Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study

Marion Dohen, Hélène Loevenbruck

Audiovisual Production and Perception of Contrastive Focus in French: a Multispeaker Study

Marion Dohen & Hélène Lœvenbruck

Institut de la Communication Parlée
UMR CNRS 5009, INPG, Univ. Stendhal, Grenoble, France
{dohen; loeven}@icp.inpg.fr

Abstract

This study examines the visual cues to prosodic contrastive focus in Hexagonal French and their role in visual speech perception. Two audiovisual corpora were recorded (from two male native speakers of French) consisting of sentences with a subject-verb-object (SVO) syntactic structure. Four conditions were studied: focus on each phrase (S,V,O) and broad focus. The corpora were first acoustically validated. Then lip area and jaw opening were extracted from the video. For each speaker, we identified a set of visible correlates of contrastive focus. The combined results showed that there were consistent visible articulatory correlates of contrastive focus across speakers: a) an increase in lip area and its first derivative on the focused item b) a lengthening of the focal syllables. There were also speaker-specific strategies in the amount of a) pre-focal anticipation or b) post-focal hypo-articulation.

Visual only perception tests were then conducted to see if the identified correlates were valid cues in perception. They showed that contrastive focus was well perceived visually for both speakers. The scores were better for the first speaker who displayed greater focal hyper-articulation. We also found that presence and salience of the visual cues enhances perception.

1. Introduction

Studies of French prosody have mainly focused on laryngeal and pulmonic correlates. A few supralaryngeal analyses exist, mostly considering tongue movements, e.g. [1], or spectral consequences of differences in articulation [2]. The few studies that have examined visual cues to prosody have focused on facial cues [3] such as eyebrow movements [4] or on head movements [5]. Only few studies have examined visible mouth correlates [6,7,8] and none have done so for French. "Visible" mouth correlates include articulatory correlates such as mouth opening and durational ones, such as lengthening. The purpose of this study is to relate tonal and visual characteristics of contrastive focus in French and to tell whether the visible correlates are used in perception.

Contrastive focus is used to emphasize a word or group of words in an utterance as opposed to another. In French, it can be either syntactic ("C'est xxx qui court." *It is xxx who runs.*) or prosodic ("XXX_F court." *XXX_F runs.*). This study deals with audiovisual prosodic contrastive focus in French.

2. Experimental material

2.1. The audiovisual data

Two audiovisual corpora were recorded, which consisted of sentences with a Subject-Verb-Object structure (SVO) and with CV syllables. Each sentence was likely to be produced as a single Intonational Phrase (IP) consisting of 3 Accentual Phrases (APs). In the broad focus condition, following Jun & Fougeron's findings [9,10], the expected default tonal pattern is {[LHiLH*]$_S$ [LHiLH*]$_V$ [LHiLL%]$_O$}. Sonorants were favoured in order to facilitate F0 tracking.

corpus 1: it consists of eight sentences:

s1.[Jean]$_{S1}$ [veut ménager]$_{V3}$ [nos jolis nouveaux navets]$_{O7}$.
'Jean wants to spare our fine new turnips.'
s2. [Romain]$_{S2}$ [ranima]$_{V3}$ [la jolie maman]$_{O5}$.
'Romain revived the good-looking mother.'
s3. [Mélanie]$_{S3}$ [vit]$_{V1}$ [les mauvais loups malheureux]$_{O7}$.
'Melanie saw the unhappy bad wolves.'
s4. [Véroniqua]$_{S3}$ [mangeait]$_{V2}$ [les mauvais melons]$_{O5}$.
'Veroniqua was eating the bad melons.'
s5. [Les mauvais loups]$_{S4}$ [mangeront]$_{V3}$ [Jean]$_{O1}$.
'The bad wolves will eat John.'
s6.[Mon mari]$_{S3}$ [veut ranimer]$_{V4}$ [Romain]$_{O2}$.
'My husband wants to revive Romain.'
s7.[Les loups]$_{S2}$ [suivaient]$_{V2}$ [Marilou]$_{O3}$.
'The wolves were following Marilou.'
s8.[Le beau marin]$_{S4}$ [vit]$_{V1}$ [Véronique]$_{O4}$.
'The good-looking sailor saw Veronica.'

corpus 2: it consists of thirteen sentences. The first four sentences correspond to s2, s4, s6 and s7 from corpus1.

