Potential audiovisual correlates of contrastive focus in French

Marion Dohen, Hélène Lœvenbruck, Marie-Agnès Cathiard & Jean-Luc Schwartz

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

Abstract

The long-term purpose of this study is to determine whether there are "visual" cues to prosody. An audiovisual corpus was recorded from a male native French speaker. The sentences had a subject-verb-object (SVO) syntactic structure. Four conditions were studied: focus on each phrase (S,V,O) and no focus. Normal and reiterant modes were recorded. We first measured F0, duration and intensity to validate the corpus. The pitch maximum over the utterance was generally on a focused syllable and duration and intensity were higher for the focused syllables. Then lip aperture and jaw opening were extracted from the video. The jaw opening maximum generally fell on one of the focused syllables, but peak velocity was more consistently correlated with focus. Moreover, lip closure duration was longer for the first segment of the focused phrase. We can therefore assume that there are visual aspects in prosody that may be used in communication.

1. Introduction

1.1. Goal

Prosody is crucial in speech communication. It is involved in the extraction of information such as the sentence structure, the type of speech act, or the speaker's emotional state. Recent studies of French prosody have mainly focused on laryngeal and pulmonic correlates of prosody. A few supralaryngeal analyses have also been carried out, mostly considering tongue movements [1] or spectral consequences of differences in articulation [2]. Very few works have examined visual cues to prosody, namely articulatory features which could be seen and used by the person who is spoken to. The purpose of this study is to relate tonal and visual characteristics of contrastive focus, to find some aspects of a "visual" prosody.

1.2. Background

Many models of the prosodic structure of French have been proposed [3,4,5,6,7,8,9]. The phonological prosodic model of Jun & Fougeron [9,10] is used in the present study. This model agrees with most descriptions of French intonation and uses a transcription system consistent with the widely-used ToBI [11]. It features two prosodic units: the Accentual Phrase (AP) and the Intonational Phrase (IP).
The AP has also been called the "rhythmic group", "intonation group", "prosodic word" or "intonème mineur". It contains one or more content words and is right-demarcated by the primary stress (H*). An initial LH (Low-High) tonal sequence can mark the initial boundary of an AP. The LHi sequence has also been called initial or secondary accent. The default tonal pattern of the AP is thus /LHiLH*/. The IP has been called "intonème majeur" or "unité intonative". The IP level can preempt the AP level. If an AP is IP-final, the default H* will be replaced by the boundary tone of the IP (L% or H%).
In this model, contrastive focus is considered to be marked by a strong Hf and by a low plateau on the subsequent syllables. Hf most often replaces Hi, that is Hf is usually on the initial syllables of the phrase, but it can sometimes also replace both Hi and H* (i.e. the rise is carried by all the syllables in the phrase and culminates on the last syllable).

2. Experimental method

2.1. The corpus

Contrastive focus has clear effects on the pitch pattern. The aim is to test if focus may be perceived visually. The corpus was made of eight sentences with a Subject-Verb-Object syntactic structure (SVO) and with CV syllables (no clusters). When possible, we privileged sonorants in order to facilitate pitch tracking. Each sentence is considered to be a single IP consisting of 3 APs. The syllable number in the subject and verb APs varies from 1 to 4 and in the object APs from 1 to 7. From s1 to s4 the syllable number of the subject grows while that of the object decreases and from s5 to s8 it is the contrary. The first four sentences have a first name subject and the four last sentences have a first name object. Four sentences have a balanced structure (about the same number of syllables in each syntactic phrase). The sentences are the following (the number next to S/V/O is the number of syllables in the phrase):

s1.[Jean]$_{S1}$ [veut ménager]$_{V3}$ [nos jolis nouveaux navets]$_{O7}$.
John 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]$_{O5}$.
Veronica 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 followed Marilou.
s8.[Le beau marin]$_{S4}$ [vit]$_{V1}$ [Véroniqua]$_{O4}$.
The good-looking sailor saw Veronica.

2.2. The audio-visual recording

The corpus was recorded from a native French speaker (male), using front and profile cameras. The video was a 25Hz signal.
Four conditions were elicited: subject focus, verb focus, object focus and no focus at all. In order to trigger focus, the speaker listened to a prompt where the sentence to be pronounced was slightly modified. He then had to perform a correction task by focusing the phrase which had been mispronounced in the prompt. The speaker was given no indication on how to produce focus (e.g. which syllables should be accented).
Four speaking modes were recorded: normal, reiterant speech, whispered and reiterant whisper. Whispered modes were visually hyperarticulated, the task being to speak in order to be

understood but not heard. Reiterant speech consisted in replacing all the syllables with /ma/. Its purpose was to be able to compare the acoustic and articulatory features of all the syllables. 256 utterances were recorded (8 sentences for 4 focus conditions and 4 speaking modes, all recorded twice).

