Gesture/speech integration in the perception of prosodic emphasis

Gaëlle Ferré

LLING – UMR 6310 & University of Nantes, France
Gaelle.Ferre@univ-nantes.fr

Abstract

It is now well established that communicative gestures are not produced randomly by speakers but that there are links between these gestures and semantic content as well as syntactic structure. In prosody, it has also been shown that hand gestures align with lexical stress in the words they accompany ([1]) and that particular gesture types like hand beats regularly accompany prosodic emphasis in speech ([2]). Although some studies have started to research possible interactions between hand gestures and prosodic emphasis, the exact influence of gestures on the perception of prosodic emphasis is still largely unknown. This issue is addressed in the present paper in which a pilot perception experiment is conducted. The paper shows that hand gestures have an impact on the perception of prosodic emphasis and that both gesture type (beats vs. pointing hand gestures) and gesture amplitude (large hand gestures as opposed to a medium-sized or a small ones) influence perception. The study therefore sheds new light on speech models and the current debate concerning the level at which gestures interact with speech.

Index Terms: prosodic emphasis, perception, hand gestures, amplitude, gesture/speech interaction

1. Introduction

A growing body of research suggests that language and gesture are part of either the same system (*Growth point theory*: [3], [4]) or two highly integrated systems in which gesture and speech are integrated at different levels of the speech process. [5] examined the effects on speech when gesture production is impeded and when speakers are free to gesture. They concluded in what is called the *Lexical Facilitation Model*, that free gesturing has an impact on speech fluency and that speech and gesture interact at an early stage of speech planning since when gesturing is impeded, speech becomes less fluent. [6] proposed another hierarchical model (the *Sketch Model*) in which gesture takes its source at the conceptualization stage of the message, much before verbalization, and is then not influenced by the verbalization stage that encompasses the grammatical and the phonological modules. Comparing different languages, [7] observe that gesture production varies with different types of syntactic encoding of the message, and that therefore, there must be some interaction between gesture and later stages in the planning of the message than what was thought before. In their *Interface Model*, they find that gestures may help retrieve words, but the syntactic structure of a language also conditions what gestures will be produced at a later stage of speech production. McNeill's work stands in contrast to the models just described as he considers that gesture and speech stem from a single communication system ([2], [4]) which cannot be accounted for by the modular models just presented ([4]).

Strangely enough, one essential component of speech, prosody, has been given very little thought in the conception of the different models (of which only the most influential are quoted here), although the growing number of experiments in multimodal prosody can give us some insight into the integration or dissociation of gesture and speech.

Several studies examined the links between gesture and the expression of prosodic focus, considering either the temporal organization of gesture and speech ([8]), or the types of gesture that accompany prosodic emphasis ([9], [10], [2]). Following the same line of research the present paper addresses two issues:

- Do hand gestures participate in the perception of prosodic emphasis?
- Does hand gesture's amplitude have an impact on the perception of prosodic emphasis?

After a presentation of the major findings in studies on the gesture/prosody relationships (section 2), the experiment design for the present study is described in section 3. Section 4 presents the results obtained both for gesture type and gesture amplitude, which are then discussed in section 5, before the conclusion in section 6.

2. Theoretical background

[9], [10] and [11] studied the links, both from a production and perception perspective, between some gestures (eyebrow raises, hand beats and head nods) and acoustic prominence, especially contrastive focus. They showed in experimental studies that these gestures facilitate the perception of prominence, but also that, when produced together with speech, they influence some acoustic parameters of speech, as shown as well by [12].

In [2], the aim of the study was to find out if the same types of gesture accompany syntactic and prosodic highlighting in the production of spoken French. Two possible ways to highlight some discourse item in French consist in either using a fronted syntactic structure (using left dislocation, topicalization, (pseudo)clefting or presentative constructions), or in marking narrow focus thanks to prosodic emphasis. The study showed that hand beats are regularly associated with prosodic emphasis whereas other gesture types accompany syntactic fronting.

Not any type of hand gesture has an impact on the perception of stress, since [13] found no direct correlation between pointing gestures and the perception of lexical stress. But they did not test the impact of pointings on the perception of other types of focus.

