Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variability

Marion Dohen°, Hélène Lœvenbruck* & Harold Hill°*

* Institut de la Communication Parlée, Grenoble, France
° Human Information Science Research Labs, ATR, Kyoto, Japan
marion.dohen@gmail.com ; helene.loevenbruck@icp.inpg.fr

Abstract

This study is a follow-up of previous studies we conducted on the visible articulatory correlates of French prosodic contrastive focus. A two speaker analysis using an automatic lip-tracking device had shown that these correlates existed and were used in visual perception. However the articulatory strategies depended on the speaker. The purpose of this study was thus to extend the analysis to other speakers, examine the similarities and variabilities and try to identify global tendencies.

We recorded five speakers of French with a 3D optical tracker using a 13 sentence (subject-verb-object) corpus and four focus conditions (S, V, O or neutral). An articulatory analysis confirmed that visible articulatory correlates exist for all the speakers. The strategies used are mainly of two types: absolute and differential. An analysis of other facial movements showed that an eyebrow raising and/or a head nod can signal focus. This association is however highly inter- and intra-speaker dependent.

1. Introduction

Deixis, or the ability to draw attention to an object or a person is crucial in communication. It can be achieved in speech through the use of focus. For example, contrastive focus is used to emphasize a word or group of words in an utterance as opposed to another. In French, it can be conveyed either by syntactic extraction (*c'est X qui* 'it is X who') or by prosodic focus by using a specific intonational contour on the constituent pointed at (*XXXf a mangé la pomme.* 'XXXf ate the apple.').

Most studies of prosodic contrastive focus in French have explored only laryngeal and pulmonic correlates i.e. essentially acoustic parameters. Yet some studies have shown that there are articulatory mouth correlates to prosodic focus in other languages (e.g. [1-7]). These correlates should be visible and there are thus reasons to think that prosodic focus is not only auditory but also visual. [8] indeed showed that there exists a set of visible cues to "phrasal stress" in English for six first names produced in three different positions in a carrier sentence. However, no similar studies have been conducted for French.

2. Background: previous studies

We previously analyzed [9] the possible articulatory visible correlates of contrastive focus in French for two speakers (A & B) using a very accurate lip-tracking device [10]. We observed lower face articulatory visual cues to contrastive focus in French for both speakers. It appeared that contrastive focus was characterized by an increase (hyper-articulation) in inter-lip area as well as in inter-lip area peak velocity for the focal constituent. The amount of hyper-articulation was however highly dependent on the speaker. Speaker A hyper-articulated much more than speaker B. Protrusion could be analyzed only for speaker B due to corpus constraints. It seems however that it is also hyper-articulated and to a greater extent than inter-lip area. It also appeared that speaker B hypo-articulated the post-focal sequence (reduced lip and jaw movements and velocity) while speaker A barely did. Durational measurements were also conducted since duration can be used as a visual cue as well. These measurements showed that the focused syllables were significantly lengthened for both speakers, the first phoneme of the focused phrase being even more significantly lengthened. For speaker A, it was also observed that the last syllable of a phrase was significantly lengthened and hyper-articulated when the following phrase was focused. This was related to an anticipation strategy. We concluded from these observations that there is a global tendency towards hyper-articulating the focused constituent but that other visible cues are produced and that they seem to depend on the speaker. This is why it seemed important to extend this study to a greater number of speakers in order to identify possible global strategies.

Visual only perception tests were also conducted using the videos of speakers A and B [9]. These tests showed that contrastive focus could be perceived through the visual modality alone and that the visual cues used for perception corresponded at least in part to those identified in the production studies. For both speakers, a few stimuli were well perceived even though the visible correlates described above were not present. This suggested that other more subtle facial correlates may intervene. Studies on other languages have indeed shown that other facial movements such as eyebrow movements [11, 12] or head movements [13-15] or both [16] could intervene. [17] also showed that F0 variations and eyebrow movements could be linked in French. This is why it seems necessary to enlarge the set of facial movements measured by the use of a complementary technique.