3. Tonal validation

For this study, only the reiterant normal mode was studied (64 utterances: 8 sentences, 2 repetitions, 4 focus conditions). 14 mispronounced utterances had to be discarded (all productions of s5, s1 4 conditions of the 1st repetition, s1 subject focus, 2nd repetition and s8 verb focus, 2nd repetition). 50 utterances were thus kept for analysis, including 37 focused ones.

Before looking at potential visual correlates of contrastive focus, we checked that the focused productions of the speaker featured the classical tonal patterns of French and hence could also potentially display significant articulatory effects. **The first test** aimed at checking if the pitch maximum over the whole utterance was on one of the focused syllables. When it was not, we carefully listened to the utterance. Due to declination, a focused object phrase (utterance-final) may actually display pitch peaks of equal (or even smaller) magnitude to those of the subject phrase (utterance-initial). Declination is however compensated for by listeners [12]. The **second test** checked that, for all the syllables of each phrase, the pitch was higher in the focused condition. In a **third test**, we examined whether the first syllable of the first content word of the focused phrase carried a Hf pitch accent, as described in [9]. Then the pitch (resp. intensity) maximum for each syllable was detected. We computed the mean of the maxima in each phrase of each utterance and the mean of the means over all the utterances (**tests 4 and 5** respectively).

Test 1 showed that the pitch maximum over the utterance was aligned with one of the focused syllables in 29 of the 37 focused utterances (78%). Careful listening showed that focus was carried by the correct phrase in 6 of the 8 outliers, and that the unexpected pitch patterns were due to declination. Therefore, only 2 utterances had to be discarded after test 1 (s1, 2nd repetition: object and verb focus).

Test 2 showed that, even though the pitch maximum over the whole utterance does not always coincide with a focused syllable (because of declination), pitch is always higher for a phrase when that phrase is focused.

Test 3 showed that the focused phrase in an utterance always carried a Hf. This accent was aligned with the first syllable of the first content word in the phrase. When the focused phrase was followed by one or more unfocused phrases, a low pitch plateau was observed after the focus. In addition, the phrase preceding the focused constituent was characterised by a low pitch plateau. The focus pattern was consistent with preceding observations [6,9,13,14]. However, when the object phrase was focused in s3, Hf was carried by 7 syllables. This resulted, for the first repetition of object focus s3, in the object phrase pronounced as two focused phrases. It was thus discarded.

Test 4 (Figure 1a) shows that, in average, the pitch maximum in a phrase was higher when it was focused. Taking declination into account, pitch was always higher on the focused phrase. Similar conclusions can be drawn from intensity (Figure 1b).

Overall, three utterances were discarded. For s1, the utterance with no focus was the only left and was thus of no use to this study. 46 utterances were kept for further analysis.

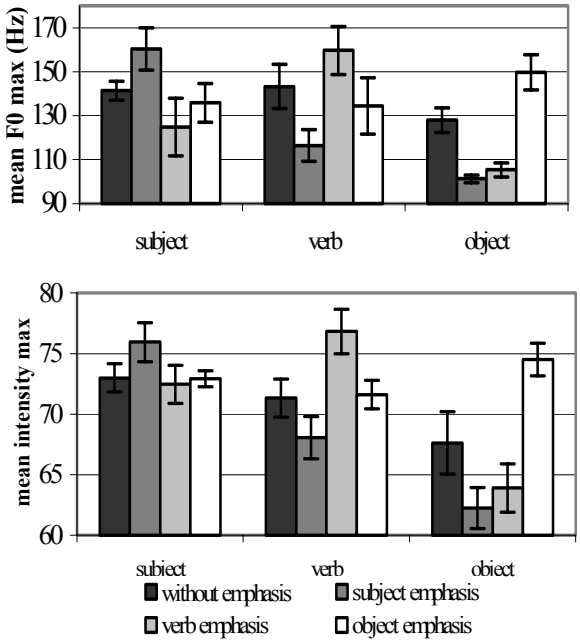

Figure 1: a) (Top) Mean value of the F0 max of each phrase (SVO) over all the reiterant utterances, in the 4 focus conditions. b) (Bottom) Mean value, over all the reiterant utterances, of the RMS max in each phrase.

