Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Can we see Focus? A Visual Perception Study of Contrastive Focus in French

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

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

Abstract

The purpose of this study was to determine whether the visual modality is useful for the perception of prosody. An audio-visual corpus consisting of four focus conditions (subject, verb, object focus and broad focus) was recorded from a male native speaker of French. A preliminary production study showed that there are visible correlates of contrastive focus in French a) increase in lip area and jaw opening on the focused syllables b) lengthening of the prefocal syllable and the focal syllables (with a considerably higher lengthening for the first segment of the focused phrase). The present perceptual study showed that a) contrastive focus was well perceived visually; b) no practice was necessary and c) subject focus was slightly easier to identify than the other focus conditions. We also found that the presence and salience of the visual cues enhances perception.

10.21437/SpeechProsody.2004-17

1. Introduction

1.1. Prosody as multigestural and multimodal

Prosody is mainly conceived of as a set of glottal and subglottal patterns resulting in variable acoustic parameters such as F0, intensity and duration. Therefore, the perceptual studies on prosody mostly deal with the auditory modality. On the visual side, glottal and subglottal gestures *per se* are essentially invisible. However, it has already been put forward for other languages that contrastive focus can be perceived visually. [1] shows that eyebrow movements are visual cues to the perception of focus and [2] and [3] found visible mouth correlates to contrastive focus for English. Prosody is multigestural and should be conceived of as multimodal.

A number of possible jaw and lip correlates of prosodic patterns should have visible consequences. The only study of the visual perception of contrastive focus that we are aware of [4] concerns English. In it, it was found that contrastive focus in English has many facial correlates varying from one speaker to another. In addition, it was shown that visual only perception of focus gave results well above chance.

In this paper we suggest a description of the visible correlates of contrastive focus in French. A perceptual test was conducted to see whether focus is perceived visually.

1.2. Background

Jun & Fougeron's model [5,6] was used in the present study. It agrees with most descriptions of French intonation and uses a transcription system consistent with the widely used ToBI [7]. It features two hierarchical prosodic units. The lower is the Accentual Phrase (AP, right-demarcated by the primary stress (H*) and possibly left demarcated by an LHi (Low-High) tonal sequence called the initial or secondary accent). The default tonal pattern of the AP is /LHiLH*/. The higher prosodic unit is the Intonational Phrase (IP) which can preempt the AP level. E.g., if an AP is IP-final, H* is 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 (f for focus) and by a low plateau on the subsequent syllables. Hf most often replaces Hi, but it can also replace both Hi and H* (i.e. the rise in F0 is carried by all the syllables in the phrase and culminates on the last syllable).

2. Experimental method

2.1. The corpus

The corpus consisted of eight sentences with a Subject-Verb-Object syntactic structure (SVO) and with CV syllables. Each sentence was likely to be produced as a single IP consisting of 3 APs. When possible, we favoured sonorants in order to facilitate the F0 tracking. For examples see 4.2.

2.2. The audio-visual recording

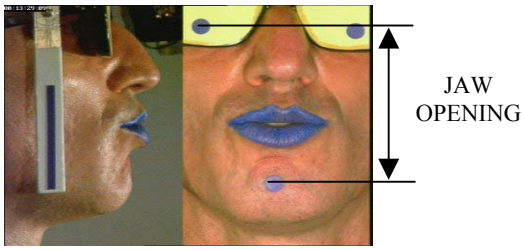

Figure 1: *Video signal recorded: measurement method.*

The corpus was recorded from a male native speaker of French with front and profile cameras (see Figure 1). Four conditions were elicited: subject-, verb- and object- contrastive focus (narrow focus) and broad focus. To elicit narrow focus, the speaker listened to a prompt in which either the subject, the verb or the object AP was incorrect. He then performed a correction task in which he contrasted the incorrect phrase in the prompt. The recording went as follows (capital letters signal focus):

Audio prompt: Denis ranima la jolie maman.
'Denis revived the good-looking mother.'
Speaker uttered*:* ROMAIN ranima la jolie maman.
'ROMAIN revived the good-looking mother.'

The speaker was given no indication on how to produce focus (e.g. which syllables should be accented). Four speaking modes were recorded: real, reiterant speech, whispered and reiterant whisper. So far, only two modes have been studied: real and reiterant. Reiterant speech was produced by replacing

all the syllables with [ma]. Its purpose is allow us to compare the acoustic and articulatory features across all syllables. 256 utterances were recorded (8 sentences, 4 focus conditions, 4 speaking modes, all recorded twice) and 128 have been studied.