3. Experimental methods

3.1. The corpus

The corpus used for this study consisted of 13 sentences with a subject-verb-object (SVO) structure, CV syllables and, whenever possible, sonorants. Below is an example of one of the sentences used (the numbers next to S, V and O correspond to the number of syllables of the phrase).

$[\text{Lou}]_{S1}$ $[\text{ramena}]_{V3}$ $[\text{Manu.}]_{O2}$
'Lou gave a lift back to Manu.'

3.2. The audiovisual recording

The corpus described above was recorded for five native speakers of French (B, C, D, E and F). Four focus conditions were elicited: subject-, verb- and object-focus (narrow focus) and a neutral version (broad focus). In order to trigger focus, the speakers had to perform a correction task. They were thus indirectly induced to produce focus on one of the phrases (S, V or O). They heard a prompt in which two speakers were talking and they were then asked to correct a phrase which had been mispronounced. The recording went as follows (where capital letters signal focus):

Audio prompt: S1: Romain ranima la jolie maman.
S2: S1 a dit : Denis ranima la jolie maman?
'S1 said: Denis revived the good-looking mother?'
Speaker utters: ROMAIN ranima la jolie maman.

The speakers were given no indication on how to produce focus (e.g. which syllables to focus). Two repetitions of each utterance (one sentence spoken in one focus condition) were recorded.

3.3. Data acquisition

For the recordings, we used a 3D optical tracking system: Optotrak (less accurate on lip contours than the system used in [9] but providing more facial data). The system consists of three infrared (IR) cameras used to record the speaker who has infrared emitting diodes (IREDs) glued to the face. The 3D coordinates of each IRED are automatically detected over time. For this experiment, we used two Optotraks in order to compensate for missing data. A total of 24 IREDs were glued to the speakers' faces. An additional 4 IREDs were attached to a head rig and were used to extract the "rigid body" movements corresponding to the head movements and thus correct for head motion. IRED positions were sampled at 60Hz and low-pass filtered. The acoustic signals were recorded simultaneously and sampled at 22kHz. Fig. 1 gives an idea of the experimental setup used.

Figure 1: *Optotrak measurement device: experimental setup.*

3.4. Preliminary data analysis

The first step was to acoustically validate the recorded data i.e. to check whether focus had actually been produced acoustically. On the one hand, it was checked that the focused utterances displayed a typical focused intonation as described in [18]. On the other hand, an informal auditory perception test was conducted in order to check that focus was indeed perceived through the auditory modality. This validation procedure showed that all the speakers had produced focus correctly from an acoustic point of view.

3.5. Measurements

3.5.1. Durational measurements

The durations of all the syllables were computed after an acoustic labeling of the corpus. The previous studies ([9]) indeed showed that the focal syllables were lengthened and that sometimes the pre-focal syllable was also lengthened.

3.5.2. Facial movements

Articulatory measurements
In our previous studies ([9]), we had mainly analyzed two articulatory features namely inter-lip area and protrusion. It was put forward that these parameters best represented the high segmental articulatory variability of real speech and would thus be the most relevant parameters in order to isolate supra-segmental originating variations. However, it is not possible to accurately compute inter-lip area from Optotrak data. This is why, in this study, we analyzed separately lip opening (difference between the z coordinates of the upper and lower middle lip markers) and lip spreading (difference between the y coordinates of the two lip corner markers). Jaw vertical movements were also analyzed using the chin marker (z coordinate). Upper lip protrusion was computed as well (x coordinate of the middle upper lip marker).

Facial movements: measurements
Based on other studies of the facial movements accompanying speech and more specifically prosody, we decided to limit our study to the head and eyebrow movements. [17] showed that eyebrow movements accompanying prosody were mainly raising movements. Therefore we decided to study the raising of both the left and the right eyebrows (z coordinates of both middle markers of the eyebrows). As for head movements, the three rotations and translations of the rigid body were available. [15, 16] found that the main movements related to prosody were nods. We therefore analyzed the rotation of the rigid body around the y axis.