s9. [La nounou]$_{S3}$ [mariera]$_{V3}$ [Li.]$_{O1}$
'The nurse will marry Li.'
s10. [Le lama lent]$_{S4}$ [lu]$_{V1}$ [Marinella.]$_{O4}$
'The slow lama read Marinella.'
s11. [Marinella]$_{S4}$ [va laminer]$_{V4}$ [Numu.]$_{O2}$
'Marinella will laminate Numu.'
s12. [Lou]$_{S1}$ [mima]$_{V2}$ [le lama.]$_{O3}$
'Lou mimed the lama.'
s13. [Le nominé]$_{S4}$ [lu]$_{V1}$ [les longs mots.]$_{O3}$
'The nominee read the long words.'
s14. [La nounou]$_{S3}$ [vit]$_{V1}$ [Lou.]$_{O1}$
'The nurse saw Lou.'
s15. [Les loups]$_{S2}$ [mimaient]$_{V2}$ [Marilou.]$_{O3}$
'The wolves mimed Marilou.'
s16. [Lou]$_{S1}$ [ramena]$_{V3}$ [Manu.]$_{O2}$
'Lou gave a lift back to Manu.'
s17. [Li]$_{S1}$ [ralluma]$_{V3}$ [les moulinets.]$_{O4}$
'Li lighted the wheels again.'

2.2. The audio-visual recording

Corpus 1 was recorded for speaker S1 (male) with front and profile cameras (see Figure 1) and was entirely analyzed. This led to an optimisation of the corpus and corpus 2 was thus recorded for speaker S2. For each corpus, four conditions were elicited: subject-, verb- and object- focus (narrow focus) and broad focus (neutral version). In order to trigger focus, the speakers had to perform a correction task by focusing a phrase which had been mispronounced in the prompt. 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 uttered**:* ROMAIN ranima la jolie maman.

The speakers were given no indication on how to produce focus (e.g. which syllables should be accented). For speaker S1, two speaking modes were recorded: real and reiterant speech. For speaker S2, only real speech was recorded. Reiterant speech was produced by replacing all the syllables with [ma]. The purpose of reiterant speech is to compare the acoustic and articulatory features across all the syllables.

The first step was to acoustically validate the corpora. It was checked, for both speakers, that the focused utterances displayed a typical focused intonation as described in [11].

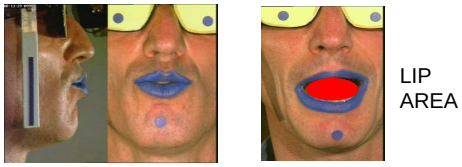

Figure 1: (left) Video signal recorded (profile and front views) and (right) symbolic representation of the lip area parameter.

2.3. Measurement techniques

Figure 1 shows an example of the images that were recorded. A program designed at Institut de la Communication Parlée (ICP) [12,13] enabled us to extract parameters describing lip shape and protrusion and jaw position from a sequence of digitalized frames. The mouth opening gesture was studied through a blue marker on the jaw (see Figure 1). The lip contour was automatically detected from the video signal and lip -height, -spreading, -area and -protrusion were derived.

3. Preliminary study: reiterant speech

Before studying real speech, a preliminary study was conducted on reiterant speech for speaker S1 [14]. The purpose was to determine a set of possible visible correlates to contrastive focus. These results showed that the **large jaw opening** gestures associated with **high opening velocities** on all the focused syllables and the **long lip closure for the first segment** of the focused group could be interpreted as a set of visual cues to the perception of focused reiterated [ma] sequences. Additional cues may be **prefocal lengthening** and **post-focal hypo-articulation**.

A visual only perception experiment showed that the visual cues described above are used for the perception of contrastive focus in French for reiterant speech.

4. Production studies

4.1. Preliminary analysis of the problem

4.1.1. Possible articulatory correlates

There are many possible visible articulatory correlates: jaw opening, lip -height, -area, -spreading, -protrusion, etc. The problem is to identify the one(s) which will vary the most significantly across conditions and the most invariantly across syllables. In our preliminary study [14] we found that the main articulatory consequence of contrastive focus is hyper-articulation. Hyper-articulation can be achieved in various ways, including increase in the amplitude of lip and/or jaw opening and closing movements, increase in lip spreading or narrowing. The parameter affected by hyper-articulation varies, depending on which syllable is uttered: for a hyper-articulated /a/ the mouth will be more opened thus the lip opening and the jaw opening will therefore be larger, for a hyper-articulated /i/, lip spreading, but not lip height will increase, and for a hyper-articulated /u/, lip protrusion will increase but not lip height nor spreading. The parameters which are most likely to be affected by hyper-articulation are thus lip height (LH), lip spreading (LS) and lip protrusion (LP). The lip area parameter (LA) takes into account the variations of both LH and LS. The articulatory parameters studied were thus LP, LA and LA's first derivative.