4. Articulatory analysis

4.1. Experimental measurements

Figure 2: Video signal recorded: measurement method.

We extracted parameters describing lip shape and protrusion and jaw position from a sequence of digitalized frames (rate of 50 frames/s) using a program designed at ICP [15,16]. Providing no complementary data in this corpus of reiterated /ma/s, the profile view was not used. The mouth opening gesture was studied through the jaw and the closing gesture through the lip. Jaw opening was the difference between the ordinates of the jaw and the right-eye reference (see Figure 2). Two utterances had to be discarded because the articulatory data had not been correctly computed (s2, 1st repetition: object focus, and no focus). Therefore, 44 utterances were examined including 33 focused utterances.

4.1.1. Peak jaw opening

The prediction was that larger jaw opening should be observed on the focused syllables. The second panel in Figure 3 shows the jaw opening (the greater the value the larger the opening). The maxima (vertical bars) were automatically detected.

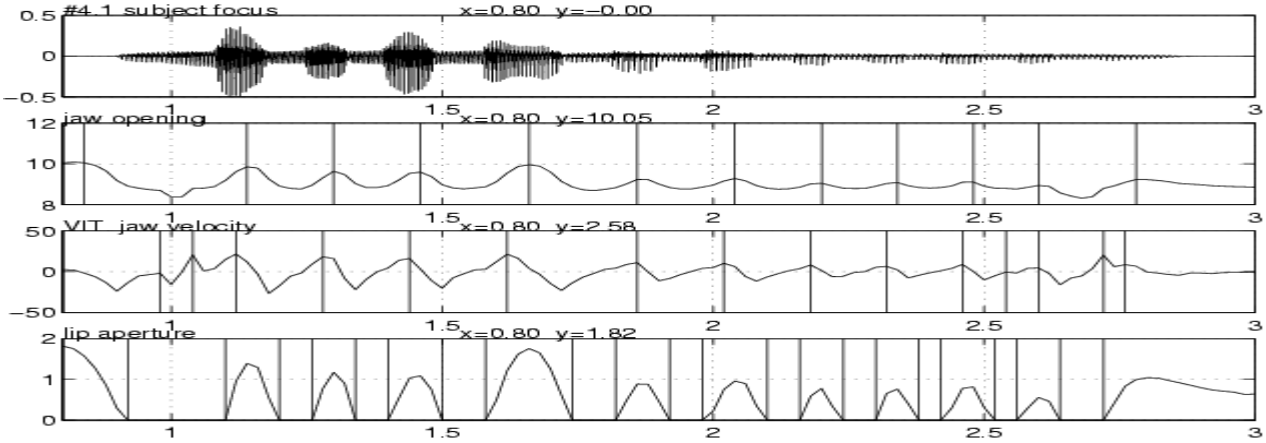

Figure 3: Traces of the acoustic signal, the jaw opening (cm), the jaw velocity and the lip aperture (cm) as a function of time (s).

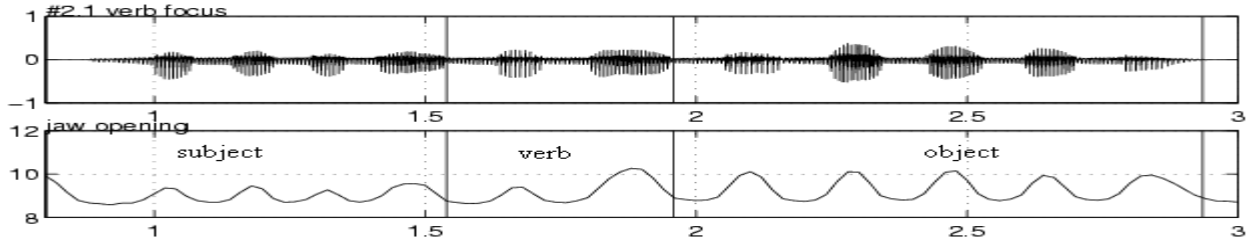

Figure 4: A case where the maximum of jaw opening is on the syllable just before the focused object phrase (syllable #6).