In a perception experiment involving computer synthesized animations, [14] and [15] showed that when head-nods and eyebrow raises accompany prominent syllables, they can aid speech perception. On the other hand, it is not clear to

10.21437/SpeechProsody.2018-7

them whether they will aid or hinder speech perception when they accompany non-prominent syllables.

[16] found that the "presence and salience of the visual cues enhances perception" (p. 2413). They studied the links between prosodic contrastive focus and some lower-face parameters. Their working hypothesis was "that the main articulatory consequence of contrastive focus is hyper-articulation." (op.cit., p. 2414). Their results, in agreement with [17], showed that larger and faster facial movements, especially mouth opening, do have an effect on the perception of both lexical stress and phrasal prosodic emphasis.

3. Experiment design

3.1. Material

In order to answer our research questions, we conducted a pilot perception experiment. 5 short sentences in French were recorded in Praat ([18]) by a participant in a quiet room: each sentence had four syllables and was pronounced without prosodic emphasis (neutral condition) and with prosodic emphasis on the second syllable, thus yielding 10 recorded sentences.

Ma maman vient.	*My mummy comes.*
Il veut venir.	*He wants to come.*
Un rien me va.	*Very little suits me.*
Le mien est là.	*Mine is here.*
La mer est loin.	*The sea is far.*

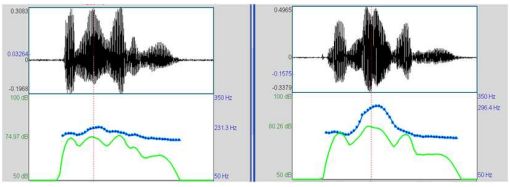

Figure 1: *Praat curves (blue dotted line = F0, green line = Intensity) of "la mer est loin" (neutral condition - left) and "la MER est loin" (emphatic condition - right) produced by the same speaker.*

As shown in Figure 1, emphatic stress was produced essentially by means of a higher pitch and intensity, and the target syllable was also longer under emphasis than in the non-emphatic condition.

In a separate recording, the speaker was filmed by a camera, producing either no gesture at all or producing a small, medium or large pointing hand gesture, as well as a small, medium or large beat hand gesture. Two examples are shown in Figure 2 below. Each audio sentence was then pasted on each video clip using Camtasia Studio, making sure that the gesture stroke was aligned with the target syllable. The speaker's face was blurred to avoid any influence from facial articulatory gestures. The sequences were then edited into an interactive video in a pseudo-random order. In order to avoid an effect of sequentiality which has a strong impact on the perception of stress, none of the sentences was repeated twice in a sequence. An extra clip was added at the end of the video that included a left-handed gesture instead of a right-handed one. There was a total of 70 sequences (+ the extra sequence one which was not taken into account in the study) in which 35 utterances were pronounced without prosodic emphasis and 35 with prosodic emphasis.

Figure 2: *Medium-sized pointing (left) and medium-sized beat (right) at their apex. Arrows indicate movement direction.*

3.2. Participants

15 undergraduate students participated in the experiment via the university intranet. None of them reported any hearing impairment and all had normal or corrected to normal vision. Their curriculum included a course on prosody and all of them were familiar with the notion of prosodic emphasis which was nonetheless explained again before the experiment started. In the experiment proper, participants saw a clip, after which the video stopped playing, showing a still screen with two buttons, "emphatic" and "non-emphatic". After clicking on one button, the video resumed playing. Subjects were recommended to use headphones, to listen to each sequence only once and were told that at the end of the video, they would have to say whether they had seen a sequence or not (the extra clip). The aim of this clip was to ensure that participants didn't close their eyes to concentrate on the audio signal only during the playing of each sequence.

4. Results

To answer our research questions, we used a series of Generalized Linear Mixed Models (GLMMs) fit by maximum likelihood estimation using the R 3.4.0 statistical programming language ([19]) and the lme4 package ([20]). Because there were five different sentences in the experiment setup which might generate variation in the perception of emphasis, we systematically included "Utterance" as a random factor in the models.

