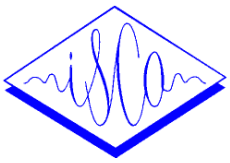

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Audiovisual Perception 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 purpose of this study is to determine whether the visual modality is useful for the perception of prosody. An audio-visual corpus was recorded from a male native French speaker. The sentences had a subject-verb-object (SVO) syntactic structure. Four contrastive focus conditions were studied: focus on each phrase (S, V or O) and no focus. Normal and reiterant modes were recorded. We first measured fundamental frequency (F0), duration and intensity to validate the corpus. Then, lip aperture and jaw opening were extracted from the video data. The articulatory analysis enabled us to suggest a set of possible visual cues to focus. These cues are a) large jaw opening gestures and high opening velocities on all the syllables of the focused phrase; b) long initial lip closure and c) hypo-articulation (reduced jaw opening and duration) of the following phrases. A perception test to see if subjects could perceive focus through the visual modality alone was developed. It showed that a) contrastive focus was well perceived visually for reiterant speech; b) no training 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.

1. Introduction

1.1. Prosody as multigestural and multimodal

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. It 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 [1-6]. On the visual side, glottal and subglottal gestures *per se* are essentially invisible, although they might have facial movement correlates [7]. The recent data by Burnham on the visual perception of tonal contrasts is a spectacular illustration of the fact that F0 could be directly perceived by the eye [8]. In [9], it is also showed that eyebrow movements are visual cues to the perception of focus.

Actually, prosody is multigestural and involves subglottal, glottal and supraglottal correlates. It displays a rich set of facial movements. Therefore, it should, without any doubt, be conceived of as multimodal.

Although most studies of French prosody have focused on laryngeal and pulmonic correlates, a few supralaryngeal studies have been carried out (e.g. [10-12]). A number of possible jaw, tongue and lip correlates of prosodic patterns were suggested. These should have visible consequences.

In this paper we give a description of the relationship between tonal and visual characteristics of contrastive focus in French. We then present a perceptive test to see whether focus is perceived visually. The purpose is to examine the visual prosodic cues to the perception of contrastive focus.

1.2. Background

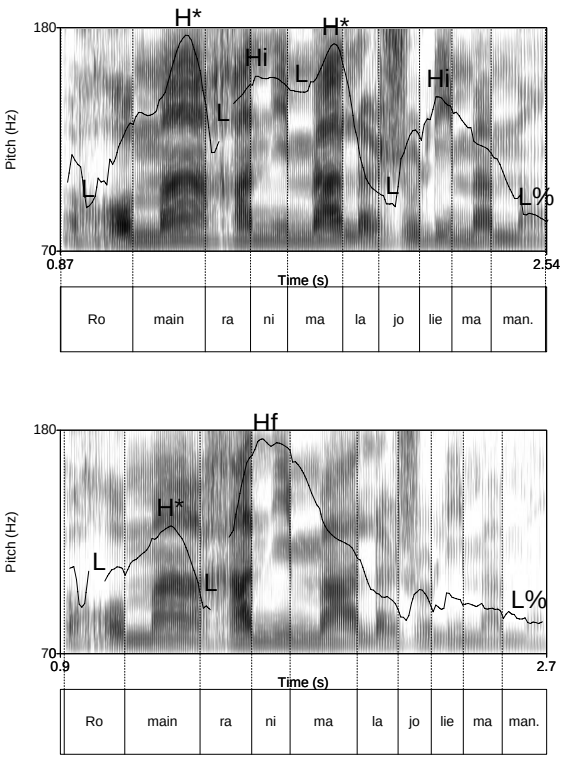

Figure1: spectrogram and F0 trace for an IP including 3 APs. a. (*top*) unfocused case. b. (*bottom*) focus on the verb AP. The utterance was $\{[\text{Romain}]_{AP}[\text{ranima}]_{AP}[\text{la jolie maman}]_{AP}\}_{IP}$ (Romain revived the pretty mother.).