4.1.2. Possible durational correlates

The major durational correlates of focus identified in the study of reiterant speech were: focal lengthening, prefocal lengthening and what was called "lenghtening of initial lip closure" for the first [ma] in the focalized phrase. Similarly, in this study, focal and prefocal duration were measured, as well as the duration of the first phoneme of the focused sequence. This last parameter will thereafter be referred to as "first segment duration". In so doing , we wanted to find out if the lengthening of the initial lip closure measured for reiterant speech (see 3) was only an artifact of the syllable used or a general correlate of contrastive focus in French.

4.1.3. Measurements

All the maxima of LA and LP were detected. The duration of all the syllables were also computed. As explained before, the fact that real speech is studied here induces a great deal of variability. Even if the LA and LP parameters will account for most hyper-articulation strategies, there still remains a comparison issue. Lip area is indeed not comparable from an /a/ to an /i/ and the same can be put forward concerning duration. In order to compare data across the corpus, a "normalization" had to be performed. Therefore, the values of all the maxima (resp. of the corresponding syllable durations) for each parameter were divided by the mean value of the maxima of that parameter (resp. the mean value of the durations) for both broad focused utterances. All the broad focus maxima (resp. durations) therefore correspond to the value 1 and for the other focus types a value above 1 implies an increase of the considered parameter and a value below 1 a decrease of the considered parameter.

4.2. Results for corpus 1 (speaker S1)

4.2.1. Articulatory measurements

Figure 2 shows the grand mean of the "normalized" values of LA over each syntactic phrase and over all the identical syntactic phrases of the corpus. For example, the 1st column was computed in the following way: all the peaks of lip area were detected and "normalized" for the subject of the broad focused utterances, the means over each subject were then calculated and the means of these means were plotted.

The mean increase of LA from a broad focus condition to a focused condition is of 48.7% (significant: $p<0.05$). The mean increase of LA's first derivative is of 49.8% (significant: $p<0.05$). Concerning the pre-focal sequence, we found a mean increase of LA (resp. LA's 1st derivative) of 23.3% (resp. 13.3%). As for the post-focal sequence, we found a mean

decrease of LA (resp. LA's 1st derivative) of 2.2% (resp. 7.9%), this was not statistically significant. LP was not studied for this speaker because there were not enough protruded syllables in this corpus.

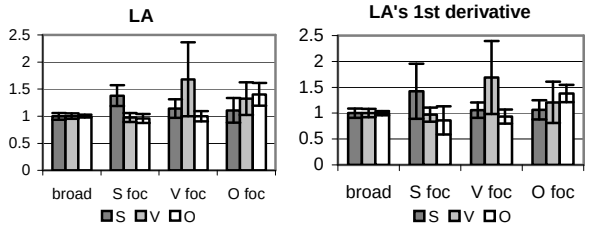

Figure 2: Grand mean of (left) the max of LA (cm²) and (right) the max of LA's 1st derivative over S, V & O.

4.2.2. Duration measurements

Focal lengthening: The mean duration of the focal syllables is significantly higher ($p<0.05$) than the duration of the same syllables in the broad focused condition. The mean lengthening from broad to narrow focus is of 33.6%.
Prefocal lengthening: The duration of the last syllable of a phrase was measured as significantly higher ($p<0.05$) when the following phrase was focused: +19.7%.
First segment lengthening: The first segment of a phrase is significantly lengthened by 53.2% ($p<0.05$) when the phrase it belongs to is focused. The first segment is therefore more lengthened than the rest of the focused phrase (only 33.6%).

4.2.3. Conclusion: S1's focus strategy

Considering the results presented above, we can summarize S1's visible correlates to contrastive focus:
- pre-focal anticipation: as had already been found and explained in the preliminary study [14], S1 displays an anticipation strategy i.e. he increases both duration and lip area just before focus.
- focal hyper-articulation: The focal syllables display a significantly larger lip area (and lip area's 1st derivative). These syllables are also significantly lengthened. We noted that both cues were used simultaneously: S1 does not either increase lip area or duration but both.
- post-focal hypo-articulation: Unlike what had been found in the preliminary study [14], S1 does not seem to hypo-articulate the post-focal sequence. The duration of the post focal syllables does not significantly change.