4.1. Perception of prosodic contrast

We first explored possible interactions among presence or absence of prosodic emphasis and perception of emphasis by participants (fixed factors = Prosody, Perception; values = yes; no). There was a significant interaction between the two factors: utterances that were effectively pronounced with prosodic emphasis were perceived as emphatic by more participants (β=12.5, SE=.32, p<.001) but utterances that were not pronounced with prosodic emphasis were also perceived as emphatic by more participants (β=1.85, SE=.28, p<.001). This means that the production of a gesture to accompany an otherwise non-emphatic utterance in prosody has an impact on how the utterance is perceived by participants. The mean number of participants who perceived emphasis in utterances that were pronounced with prosodic emphasis is 14.3, whereas the mean number of participants who perceived emphasis in utterances that were not pronounced with prosodic emphasis is 1.85.

4.2. Gesture type

It was expected that the type of gesture produced would have an influence on the perception of prosodic emphasis, i.e. that

beats would influence the perception of emphasis more than pointing gestures. We could not make any prediction concerning a potential effect of pointing gestures as opposed to no gesture at all. In order to test this, we explored possible interactions between gesture types (fixed factor = Type; values = none, pointing, beat) and the number of participants who perceived emphasis in utterances. We found a significant interaction between beats and the perception of prosodic emphasis (β =8.4, SE = 1.19, $p < .001$) in the way that utterances accompanied by beats are judged as emphatic by more participants. On the contrary, the addition of a pointing gesture or the absence of gesture in the utterance has no significant effect on the perception of emphasis.

4.3. Gesture amplitude

The initial expectation regarding gesture amplitude was that regardless of gesture type, more and more participants would perceive some prosodic emphasis in utterances as co-speech gestures grew larger. In order to test this prediction, we explored possible interactions between gesture amplitude (fixed factor = Amplitude; values = none, small, medium, large) and the number of participants who perceived emphasis in utterances. We found a significant interaction between large gestures and the perception of prosodic emphasis (β =8.8, SE = 1.4, $p < .001$) in the way that utterances accompanied by a large gesture are judged as emphatic by more participants. Although the perception of prosodic emphasis increases linearly with gesture size as shown in Figure 3 below, the model doesn't yield significant results for small or medium-sized gestures. This means that if gestures are not distinguished by type, they need to be large to influence the perception of prosodic emphasis.

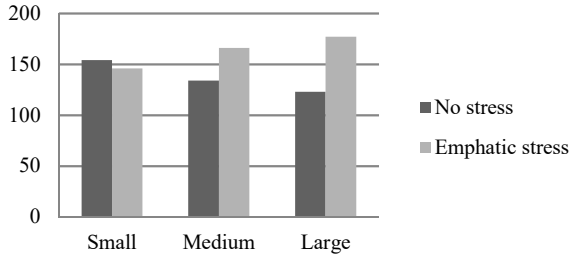

Figure 3: *Number of utterances perceived with prosodic emphasis vs. no emphasis when accompanied with a small, medium or large hand gesture.*

4.4. Emphasis weight

In order to test the interactions between the different parameters in our utterances (real presence/absence of prosodic emphasis, gesture type, gesture size), we calculated an emphasis weight for each utterance used in the experiment. The weight of each utterance was a sum of the different scores assigned for each parameter:

- Absence of prosodic emphasis was noted 0 as opposed to presence of prosodic emphasis noted 1.
- No gesture scored 0, a pointing gesture was noted 1 and a beat counted as 2.
- In terms of gesture size, no gesture counted as 0, a small gesture counted as 1, a medium-sized gesture as 2 and a large gesture as 3.

For all the utterances, emphasis weight thus ranged from 0 (absence of prosodic emphasis and no accompanying gesture) to 6 (presence of prosodic emphasis in an utterance accompanied by a large beat gesture). Intermediate scores were: 1 → presence of prosodic emphasis but no gesture, 2 → small pointing gesture but no prosodic emphasis, 3 → no prosodic emphasis + medium sized-pointing gesture, or prosodic emphasis + small pointing gesture, or no prosodic emphasis + small beat gesture, 4 → no prosodic emphasis + large pointing gesture, or prosodic emphasis + medium-sized pointing gesture, or no prosodic emphasis + medium-sized beat, or prosodic emphasis + small beat, 5 → no prosodic emphasis + large beat, or prosodic emphasis + medium-sized beat, or prosodic emphasis + large pointing gesture.