Many phonological models of the prosodic structure of French have been proposed [1,13-18]. Jun & Fougeron's model [18, 19] 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 [20]. It features two hierarchical prosodic units: the lowest is the Accentual Phrase (AP) and the highest is the Intonational Phrase (IP).

The AP contains one or more content words and is right-demarcated by the primary stress (H*). An initial LHi (Low-High) tonal sequence, also called the initial or secondary accent, can mark the initial boundary of an AP. The default tonal pattern of the AP is /LHiLH*/ as realized on the second AP of Figure1a). The IP level 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%) as shown in the last AP of Figure 1a).

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 (Figure 1b), 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 was made 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. In the no-focus condition, the default tonal pattern is thus supposed to be $\{[LHiLH^*]_S\ [LHiLH^*]_V\ [LHiLL\%]_O\}$. When possible, we favoured sonorants in order to facilitate the F0 tracking. For examples of the corpus items see 2.4.2.

2.2. The audio-visual recording

The corpus was recorded from a male native speaker of French with front and profile cameras. The video was a 25Hz signal. Four conditions were elicited: subject-, verb- and object- focus and no focus. In order to trigger focus, the speaker listened to a prompt in which 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. Reiterant speech was produced by replacing all the syllables with [ma]. The purpose of reiterant speech is to be able to compare the acoustic and articulatory features across all the syllables. Whispered speech elicits visible hyper-articulation, the task being to speak in order to be understood by one person but not heard by others. 256 utterances were recorded (8 sentences, 4 focus conditions, 4 speaking modes, all were recorded twice).

2.3. Tonal and articulatory analyses

The following study was made only for reiterant speech since the different articulatory features could easily be compared from one syllable to another (always the same syllable: [ma]).

2.3.1. Tonal validation

This preliminary study aimed at confirming that the speaker had pronounced the focused phrases with a typical focused intonation. For each production, it was therefore checked that the F0 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 F0 peaks of equal (or even smaller) magnitude to those of the subject phrase (utterance-initial). Declination is however known to be compensated for by listeners [21]. It was also checked that F0 was higher on the focused syllables. We verified that the first content word syllable of the focused phrase carried a Hf accent, as described in [18]. The F0 (resp. intensity) maximum for each syllable was detected. We computed the mean of the means over all the utterances. This showed that, in average, the F0 maximum in a phrase was higher when it was focused. Taking declination into account, F0 was always higher on the focused phrase. Similar conclusions were drawn from intensity. The items therefore clearly contained cues to focus structure consistent with previous observations [15,18,22,23].

2.3.2. Articulatory analysis

Figure 2 shows an example of the images that were recorded. We extracted parameters describing lip shape and protrusion and jaw position from a sequence of digitalized frames (with a rate of 50 frames/s, since one video image consists of two interleaved frames) using a program designed at ICP (Institut de la Communication Parlée) [24, 25]. The mouth opening gesture was studied through the jaw (see Figure 2) and the closing gesture through the lips. Lip heigth was defined as the distance between upper and lower lips. The internal lip contour was detected from the video signal.

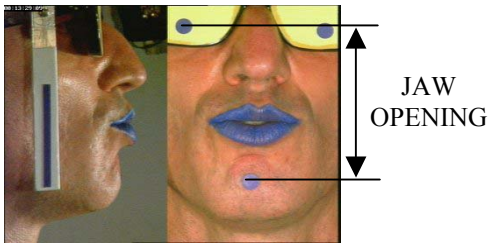

Figure 2: Video signal recorded: measurement method.