4.3. Results for corpus 2 (speaker S2)

4.3.1. Articulatory measurements

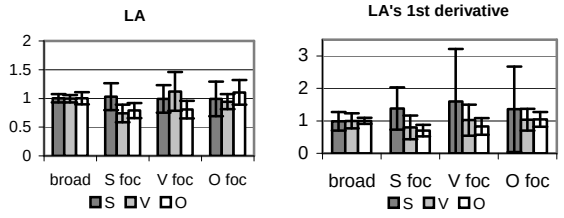

Figure 3: Grand mean of (left) the max of LA (cm²) and (right) the max of LA's 1st derivative over S, V & O.

Figure 3 corresponds to Figure 2 for speaker S2. The mean increase of LA from a broad focus condition to a focused condition is of 8.5% (significant: $p<0.05$). The mean increase of LA's first derivative is of 14.9% (significant: $p<0.05$). Concerning the pre focal sequence, we found a mean decrease of LA of 3.3% (not statistically significant) but an increase of LA's 1st derivative of 31.5%. As for the post focal sequence, we found a mean decrease of LA (resp. LA's 1st derivative) of 22.2% (resp. 22.3%). S2 also increases lip protrusion by 11.2% for the focused syllables.

4.3.2. Duration measurements

Focal lengthening: The mean lengthening of the syllables from the broad focus case to the focus case is of 20.4% (significant: $p<0.05$).
Prefocal lengthening: The duration of the syllable preceding a focused phrase did not significantly change when the following phrase was focused (decrease of 2%; not significant).
First segment lengthening: The mean lengthening of the first segment of a focused phrase is of 20.4%. This exactly corresponds to the focal lengthening (20.4%). For speaker S2, the first segment is thus not more lengthened than the rest of the focused phrase.

4.3.3. Conclusion: S2's focus strategy

Considering the results presented above, we can summarize S2's visible correlates to contrastive focus:
- pre-focal anticipation: the results show that S2 does not develop an anticipation strategy (no rise in lip area or duration just before focus).
- focal hyper-articulation: The focal syllables display a significantly larger lip area (and lip area's 1st derivative) and lip protrusion. These syllables are also significantly lengthened. We noted that S2 increases both LA and duration in only 40% of the cases and increases only one of the two parameters in 40% of the cases. In 20% of the cases he increases neither.
- post-focal hypo-articulation: the post-focal sequence displays an important decrease in lip area and its first derivative. The duration of the post-focal sequence however does not significantly change.

4.4. Comparison between S1 and S2

The rise for lip area and duration in S2 is not as important as for S1 (LA: S1: 48.7% S2: 8.5%; LA's 1st derivative: S1: 49.8% S2: 14.9%; duration: S1: 33.6% S2: 20.4%).

5. Perception studies

Visual only perception tests were conducted to check if the visible correlates identified above are used for perception.

5.1. Description of the experiments

The participants were told that they would be witnessing a conversation between two speakers. The first speaker would pronounce an utterance which they would first hear (audio prompt). They were told that one element (Subject, Verb or Object) in this sentence was misunderstood by the second speaker, who would therefore repeat the sentence as a question. This question would neither be heard nor seen by the

participants. The first speaker would then repeat the sentence and put focus on the misunderstood phrase. The participants saw a video recording of that speaker but heard no sound. Below is an example of how the test went:

Speaker 1 (audio only): Romain ranima la jolie maman.

Speaker 2 (nothing): Denis ranima la jolie maman ?

Speaker 1 (video only): ROMAIN ranima la jolie maman.

The participants were told that, in some cases, there was no misunderstanding (corresponding to a broad focus case). They were asked to determine which phrase (S, V, O or broad) had been misunderstood and thus focused. The participants used a highlighter pen to mark the constituent they perceived as focused on an answer sheet presented as below and highlighted the empty cell when they perceived broad focus.

Romain	ranima	la jolie maman.	