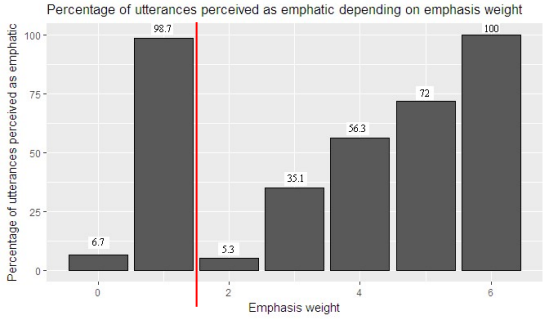

Figure 4: *Percentage of utterances perceived as emphatic by participants depending on their emphasis weight.*

We first found a positive correlation (.37) between emphasis weight and the perception of prosodic emphasis by participants, although this correlation is not extremely high. The reason why the correlation is not as high as one could have expected appears in Figure 4 above which can be analyzed as showing two distinct perception patterns delimited by the red line. Left of the line, for emphasis weights 0 and 1, gestures are not involved. When utterances had no prosodic emphasis and no gesture (score 0), the percentage of utterances perceived as emphatic by the participants was low (6.7 %). Reversely, when utterances were pronounced with prosodic emphasis without any accompanying gesture, the percentage of utterances perceived as emphatic by the participants was high (98.7 %). All the weights right of the red line in Figure 4 involve utterances accompanied by gestures, and this is when perception becomes linear. The higher the emphasis weight, the higher the percentage of utterances perceived as emphatic.

Statistically speaking, the model shows an effect of emphasis weight on the perception of prosodic emphasis for score 0 (β =-9.8, SE = 2.6, $p < .001$), score 2 (β =-10, SE = 1.8, $p < .01$), score 3 (β =-5.5, SE = 1.4, $p < .001$) and score 5 (β =10, SE = 1.3, $p < .001$). Score 6 is not significantly different from score 1. Score 4 is not distinguished either from 3 and 5.

5. Discussion

From this experiment we may conclude that:

(a) Only beats tend to influence the perception of prosodic emphasis but not pointings, although gesture duration (and perhaps velocity) could have an impact on perception as well.

(b) The fact that a sequence contains a gesture as opposed to no gesture at all is not sufficient to induce the perception of prosodic emphasis since there was no difference in between the no gesture and the small or medium gesture conditions.

(c) Gesture amplitude has a direct influence on the perception of prosodic emphasis since utterances which do not contain any prosodic emphasis are more likely to be perceived as emphatic when they contain a large gesture than a medium or a small one. Perception of prosodic emphasis becomes linear when a gesture accompanies an utterance.

Emphatic stress is therefore not only embodied in its production, in that it involves other articulators than the mere vocal tract, it is also embodied in its perception. Participants seeing someone produce a larger hand gesture are convinced that the sentence is uttered with prosodic emphasis even when this is acoustically not the case. This effect is in a way comparable to the so-called McGurk effect ([21]) that creates "auditory-visual illusions" (p. 747) in perception. The authors state that there is a strong influence of the visual modality on the perception of speech in adult subjects and that the item perceived as a result from the auditory-visual combination is a compromise between what has effectively been uttered and the way it has been uttered. The same probably occurs in our study: participants were not expected to rank the degree of stress but just say whether each utterance contained a prosodic emphasis or didn't. Had they been expected to rank the degree of stress, then they would probably have graded utterances with matching emphatic stress and large gesture as uttered with a strong prosodic emphasis, whereas utterances with a large gesture but no real stress would have been considered as uttered with a weaker prosodic emphasis. It will be interesting to test this hypothesis in further work on this particular issue.

Whereas facial gestures help perceive stress through the realization of phonemes, hand gesture amplitude rather helps perceive the presence of rhythmic beats. Yet, a future perception experiment will have to test whether the amplitude of a gesture not directly connected to speech would have the same effect on the perception of prosodic emphasis to determine whether there must be some kind of linguistic integration of the different body movements for amplitude to induce prosodic emphasis or whether amplitude per se may convey prosodic emphasis.