Data shaping
The area under the curve of variation of each parameter over time was automatically detected for each phrase and then divided by the duration of the phrase. This normalized area represents the mean amplitude of the parameter considered. After this computation, we get three values per utterance and per parameter considered.

3.5.3. The comparison issue: normalization

In order to be able to isolate and compare supra-segmental articulatory variations for different segmental constituents, we used a normalization technique. This first consisted in calculating the mean of the two normalized areas detected for each constituent (SVO) of the neutral versions of the sentence (two values for each constituent). Then all the other normalized area values corresponding to the same constituent in the same sentence but uttered in a focused version were divided by this neutral mean. After this normalization, a value of 1 corresponds to no variation of the considered parameter compared to the neutral version, a value above 1 corresponds to an increase and a value below 1 corresponds to a decrease.

4. Results

For both articulatory and facial movement parameters, we analyzed the inter- and intra-utterance contrasts related to focus (*inter:* comparison of a constituent in its focused and neutral versions; *intra:* comparison of a focused constituent with the other constituents of the same utterance).

4.1. Articulatory and durational analysis

The results obtained after the measurements and the data reshaping are given in Fig. 2 for each parameter and each speaker and summarized below. All the results presented below are significant ($p<0.01$). The expression largest (resp. smallest) visible marking of focus corresponds to the largest (resp. smallest) value on the focused constituent (i.e, foc on Fig. 2).

Speaker B – focal lengthening (intra: +38.7% inter: +34.3%); focal hyper-articulation (except for lip spreading); post-focal hypo-articulation of all the parameters; largest visible marking for protrusion and duration.

Speaker C – focal lengthening (intra: +30.5% inter: +34.8%); focal hyper-articulation; slightly significant post-focal hypo-articulation for lip opening and jaw movements; pre-focal anticipation; largest visible marking for protrusion and duration.

Speaker D – focal lengthening (intra: +25.3% inter: +29.8%); focal hyper-articulation (except lip spreading); post-focal hypo-articulation only for lip opening and protrusion; pre-focal anticipation only for lip opening; largest visible marking for protrusion; smallest visible marking for lip spreading.

Speaker E – focal lengthening (intra: +16.8% inter: +23.9%); focal hyper-articulation; post-focal hypo-articulation; pre-focal anticipation only for protrusion; largest visible marking for protrusion; smallest visible marking for lip opening and spreading.

Speaker F – focal lengthening (intra: +43.8% inter: +49%); focal hyper-articulation; pre-focal anticipation only for protrusion; largest visible marking for protrusion; smallest visible marking for lip opening.