5.1.1. Test 1: speaker S1

We used four sentences from the corpus for their nearly balanced structures (almost the same number of syllables in S, V and O): s2, s4, s6 and s7. A total of 32 sentence pairs (1 pair: audio only unfocused utterance and visual only focused utterance) were available (4 sentences, 4 focus conditions, 2 repetitions). Five tests consisting of five random combinations of the 32 pairs were presented to each participant. These five tests were the same for all participants but the presentation order was different. Therefore, each person was presented with a total of 160 pairs of sentences. Both front and profile views were shown at the same time. A total of 33 native speakers of French participated in the experiment.

5.1.2. Test 2: speaker S2

We used nine sentences from the corpus for their nearly balanced structures: from s9 to s17. A total of 72 sentence pairs were available (9 sentences, 4 focus conditions, 2 repetitions). Two tests consisting of two random combinations of the 72 pairs were presented to each participant. The participants were tested on both views (front and profile) separately: some were presented with the first test front and the others with the first test profile and vice-versa for the second test. Therefore, each person was presented with a total of 144 pairs of sentences.

A total of 27 native speakers of French participated in the experiment.

5.2. Results

The results showed that the participants successfully perceived the focus through the visual modality alone. For test 1, the percentage of correct answers was of 71.45% and for test 2 it was of 43% (chance level for both tests: 25%). The scores for the test using speaker S1 were better, a finding which we expected since hyper-articulation was more salient for that speaker. However both scores are well above chance. A detailed analysis of the results showed that poorly perceived stimuli corresponded to unsalient visible correlates (articulatory and durational). This supports the hypothesis that the correlates perceived are those identified in the production studies. However, it was also found that the stimuli with the highest scores displayed all the correlates but none of them was either very unsalient nor very salient. This could mean that all the correlates are necessary to best identify focus even if they are not highly significant.

6. Discussion & Conclusions

The measurements and experiments suggest that there are lower face visual correlates of contrastive focus in French which can intervene in audiovisual speech perception. It is highly possible that more subtle facial correlates are also used in the visual perception of focus. Those could be head and/or eyebrow movements as suggested in [4,5,8]. We are currently analyzing Optotrak data to assess this issue.

7. Acknowledgements

We thank G. Rolland for designing and recording corpus 1 and C. Savariaux and A. Arnal for their technical help. We also thank M.-A. Cathiard, J.-L. Schwartz and P. Welby for their comments on our work.

8. References

[1] Lœvenbruck H., 1999. An Investigation of Articulatory Correlates of the Accentual Phrase in French. *Proceedings of ICPhS'99*. San Francisco, 1, 667-670.

[2] Tabain M., 2003. Effects of prosodic boundary on /aC/ sequences: articulatory results. *JASA* 113(5), 2834-2849.

[3] Burnham D., 2001. Visual discrimination of Cantonese tones by tonal but non-Cantonese speakers and by non-tonal language speakers. *AVSP'01*,. Denmark, 155-160.

[4] Granström B., House D. & Lundeberg M., 1999. Prosodic Cues in Multimodal Speech Perception. *Proceedings of ICSLP'99*. San Francisco, 1, 655-658.

[5] Munhall, K.G., Jones, J.-A., Callan, D., Kuratate, T. & Vatikiotis-Bateson, E., Visual prosody and speech intelligibility: Head movement improves auditory speech perception, *Psychological Science*, 15(2), 133-137, 2004.

[6] Bernstein L.E., Eberhardt S.P. & Demorest M.E., 1989. Single-channel vibrotactile supplements to visual perception of intonation and stress. *JASA* 85, 397-405.

[7] De Jong K., 1995. The supraglottal articulation of prominance in English: linguistic stress as localized hyper-articulation. *JASA* 97, 491-504.

[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. *Proceedings of 15th ICPhS*, 2071-2074.

[9] Jun S.-A., Fougeron C., A Phonological Model of French Intonation. In *Intonation: Analysis, modelling and technology*, A. Botinis (Ed.). Dordrecht: KAP, 209-242, 2000.

[10] Jun S.-A., Fougeron C., 2002. Realizations of Accentual Phrases in French Intonation. *Probus* 14, 147-172.

[11] Dohen M. & Lœvenbruck H.,. Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French. *ICSLP 2004*, Jeju Island (Korea), October 2004.

[12] Lallouache M.-T., 1991. *Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours de lèvres*. PhD Thesis, Institut National Polytechnique de Grenoble.

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

[14] Dohen M., Lœvenbruck H., Cathiard M.-A. & Schwartz J.-L., Visual perception of contrastive focus in reiterant French speech, *Speech Com* 44, 155-172, 2004.