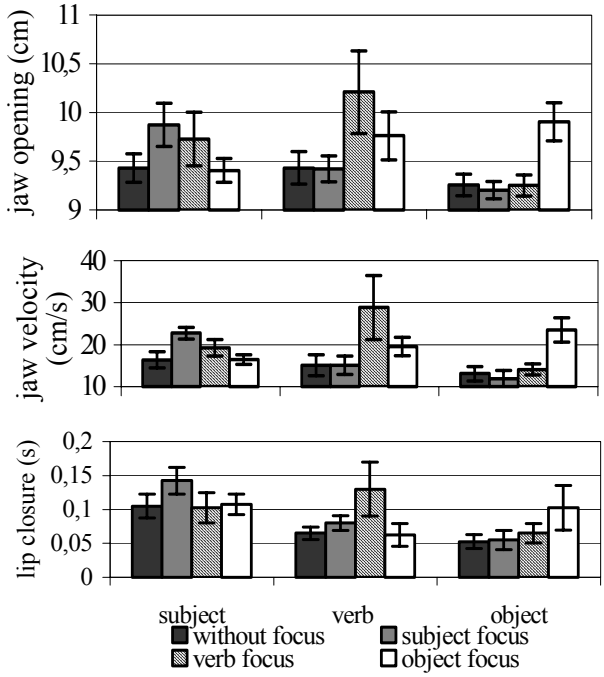

Figure 3: *a. (top)* mean peak jaw opening over all the syllables of the AP (cm) *b. (middle)* mean peak jaw velocity (cm/s) *c. (bottom)* lip closure duration (s) for the first segment of each phrase (SVO).

Jaw opening gesture

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 (see Figures 3.a & 3.b).

There were a few exceptions in which the syllable preceding the focused phrase showed a larger opening. We suggest two explanations for this. The first one is that the syllable just before the focused phrase is the last syllable of an AP and therefore probably bears some of the articulatory correlates of primary stress (H*). It has been suggested [11] that those articulatory correlates include a large articulatory (jaw or tongue) movement, a high peak velocity and a long duration. In the following, these three characteristics will be referred to as "**hyper-articulation**". The second explanation is that the speaker simply slows down just before focus thereby allowing a larger opening gesture (more time is available for the jaw

opening gesture). This would correspond to an anticipatory strategy in order to prepare to focus. This interpretation, rather than the first one, is supported by the fact that jaw opening velocity was not always higher for these cases (a high jaw velocity is usually correlated with hyper-articulation [26]). The anticipatory hypothesis is further grounded by the fact that this large jaw opening on the last syllable of an AP was more clearly observed on the prefocal AP. For example, when the focus is on the object, we observe the large jaw opening correlate on the last syllables of both the subject and verb APs. However, the jaw opening is much larger for the verb AP (prefocal AP).

Lip closure gesture

The mouth closing gesture analysis showed that lip closure was longer for the first segment of the first syllable of the focused phrase (see Figure 3.c.).

Post focal hypo-articulation

Overall, we noticed for all the focus conditions and all the utterances, that when a phrase is focused, the subsequent phrases are hypo-articulated; in other words, the jaw opening gesture is reduced and uniform, and the duration of the syllables is smaller (see the object AP in Figure 4 which displays a verb focus utterance).

Summary of the articulatory correlates of focus

These results allow us to sketch a model of the articulatory correlates of contrastive focus for this speaker. Figure 4 illustrates this speaker's strategies. We suggest 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 could be interpreted as a set of visual cues to the perception of focused reiterated [ma] sequences. Additional cues may be that the syllable preceding the focused AP is always of high jaw opening magnitude and duration and that the post-focal phrases are hypo-articulated.