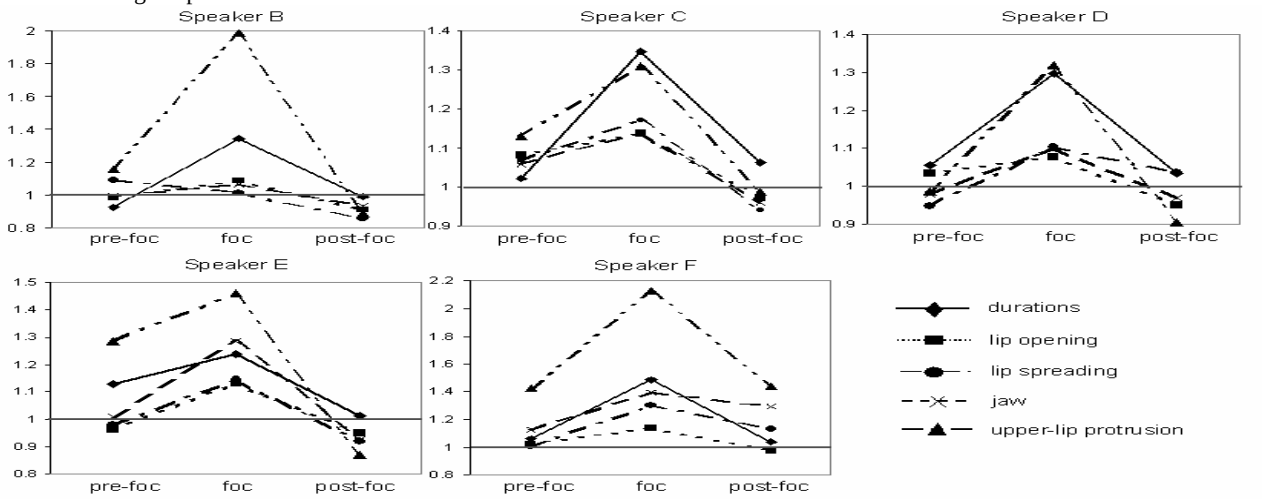

Figure 2: *Durational and articulatory measurements for all five speakers: normalized values corresponding to the pre-focal, focal and post-focal sequences (the dark horizontal lines correspond to the neutral case i.e. 1).*

4.2. Analysis of the other facial data

Eyebrow movements (raising) – There appears to be a link between eyebrow raising and the production of prosodic contrastive focus only for three out of the five speakers (B, C & E). However, this eyebrow raising is not systematic and does not occur whenever focus is produced. Speaker B is the one for which the combined productions are the most frequent. However, the amplitudes of the movements are very small (largest movement: 2mm). The other speakers either never raise their eyebrows, or do it on a random basis with no particular link to the production of focus.

Head movements – Speaker B is the only one for whom we can observe a correlation between head nods and focus production. This correlation is however not systematic and the amplitudes and temporal alignment of the movements are highly variable. The other speakers also move their heads but these movements seem to be produced randomly.

5. Discussion: modeling the production of visible correlates of prosodic contrastive focus in French

The production study described above along with that described in [9] have shown that there are potential visible articulatory correlates to the production of prosodic contrastive focus in French. One of the main conclusions that can be drawn is the fact that focus affects the whole utterance and not only the specific focused constituent. A number of visible articulatory gestures are indeed affected by focus and its position inside the utterance. The way and the extent to which these articulatory gestures are affected depend on the speaker. However, after having studied the productions of six different speakers, we have managed to extract two main strategies of the visual signaling of focus that satisfactorily represent all the productions.

Absolute visual signaling strategy: the focal constituent is lengthened and hyper-articulated to a large extent (inter-lip

area, protrusion and jaw movements). Previous studies ([9]) showed that the peak velocities were also increased which signals an increase of the underlying articulatory effort during the gestures [19]. The speakers using this strategy therefore concentrate their efforts on the hyper-articulation of the focal constituent. Some speakers also slightly anticipate focus. Fig. 3 illustrates this strategy.

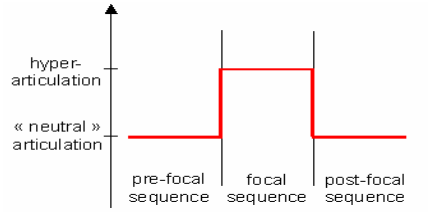

Figure 3: *Schematic representation of the absolute visual signaling strategy.*

Differential visual signaling strategy: in this case, the focal constituent is also lengthened and hyper-articulated but to a smaller extent. Focus is also sometimes anticipated. Additionally, the post-focal sequence is hypo-articulated compared to the neutral case. An important visible contrast is thus created inside the utterance: the focal hyper-articulation is not very distinct but is reinforced by the post-focal hypo-articulation. Fig. 4 illustrates this strategy.

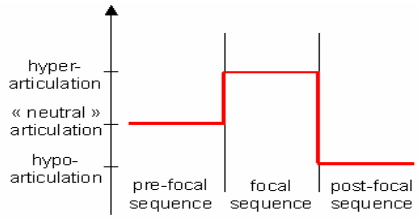

Figure 4: *Schematic representation of the differential visual signaling strategy.*

We observed that the visible articulatory parameter that was the most hyper-articulated under focus was protrusion.

We also found that there could be a link between prosodic contrastive focus and head (nod) and/or eyebrow (raising) movements. However this link is far from being systematic, particularly for the head movements. There are important inter- and intra-speaker variations concerning the realization of these movements, their amplitude or their synchronization with respect to the acoustic signal.