Our data on perceived amplitude presented in the previous section, in which utterances that initially contained no prosodic emphasis are perceived as emphatic when they are accompanied with large hand gestures can be seen as cases of gesture-speech mismatch. This is a prosodic mismatch (the visual modality conveying strong prosodic stress whereas the vocal modality conveys no phrasal stress on the syllable tested), but mismatch can occur in other modalities as well. For instance, [22] tested cases of semantic mismatches and observed that "when gesture and speech convey the same information, they are easier to understand — they are faster and produce fewer errors — than when they convey different information (...) with strong incongruities between speech and gesture bidirectionally affecting integration to a greater extent than weak incongruities. Second, this integration is obligatory: people cannot help but consider one modality (gesture) when processing the other (speech)" (pp. 266-267). They concluded that "gesture influences the processing of speech, speech influences the processing of gesture, and this integration is mandatory" (p. 261). This conclusion is in line with observations in neural imaging in language comprehension. [22] found that "a classical language area, Broca's area, is not only recruited for language-internal processing but also when action observation is integrated with speech. These findings provide direct evidence that action and language processing share a high-level neural integration system. (...) In conversational settings, the brain therefore continuously integrates several streams of language and action-related information that contribute to the listener's understanding of a speaker's message." [23] explain this integration of gesture and speech stating that "both sources of information, the optical and the acoustic, provide information apparently about the same event of talking". Although they mainly concentrated on phonetic articulatory gestures, the explanation can be extended to other gesture types which are not directly connected with articulation proper, but are related to speech at other levels (semantic, syntactic, prosodic, etc...) of production. This is noted as well by [12] who observed in their study that "word and corresponding-in-meaning symbolic [hand] gesture influenced each other when they were emitted simultaneously." [24] rightly note that "we understand spoken speech in terms of how it is produced rather than in terms of its acoustic properties", which may well explain the mismatches between perception and the actual production of speech. What we perceive is a multimodal ensemble ([25]) and if a hand gesture that accompanies speech cues prosodic emphasis in its amplitude or in its type, then prosodic emphasis will be part of the multimodal construction and will be perceived even if it has no acoustic correlate in the utterance.

6. Conclusions

This paper has shown that hand gestures have an influence on participants' perception of prosodic emphasis, even in utterances initially pronounced without prosodic emphasis. Some gestures however have a larger impact on perception than others. Hand beats, which have been shown to be associated with narrow focus in speech production ([9], [10], [11]), also influence the perception of speech, contrary to pointing gestures which are not associated with prosodic emphasis in production. Besides, large gestures contribute more to the perception of prosodic emphasis than medium-sized or small ones. It has been shown as well that perception of prosodic emphasis is linear when gesture is involved: the more utterances are loaded with emphasis cues, the more participants perceive them as prosodically emphatic irrespective of whether utterances are in fact pronounced with prosodic emphasis. It would be interesting in further research to design experiments with Likert scales so that participants could rate the degree of emphasis they perceive in utterances.

Our results also show that perception of prosodic emphasis is more in line with the *Interface Model* ([7]), since there is some interaction between gesture and perception of prosodic phenomena at a much later stage than mere message conceptualization or semantic encoding, thus ruling out the *Lexical Retrieval Model* ([5]) and the *Sketch Model* ([6]) mentioned in the introduction. Prosodic emphasis perception is also treated differently when it includes gesture than when it does not which seems to rule out the idea that gesture and speech are perceived as a whole which would be the case if the two formed a single system as posited by the *Growth Point Theory* ([3]).

7. Acknowledgements

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

8. References

[1] D.P. Loehr. "Gesture and Intonation." PhD Thesis, Georgetown University, 2004.

[2] G. Ferré, "A Multimodal Approach to Markedness in Spoken French." *Speech Communication* 57, Special Issue on Gesture and Speech in Interaction, pp. 268-82, 2014.

[3] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago and London: The University of Chicago Press, 1992.

[4] D. McNeill, *Gesture and Thought*. Chicago and London: The University of Chicago Press, 2005.