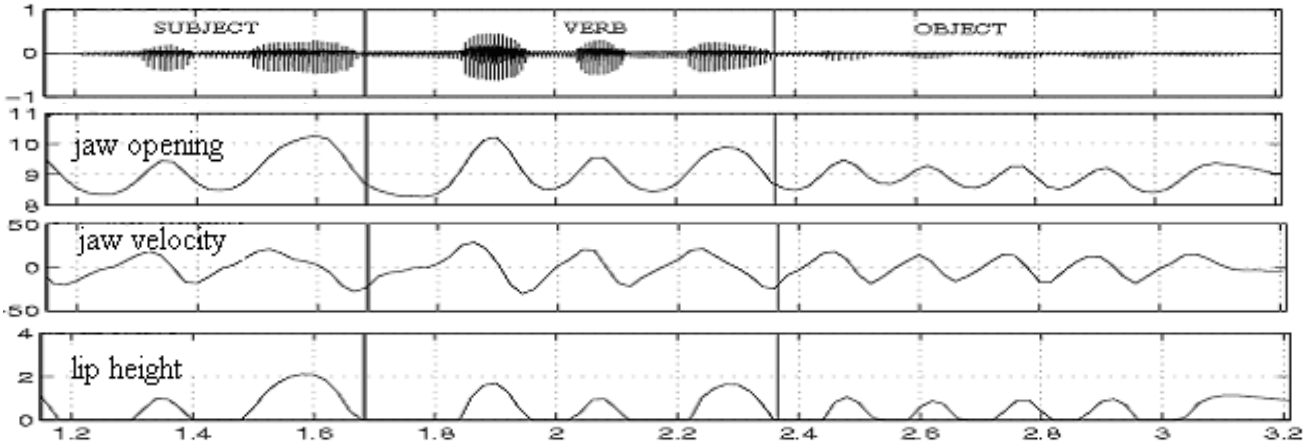

Figure 4: Traces of the acoustic signal the jaw opening (cm) and velocity (cm/s) and the lip height (cm) as a function of time (s). The utterance pronounced was [mama] [MAMAMA] [mamamamama].

2.4. Perceptual experiment

2.4.1. Aim of the experiment

The purpose of this experiment was to find out whether there are visual cues to the perception of contrastive focus in French. As it was explained before, we observed articulatory visual correlates of focus but the aim here is to know if they are relevant for the perception of prosody.

2.4.2. Description of the experiment

For this study, we used four sentences from the corpus. The sentences are the following (the number next to S/V/O is the number of syllables in the phrase):

(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 followed Marilou.

The participants were first given an audio prompt. They heard one of the four sentences in a normal non-reiterant mode. They were told that one syntactic phrase (Subject, Verb or Object) in this sentence was erroneous. Therefore, they would have to observe another speaker correcting the sentence by replacing the erroneous phrase with the correct one. By doing so, the speaker would put focus on the phrase considered. The erroneous phrase always had the same number of syllables as the correct one. The correction was made in a reiterant normal speaking mode and the participants saw the front and profile views of the speaker (as in Figure 2) on a video monitor but heard no sound. Below is an example of how the test went:

Participant hears: Romain ranima la jolie maman.
Correct sentence is: Denis ranima la jolie maman.
Participant sees: MAMA mamama ma mama mama.

The subjects were told that, in some cases, there was no correction by the second speaker thus no focus at all. They were asked to determine which phrase (S, V, O or none) had been corrected. They were given an answer sheet with the sentences presented in the following fashion:

[Romain] [ranima] [la jolie maman].

The subjects used a highlighter pen to mark the constituent they perceived as focused and highlighted nothing when they perceived no focus. For the erroneous sentence, we used the normal non-reiterant non-focused utterance from the corpus described in 2.1 and for the correction we used all pronunciations of the sentence in the reiterant normal mode. The speaker of the erroneous sentence was therefore the same as that of the corrected sentence (corpus uttered by only one speaker). Listeners were however unaware of this since the

first sentence was presented in the auditory mode and the corrected sentence, in the visual mode.
The non-reiterant speech condition was chosen for the erroneous sentence so that participants would treat the task as linguistic and not only as melodical or rhythmical. The first pronunciation was presented only auditorily to avoid the subjects from performing a mapping of the two utterances (same speaker) which could have altered natural perception.
A total of 32 sentence pairs (1 pair: audio erroneous sentence and visual corrected sentence) 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.

2.4.3. The participants

A total of 28 native French speakers (11 males and 17 females) aged 18 to 55 initially participated in the experiment. They came from various regions of France. Once each participant was over with the test, he/she was asked what they had done. Three participants (1 male and 2 females) had misunderstood the task and their results were thus discarded.