6. Conclusion

We found that the most hyper-articulated parameter under focus was protrusion. [20] have found that French protruded vowels were more intelligible visually than open vowels. Therefore it seems that the most visible parameter is also the most marked visually.

No consistent observation could be made as to the realization, the amplitude or the synchronization of eyebrow and head movements in correlation to the production of focus. However, only eyebrow raising and head nods were analyzed here. Other studies should be conducted to further investigate these potential links exploring other components of the eyebrow and head movements for example or other facial movements.

7. Acknowledgements

We thank Ishi-san and G. Vignali for their help with the recordings as well as the five patient speakers. We also thank J.-L. Schwartz for his comments on our work.

8. References

[1] Kelso, J.A.S.; Vatikiotis-Bateson, E.; Saltzman, E.; Kay, B.A., 1985. A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modelling. *JASA* 77(1), 266-280.

[2] Summers Van, W., 1987. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *JASA* 82(3), 847-863.

[3] De Jong, K., 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *JASA* 97(1), 491-504.

[4] Harrington, J.; Fletcher, J.; Roberts, C., 1995. Coarticulation and the accented/unaccented distinction: evidence from jaw movement data. *Journal of Phonetics* 23, 305-322.

[5] Erickson, D., 1998. Effects of Contrastive Emphasis on Jaw Opening. *Phonetica* 55, 147-169.

[6] Erickson, D.; Maekawa, K.; Hashi, M.; Dang, J., 2000. Some articulatory and acoustic changes associated with emphasis in spoken English. *ICSLP 2000*, vol. 3, 247-250.

[7] Cho, T., 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ , i/ in English. *JASA* 117(6), 3867-3878.

[8] Keating, P.; Baroni, M.; Mattys, S.; Scarborough, R.; Alwan, A.; Auer, E.T.; Bernstein, L.E., 2003. Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *ICPhS 2003*, 2071-2074.

[9] Dohen, M.; Lœvenbruck, H., 2005. Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study. Interspeech *2005*, 2413-2416.

[10] Audouy, M., 2000. *Traitement d'images vidéo pour la capture des mouvements labiaux*. Final engineering report, Institut National Polytechnique de Grenoble.

[11] Krahmer, E.; Ruttkay, Z.; Swerts, M.; Wesselink, W., 2002. Pitch, Eyebrows and the Perception of Focus. S*peech Prosody 2002*, 443-446.

[12] Granström, B.; House, D., 2005. Audiovisual representation of prosody in expressive speech communication. *Speech Communication* 46, 473-484.

[13] Hadar, U.; Steiner, T.J.; Grant, E.C.; Rose, F.C., 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech* 26, 117–129.

[14] Cerrato, L.; Skhiri, M., 2003. Analysis and measurement of communicative gestures in human dialogues. *AVSP 2003*, 251-256.

[15] Munhall, K.G.; Jones, J.A.; Callan, D.E.; Kuratate, T.; Vatikiotis-Bateson, E., 2004. Visual Prosody and Speech Intelligibility – Head Movement Improves Auditory Speech Perception. *Psychological Science*, 15(2), 133-137.

[16] Graf, H.P.; Cosatto, E.; Strom, V.; Huang, F.J., 2002. Visual Prosody: Facial Movements Accompanying Speech. *5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)*, 381-386.

[17] Cavé, C.; Guaïtella, I.; Bertrand, R.; Santi, S.; Harlay, F.; Espesser, R., 1996. About the Relationship between Eyebrow movements and F0 Variation. *ICSLP 96*, vol. 4, 2175-2179.

[18] Dohen, M.; Lœvenbruck, H., 2004. Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French. *ICSLP 2004*, 1313-1316.

[19] Nelson, W.L., 1983. Physical principles for economies of skilled movements. *Biological Cybernetics* 46(2), 135-147.

[20] Benoît, C.; Mohamadi, T.; Kandel, S., 1994. Effects of Phonetic Context on Audio-Visual Intelligibility of French. *JSHR* 37, p. 1195-1203.