2.3. Complementary production study

The acoustic and articulatory analyses of the corpus are described in another paper published in this conference [13].
The acoustic analysis showed that the recordings clearly contained tonal cues to focus structure consistent with previous observations [5,9,10,11]. The articulatory and temporal analysis enabled us to identify a set of **visible correlates** of contrastive focus in French:

- increase in lip area and jaw opening
- hypo-articulation of the post-focal sequence (significant reduction in lip area and jaw opening),
- lengthening of the prefocal syllables, the focal syllables (especially of the first segment of a focused phrase).

3. Preliminary perceptual study: reiterant speech

A preliminary study, described in [12], was conducted on the reiterant speech corpus. A perceptual experiment was carried out in which participants were presented with purely visual stimuli and had to identify the focus condition (S, V, O or broad). The results showed that the subjects successfully perceived the focus through the visual modality alone (86% of correct answers on average for a 25% chance level). Participants most often mismatched focus conditions with broad focus rather than the reverse. Subject focus was found to be significantly easier to detect than any other focus condition and when it was not detected it was most often mismatched with broad focus. This finding is supported by the fact that the verb and object following the focused subject are hypo-articulated: they could not be identified as focused, and thus if the focused subject was not identified, the most expected answer would be broad focus. Taken together, these results enabled us to assert that the visual modality is relevant for the perception of contrastive focus in reiterant speech in French. Moreover, it was noticed that for the stimuli with high error rates, the visible correlates were not "marked", i.e. not fully consistent with the description in 2.3. Those with low error rates corresponded to utterances for which the correlates were very clear and marked. Thus, it was assumed that there are visual cues to the perception of focus and that these cues may correspond to the correlates we measured.

4. Perception experiment: experimental method

4.1. Aim of the experiment

The preliminary study for reiterant speech [12] showed that there are visual cues to the perception of contrastive focus in French for reiterant speech. But is this also true for real speech? As it was explained before, we also observed visible correlates of focus for real speech [13]. But are these correlates perceived?

4.2. Description of the experiment

For this study, we used four sentences from the corpus for their nearly balanced structures (similar number of syllables in S, V and O). The sentences are the following:

(1) [Romain]$_{S2}$ [ranima]$_{V3}$ [la jolie maman]$_{O5}$.
'Romain revived the good-looking mother.'

(2) [Véroniqua]$_{S3}$ [mangeait]$_{V2}$ [les mauvais melons]$_{O5}$.
'Veronica was eating the bad melons.'

(3) [Mon mari]$_{S3}$ [veut ranimer]$_{V4}$ [Romain]$_{O2}$.
'My husband wants to revive Romain.'

(4) [Les loups]$_{S2}$ [suivaient]$_{V2}$ [Marilou]$_{O3}$.
'The wolves were following Marilou.'

The participants were told that they would be witnessing a conversation between two speakers. The first speaker would pronounce an utterance (one of the 4 sentences) 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 the front and profile views of this speaker (as in Figure 1) on a video monitor as he was uttering the repetition 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 subjects were told that, in some cases, there was no misunderstanding by the second speaker (i.e. broad focus). They were asked to determine which phrase (S, V, O or broad) had been misunderstood and thus focused. The subjects 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 no correction (broad focus).

Romain	ranima	la jolie maman.	

For the 1st sentence (audio prompt), we used the broad-focused utterances from the corpus and for the repetition we used all pronunciations of the sentences in the real mode.
The audio prompt was used in order for the participants to have an audio reference. A total of 32 sentence pairs (1 pair: audio only broad focused utterance and visual only 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. The tests were the same for all participants but the presentation order was different. Thus, each person was presented with a total of 160 pairs of sentences.

A total of 33 native speakers of French (8 males and 25 females) aged 18 to 52 participated in the experiment.

5. Results

5.1. General results

Figure 2 gives the percentage of correct answers (the focus condition was correctly identified) for each participant. Each bar corresponds to one subject. The average percentage of correct answers over all the 33 participants was 71.45%. Since this is much better than chance (25%), it can be assumed that the participants were sensitive to visual information on contrastive focus. These high scores were surprising since most participants found the test difficult. This suggests that the visual cues to prosody could be used in a non-explicit way. We also checked that the scores were independent of the order of the stimuli.

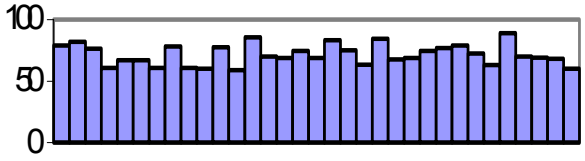

Figure 2: *Percentages of good answers for each participant*.