4.1.2. Peak jaw velocity

Higher peak jaw openings may not always characterize focused syllables. When speaking rate increases, independent of the degree of focus, the amount of time devoted to opening the jaw decreases, and the movement may be reduced. Peak velocity can be a better correlate of focus ([17, 18, 19]). Our prediction was that higher peak velocities should mark focused syllables. Velocity was the derivative of the jaw opening trace. The third panel in Figure 3 shows the jaw velocity trace (the greater the absolute value the fastest the opening or closing). The positive maxima, which correspond to the opening gesture (/m/ to /a/) and are referred to as peak opening velocity, were automatically detected and are labelled by the vertical bars.

4.1.3. Lip closure

A close examination of the video showed that there was a longer lip closure at the beginning of the focused phrase. The fourth panel in Figure 3 shows the lip aperture (zero plateaus: closed lips). Vertical bars mark the beginning and end of the closure plateaus. We studied the duration of the plateau for the first /m/ of each phrase, hereafter referred to as lip closure.

4.2. Results

4.2.1. Peak jaw opening

Location of the maximum of peak jaw opening
The maximum of the peaks of jaw opening over the whole utterance was generally on one of the focused syllables. In 8 cases over 33 however, the maximum was on the syllable just preceding the beginning of the focused phrase (Figure 4). In these cases, we noticed that the duration of the syllable just before the focused phrase was longer than the average syllable duration in the utterance. This suggests a possible anticipatory strategy, consisting in slowing down during the syllable preceding the focused phrase to prepare the focus. Therefore, it seems quite interesting to study velocity.

Mean maximum of peaks
The mean of the peaks on the syllables of each phrase was calculated for each sentence as well as the mean of the means over all the utterances (Figure 5). On average, mean peak jaw opening is always higher for the focused group but the results are not significant. This seems contradictory with the acoustic validation presented above. Peak jaw opening may thus not adequately represent focus.

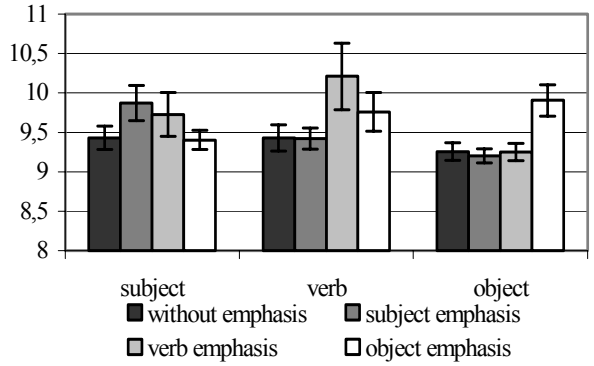

Figure 5: Mean peak jaw opening.

4.2.2. Peak opening velocity

Location of the maximum of peak opening velocity
We checked that the maximum of all the opening velocity peaks of the utterance was on one of the focused syllables. There were only 4 utterances for which it was not valid. Once more, in these cases, the maximum was on the syllable before

the beginning of the focused phrase. As for jaw opening, we averaged the peaks over each phrase and calculated the mean of the means over all the utterances (Figure 6).

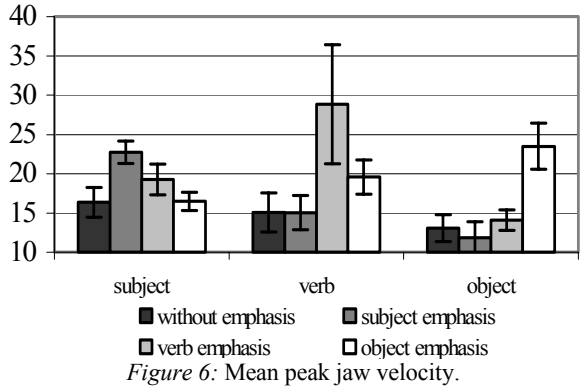

Figure 6: Mean peak jaw velocity.

These results show that, on average, peak opening velocity is significantly higher for the focused syllables and that standard deviations are acceptable, except for the focused verb phrase. The observed higher peak velocity is an artefact of the syllable number. In s3 & s8, the verb phrase was monosyllabic. A single syllable carried the focus and therefore the mean peak velocity on this syllable was higher than if there were many syllables (in a focused phrase, not all syllables show high peak velocity). The subject and object phrases did not include such monosyllabic exemplars (s1 & s5 had been discarded).