[5] R.M. Krauss, Y. Chen, and R. F. Gottesman. "Lexical Gestures and Lexical Access: A Process Model." In *Language and Gesture*. Ed. McNeill, D. Cambridge: Cambridge University Press, pp. 261-283, 2000.

[6] J.P. de Ruiter, "The Production of Gesture and Speech." In *Language and Gesture*. Ed. McNeill, D. Cambridge: Cambridge University Press, pp. 284-311, 2000.

[7] K., Sotaro, and A. Özyürek. "What Does Cross-Linguistic Variation in Semantic Coordination of Speech and Gesture Reveal?: Evidence for an Interface Representation of Spatial Thinking and Speaking." *Journal of Memory and Language* 48, pp. 16-32, 2003.

[8] K.A. Wilmes, "Hands in Focus: Focus Marking by Speech Accompanying Gestures." PhD Thesis: University of Osnabrück, Germany, 2009.

[9] 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* 57, pp. 396–414, 2007.

[10] M. Swerts, and E. Krahmer. "Facial Expression and Prosodic Prominence: Effects of Modality and Facial Area." *Journal of Phonetics* 36, pp. 219-38, 2008.

[11] E. Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink. "Pitch, Eyebrows and the Perception of Focus." In Proceedings of *Speech Prosody 2002*. Ed. Bel, B. and I. Marliens, Laboratoire Parole et Langage, pp. 443-446, 2002.

[12] P. Bernardis, and M. Gentilucci. "Speech and Gesture Share the Same Communication System." *Neuropsychologia* 44, pp. 178-190, 2006.

[13] A. Jesse, and H. Mitterer. "Pointing Gestures Do Not Influence the Perception of Lexical Stress." In Proceedings of *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*. Ed., pp. 2445-2448, 2011.

[14] S. Al Moubayed, J. Beskow, and B. Granström. "Auditory Visual Prominence. From Intelligibility to Behavior." *Journal of Multimodal User Interfaces* 3, pp. 299-309, 2010.

[15] S. Al Moubayed, J. Beskow, B. Granström, and D. House. "Audio-Visual Prosody: Perception, Detection, and Synthesis of Prominence." In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*. Ed. Esposito, A.M. Heidelberg: Springer Verlag, pp. 55-71, 2010.

[16] M. Dohen, and H. Lœvenbruck. "Audiovisual Production and Perception of Contrastive Focus in French: A Multispeaker Study." In Proceedings of *Interspeech/Eurospeech*. Ed., pp. 2413-2416, 2005.

[17] R. Scarborough, P. Keating, M. Baroni, T. Cho, S. Mattys, A. Alwan, E. Auer, and L.E. Bernstein. "Optical Cues to the Visual Perception of Lexical and Phrasal Stress in English." In Proceedings of *Speech Prosody 2006*. Ed., TUDpress Verlag, pp. 217-220, 2006.

[18] P. Boersma, and D. Weenink. "Praat: Doing Phonetics by Computer (Version 5.1.05) [Computer Program]." 2009.

[19] R Core Team. A language and environment for statistical computing. r foundation for statistical computing. [online: http://www.r-project.org], 2012.

[20] D. Bates, M. Maechler, B. Bolker, and S. Walker. Linear mixed-effects models using eigen and s4 [online: http://cran.r-project.org], 2014.

[21] H. McGurk, and J. MacDonald. "Hearing Lips and Seeing Voices." *Nature* 264, pp. 746-748, 1976.

[22] S.D. Kelly, A. Özyürek, and E. Maris. "Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension." *Psychological Science* 21, pp. 260-267, 2010.

[23] C.A. Fowler, and L.D. Rosenblum. "The Perception of Phonetic Gestures." *Speech Research* 99-100, pp. 102-117, 1989.

[24] M. Gentilucci, and M.C. Corballis. "From Manual Gesture to Speech: A Gradual Transition." *Neuroscience and Biobehavioral Reviews* 30, pp. 949-960, 2006.

[25] N.J. Enfield. *The Anatomy of Meaning. Speech, Gesture, and Composite Utterances*. In. Cambridge: Cambridge University Press, 2009.