5.2. Influence of practice

The subjects all took part in the same five tests but in different orders. The purpose here was to examine the results of each trial not relative to the order of the stimuli it contains, but to its position in the experiment. This could give an indication of increased ability due to practice. Did the participants improve their performance throughout the experiment? The tests can be compared to one another since they are made of exactly the same stimuli presented in different orders and that this order has no significant influence on the results. A one-way analysis of variance (ANOVA) shows that the five means are not significantly different ($F(4,160) = 0.55$, $p = 0.7$). We can therefore assume that the features taken into account by the participants are not a matter of practice.

5.3. Differences between syntactic phrases

This statistical analysis aimed at examining whether the focus position (S, V or O) had an influence on the performance of the subjects. A one-way ANOVA shows that the results for the four focus cases are significantly different ($F(3,96)=11.3$, $p<0.01$). Multiple comparisons show that the average score for the subject focus condition is significantly higher than that for verb and object focus ($p < 0.001$). The results for the verb, object and broad focus conditions are not significantly different and the same conclusion can be drawn for the subject and broad focus conditions.

Thus, it can be assumed that the subject focus condition was easier to detect for the participants. Actually, it could have been expected that differences due to focus would be more marked if placed in the middle of the sentence and thus more easily detected (some participants even reported this). The articulatory data supports this expectation since peaks of lip area and jaw opening are greater in magnitude for the verb focus conditions. Therefore, it should be more difficult to identify a subject focus condition. However, as explained in 2.3, when a phrase is focused the subsequent phrases are hypo-articulated (reduced jaw opening and lip area). For the subject focus condition, this hypo-articulation is observed throughout the verb and object APs. The difference between the hyper-articulation of the subject and the hypo-articulation of the other APs is probably a strong cue to the location of focus is.

5.4. Analysis of the error trends

The purpose here is to examine the trends in which the participants mismatched a stimulus to a focus condition. Did they make a given mismatch more often than another one? Figure 3 shows the percentages of each type of matching.

stimulus \ answer	S	V	O	BROAD
S	**80.2**	3.9	0.2	15.8
V	7.4	**65.7**	0.8	26.1
O	0.6	25.4	**65**	9
BROAD	9	11.5	4.5	**74.9**

Figure 3: *Confusion matrix providing the percentages of each type of matching made by the subjects. E.g. 80.2% of the S focus stimuli were indeed identified as focused on the subject. (S, V, O: subject, verb and object focus, BROAD: broad focus).*

It seems that an important part of the mismatching was toward a broad focus condition (for 15.8% of the subjects, 26.1% of the verbs and 9% of the objects). Thus participants more likely gave a broad focus interpretation when there was in fact narrow focus than the reverse. These confusions of a narrow focus case with a broad focus case were predictable since the visible correlates of contrastive focus were more or less marked depending on the stimuli. Nevertheless, there were also confusions between the broad focus conditions and the subject (9%) and verb (11.5%) focus utterances.

Figure 3 also shows a high mismatching of object focus stimuli with verb focused utterances (25.4%). This is interesting since the same observation had been made for the perceptual tests conducted on reiterant speech (see [12] for details). We had explained this by the fact that, when a phrase was focused, focus was distributed over all the syllables of the phrase and thus the longer the focused phrase was the less significant was its hyper-articulation. In the corpus, the object phrases could be quite long compared to the other phrases ((1) and (2) have 5 syllable objects, (3) has a 2 syllable object and (4) has a 3 syllable object), thus the hyper-articulation of the longer phrases when they were focused was not as marked as for shorter phrases. This is supported by the articulatory data for the lip area and the jaw opening: the object is not as hyper-articulated when focused as are the other phrases. An explanation for the fact that hyper-articulation is distributed over all the syllables could be related to articulatory effort or timing. The strong increase in lip area, jaw opening and duration probably requires less effort for short phrases than long ones. However, this effort notion alone does not explain why the mismatching is mostly made toward a verb focus condition. In this corpus, when the number of syllables of the object is large, that of the verb is small. Thus, when the object is long, the focus pattern is uniformly slightly hyper-articulated whereas the syllable before the beginning of the object phrase carries the H* articulatory correlate of the short verb. This accented syllable can therefore appear as more marked than the subsequent syllables and the participant may identify a verb focus.