4.2.3. Lip closure

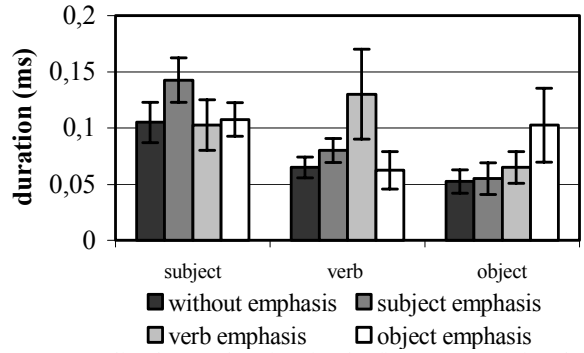

Figure 7: lip closure duration for the first segment of each phrase (SVO).

We calculated the mean of the duration of the lip closure over all the utterances. The results (Figure 7) show that lip closure duration is significantly higher for the focused phrase.

5. Conclusion

The corpus under analysis consisted of reiterant speech under 4 focus conditions. The 1st study showed that the speaker had pronounced the focused phrases with a classical intonation. The 2nd analysis aimed at examining whether there were also visual prosodic correlates. The analysis of the mouth opening gesture showed that, in general, the jaw was more opened for the focused syllables and the opening gesture was faster. There were a few exceptions for which the syllable preceding the focused phrase could show a larger opening. Our explanation for this is that the speaker seems to slow down just before focus allowing a larger opening gesture (more time is available for the jaw opening gesture). This interpretation is supported by the fact that jaw opening velocity was not always higher for these cases (a high jaw velocity is often correlated with hyper articulation [19]). The mouth closing gesture analysis showed that lip closure was longer for the 1st segment of the 1st syllable of the focused phrase.

As a general conclusion, we suggest that the large jaw opening gesture associated with a high opening velocity and the long lip closure could altogether be visual cues to the perception of focused reiterated /ma/ sequences. This will be further studied through perceptual experiments.

6. Acknowledgments

We thank G. Rolland for designing and recording the corpus and C. Savariaux for his technical help with the video data.

7. References

[1] Lœvenbruck, H. 'An investigation of articulatory correlates of the accentual phrase in French', *Int. Congr. Phon. Sc*., San Francisco, Vol. 1, 667-670, 1999.

[2] Tabain M. (in press). 'Effects of prosodic boundary on /aC/ sequences: Articulatory results'. *J. Acoust. Soc. Am*.

[3] Rossi M. 'L'intonation et l'organisation de l'énoncé'. *Phonetic, 42*, 135-153, 1985.

[4] Vaissière J. 'Langues, prosodies et syntaxe'. *A.T.A.L.A*, 38 (1), 53-82, 1997.

[5] Di Cristo A. 'Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français'. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 9-24, 1993.

[6] Di Cristo A. 'Intonation in French'. *Intonation systems: a survey of twenty languages*. Hirst D. & Di Cristo A. (eds.). CUP, 195-218, 1998.

[7] Mertens P. 'Intonational grouping , boundaries and syntactic structure in French'. *Proceedings of the ESCA Workshop on Prosody*, Lund, 41, 155-159, 1993.

[8] Post B. *A Phonological Analysis of French Intonation*. MA thesis, University oh Nijmegen, 1993.

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

[10] Jun S.-A. & Fougeron C., in press. *Realizations of Accentual Phrases in French Intonation*.

[11] Silverman KE. & al. 'ToBI: A standard for labelling English prosody'. *ICSLP 92*, 867-869, 1992.

[12] Liberman M. & Pierrehumbert J. 'Intonational invariance under changes in pitch range and length'. *Language sound to structure: studies in phonology presented to Morris Halle by his teacher and students.* Aronoff M. & Oehrle R. (eds.), MIT Press. 157-233, 1984.

[13] Touati P. 'Structures prosodiques du suédois et du français'. *Working Paper,* 21. Lund University Press, 1987.

[14] Clech-Darbon A., Rebuschi G. & Rialland A. 'Are there Cleft Sentences in French?'. *The Grammar of Focus,* Tuller L. & Rebuschi G. (eds), Amsterdam: Benjamins, 83-118, 1999.

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

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

[17] Kelso J.A.S & al. 'A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling'. *J. Acous. Soc. Am.*, 77 (1), 266-280, 1985.

[18] Beckman M. E., Edwards J. & Fletcher J. 'Prosodic structure and tempo in a sonority model of articulatory dynamics'. *Papers in Laboratory Phonology II, Gesture, segment, prosody. G. J. Docherty & D. R. Ladd* (eds.), Cambridge University Press. 68-86, 1992.

[19] De Jong K. 'The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation'. *J. Acoust. Soc. Am.*, 97 (1), pp 491-504, 1995.