech information to the interlocutor or gathering feedback on the interlocutor's understanding.

In noise, the pointing gestures were initiated earlier and the gestural holds were lengthened. Onset and offset strokes were not affected. Whereas speech lengthening in noise is due to overall slower articulation, it is not the case for gesture. The dynamics of gesture production are unchanged in noise (onset and offset strokes not lengthened) but their timing is (gestural hold lengthened by the same amount as general utterance lengthening). Even though the pointing gesture is not affected by noise per se, since it is conveyed through the visual channel, its timing is modified in noise. The manual gesture thus seems to adapt to the changes in speech production in order to maintain the speech / gesture coherence ([7,9]).

Concerning the temporal coordination between gesture and speech, three different predictions could be made: the pointing gesture could occur at the same time as the target word (corresponding to the drawing on the card the interlocutor should pick up), the demonstrative (corresponding to the speech element conveying designation) or something else. The results show that the apex of the pointing gesture mainly occurs within the target word. This suggests that participants tend to group complementary information (target image through speech and location through pointing) rather than communicative functions (demonstrative and pointing which both convey designation).

19 out of 20 participants did not change communicative strategies from silent to noisy conditions, one of them did. Even though this is not statistically relevant, the change in strategy is particularly interesting since, in noise, this participant started using iconic gestures aligned with the target word combined with pointing gestures aligned with the demonstrative. In this case, an iconic gesture was used to compensate for the lack of intelligibility of the target word due to noise. This strategy was efficient since the interlocutor made much less errors than with the other participants.

Finally, as in our previous studies ([11,12]), we found that the tightest alignment between gestural and speech features was between the apex of the pointing gesture and a vocalic articulatory target of the target word. This is in line with [9] but contrary to what was observed by Krivokapic et al. [10] who found tightest alignments between pointing gesture and tone gestures. Note however that supra-laryngeal movements (lips) and F0 peak appear to be correlated [19].

Even if one of the objectives of the current study was to investigate speech / gesture coordination in a more natural setting than in previous studies, i.e. in an interactive setting, we still had to use fixed speech material in order to address the research questions raised. The next step would be to use the same paradigm but with no constraint on the speech material. This would make it less possible to use replicable statistical analyses but would be very interesting to study how speech and gestures are naturally combined in a designation task.

5. Acknowledgements

We would like to thank Hugues Naigeon for the drawings as well as all the participants to this study, Coriandre Vilain for his technical assistance and Jean-Luc Schwartz for his scientific advice.

6. References

[1] S. Kita, *Pointing, Where Language, Culture, and Cognition Meet. Lawrence Erlbaum Associates*, 2003.

[2] H. Diessel, "Demonstratives, joint attention, and the emergence of grammar", *Cognitive Linguistics*, vol. 17, no. 4, pp. 463-489, 2006.

[3] E. O. Selkirk, "The grammar of intonation", In E. O. Selkirk (Ed.), *Phonology and syntax: the relation between sound and structure*, The MIT Press, 1984, pp. 197-296.

[4] D. McNeil, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[5] D. McNeil, Language and Gesture. Cambridge University Press, 2000.

[6] P. Wagner, Z. Malisz, and Kopp, S., "Gesture and speech in interaction: An overview", *Speech Communication*, vol. 57, pp. 209-232, 2014.

[7] W. J. Levelt, G. Richardson, and W. La Heij, "Pointing and voicing in deictic expressions", *Journal of Memory and Language*, vol. 24, no. 2, pp. 133-164, 1985.

[8] J. P. de Ruiter, *Gesture and speech production*, PhD dissertation, Catholic University of Nijmegen, Netherlands, 1998.

[9] 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, 2008.

[10] J. Krivokapic, M. Tiede, M. E. Tyrone, and D. Goldenberg "Speech and manual gesture coordination in a pointing task", In *Proceedings of Speech Prosody 2016*, Boston, USA, 2016, pp. 1240-1244.

[11] 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", *Proceedings of Speech Prosody 2010-5th International Conference on Speech Prosody*, Chicago, USA, 2010, pp. 100110-1.

[12] B. Roustan, and M. Dohen, "Gesture and speech coordination: The influence of the relationship between manual gesture and speech", *Proceedings of Interspeech 2010,* Japan, 2010.

[13] E. Lombard, "Le signe de l'élévation de la voix", *Annales des maladies de l'oreille et du larynx*, vol. 37, pp. 101-119, 1911.

[14] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and Articulatory Study of Lombard Speech: Global effects on the Utterance", in *INTERSPEECH 2006*, Pittsburgh, USA, 2006.

[15] R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory", *Neurospychologia*, vol. 9, no. 1, pp. 97-113, 1971.

[16] J. Zeiliger, J.-F. Serignat, D. Auteserre, and C. Meunier, "BDBRUIT, une base de données parole de locuteurs soumis à du bruit", *Proceedings of the 10th Journées d'Études sur la Parole*, Lanion, France, 1994, pp. 287-290.

[17] P. Boersma, and D. Weenink, « Praat: doing phonetics by computer » [Computer program]. Version 6.0.26, retrieved from http://www.praat.org/, 2017.

[18] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research", in: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 2006. (http://tla.mpi.nl/tools/tla-tools/elan/)

[19] M. D'Imperio, R. Espesser, H. Lœvenbruck, C. Menezes, N. Nguyen, and P. Welby, "Are tones aligned with articulatory events? Evidence from Italian and French", in J. Cole (Ed.), *Papers in Laboratory Phonology 9*. Mouton Gruyter, 2007, pp. 577-608.