5.5. Further analysis of the results for each stimulus

In this section, the poorly and correctly identified stimuli were closely analyzed. Out of 24 narrow focused stimuli (among the 32 stimuli 8 were broad focused), 4 were *poorly perceived* (percentage of correct answers less than or close to 25%), 9

were *well perceived* (between 60% and 80%) and 11 were *very well perceived* (more than 80%). For the *poorly perceived* stimuli, the visible correlates measured were on the whole not very marked. In average, lip area (resp. jaw opening) was smaller than for the well perceived stimuli: 3.15 vs. 5.53 cm² (resp. 3.34 vs 5.76 cm). The prefocal duration was also shorter (193 vs. 221ms). Another interesting fact was noticed, namely that three *well* and five *very well perceived* stimuli had some non marked correlates. For most of these outliers even if one or several correlates were non marked, at least one of the other correlates was on the contrary very marked. The three *well perceived* outliers and two of the *very well perceived* outliers had non marked durational correlates but at least one very marked articulatory one. For one of the *very well perceived* outliers the articulatory correlates were not marked but the durational ones were very marked. Only for two outliers, similar conclusions could not be drawn. Moreover these outliers had the highest scores (98.2% and 98.8% of correct answers). Actually, for these stimuli none of the parameters were really low, they simply were not highly marked. This could mean that contrastive focus is easier to detect if all the correlates are present even if they are not particularly highly marked. It could also imply (as proposed in [4]) that other more subtle correlates are implied in the perception of contrastive focus (for example post-focal hypoarticulation was only qualitatively evaluated here and could give further information if quantified).

6. Discussion and Conclusion

After a preliminary study on reiterant speech, it was found that there were visual cues to contrastive focus in French and that they were very well perceived (86% of correct focus identification for a chance level of 25%). The present study aimed at checking if this was also true for real non-reiterant speech. The corpus under analysis consisted of real speech under 4 focus conditions (subject, verb, object contrastive focus and broad focus).

The visible correlates identified after a preliminary production study were the following:

- Increase in the lip area and the jaw opening when the phrase was focused;
- Lengthening of the prefocal syllable, the first segment of the focused phrase and the focal syllables;
- Post-focal hypo-articulation;

A perception experiment was conducted to test whether these visible articulatory and durational correlates were cues for the perception of contrastive focus. Participants were presented with purely visual stimuli and had to identify the focus condition. The results showed that the subjects successfully perceived focus through the visual modality alone at a level well above chance (71.45% of correct answers on average for a chance level of 25%). There was no effect of practice on the performances. Subject focus was found to be significantly easier to detect than any other focus condition and most often mismatched with a broad focus condition. The fact that poorly perceived stimuli corresponded to non marked visible correlates supports the fact that the correlates perceived must be those identified. However, it was also shown that some stimuli were well perceived even without highly marked visible correlates. The two stimuli that had the highest scores actually displayed all the correlates but only with average values. This could mean that all the correlates are necessary to best identify focus even if they are not highly significant. Moreover it is possible that other more subtle correlates are used in the visual perception of focus as suggested above and in [4]. This is important for future studies. Moreover [2,4] showed that articulatory strategies used to signal focus are speaker dependent. Our next studies will therefore consist of recording other speakers to see if we get similar production and perception results.

7. Acknowledgements

We thank Guillaume Rolland for designing and recording the corpus and Christophe Savariaux and Alain Arnal for their technical help with the video data. We thank Pauline Welby for her comments on this article. We are also thankful to the experimental participants to the test and those who helped us recruit them.

8. References

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

[2] De Jong K., 1995. The supraglottal articulation of prominence in English: linguistic stress as localized hyper-articulation. *Journal of the Acoustical Society of America* 97, 491-504.

[3] Bernstein L.E., Eberhardt S.P. & Demorest M.E., 1989. Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America* 85, 397-405.

[4] 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*. Barcelona, 2071-2074.

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

[6] Jun S.-A., Fougeron C., 2003. Realizations of Accentual Phrases in French Intonation. *Probus* 14, 147-172.

[7] Beckman M. E., Hirschberg J. & Shattuck-Hufnagel S.. The original ToBI system and the evolution of the ToBI framework. *Prosodic typology and transcription: a unified approach*. S.-A. Jun (ed.), Oxford University Press.

[8] Liberman M., Pierrehumbert J., 1984. Intonational invariance under changes in pitch range and length. In *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.

[9] Di Cristo A. 1998. Intonation in French. In *Intonation systems: a survey of twenty languages*, Hirst D. & Di Cristo A. (Eds.). Cambridge University Press, 195-218.

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

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

[12] Dohen M., Loevenbruck H., Cathiard M.A., Schwartz J.L., 2003. Audivisual Perception of Contrastive Focus in French. *Proceedings of the AVSP 2003 Conference*. St Jorioz, France, 245-250.

[13] Dohen M., Loevenbruck H., Cathiard M.A., Schwartz J.L., 2004. Identification of the Possible Visible Correlates of Contrastive Focus in French. *Proceedings of the SP2004 Conference*. In Press.