3. Results

3.1. General results

Figure 5 gives the percentage of correct answers (the focus or no focus condition was correctly identified) for each participant. Each bar corresponds to one subject. The average percentage of correct answers over all the 25 participants was 86%. Since this is much better than chance (25%), it can be assumed that the participants were sensitive to visual information on contrastive focus. It was surprising to obtain such high scores since the participants all found the test difficult. This suggests that the visual cues to prosody must be used in a non-explicit way. We also checked that the scores were independent of the order of the stimuli.

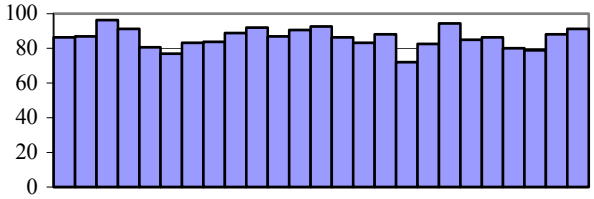

Figure 5: Percentages of good answers for each participant.

3.2. Influence of training

The subjects all took part in the same five tests but in different orders. The purpose here is 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 owing to training. 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. Figure 6 shows the average percentage of correct answers over all the participants for each trial. The five means are not considerably different. However, a one-way analysis of variance (ANOVA) shows that the hypothesis of equality of the 5 means can be rejected ($F(4,120) = 4.11$, $p < 0.01$). Pairwise comparisons (separate t-tests) show that the mean percentage of correct answers of the first trial was significantly smaller than that of the fourth and fifth trials ($p < 0.001$). The results for the 2nd, 3rd, 4th and 5th trials are not significantly different ($p = 0.04$).
Although participants' performance improved after the first trial, the scores, even for the first test, were well above chance (over 80%). After the second trial, there was no significant improvement. This may reflect the fact that the participants needed a first trial to get familiar with the task since they found it difficult at first. We can assume that the features taken into account by the participants must be perceived without explicit knowledge and are not a matter of training.

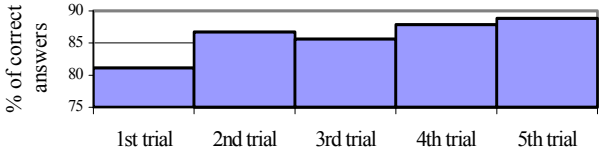

Figure 6: Average percentage of correct answers for each trial in its order, over all the participants.

3.3. Differences between the syntactic phrases

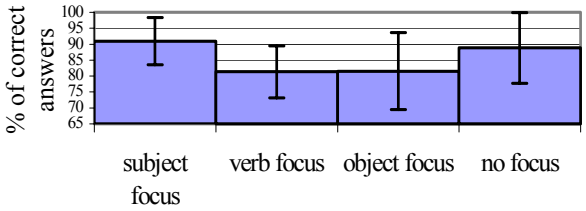

Figure 7: Average percentage of correct answers for each focus condition.

This statistical analysis aimed at examining whether the focus position (S, V or O) had an influence on the performance of the subjects. Figure 7 shows the average percentage of correct answers for subject, verb, object and no focus conditions. A one-way ANOVA shows that the results for the four focus cases are significantly different ($F(3,96) = 6.46$, $p < 0.01$). Separate t-tests for each pair of focus conditions show that the average score for subject focus is significantly higher than that for verb and object focus ($p < 0.001$). The results for the verb and object focus conditions are not significantly different ($p = 0.9455$) and the same conclusion can de drawn for the subject and no focus conditions ($p = 0.446$).
Thus, it can be assumed that the subject focus and no focus conditions were easier to detect for the participants. Actually, it could have been expected that differences due to focus would be more salient if placed in the middle of the sentence and thus more easily detected. The articulatory data supports this expectation since peaks of jaw opening and jaw opening velocity are greater in magnitude for the verb and object 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 duration). 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 where focus is.

3.4. 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 8 shows the percentages of each type of matching.
It can be noticed that the subject focus stimulus is most commonly confused with a no focus stimulus (6.9% > 2.1% and 6.9% > 0.1%). 16% of the verb focus stimuli were mismatched with a no focus case, and 8.2% of the object focus stimuli were also associated to a no focus case. It therefore seems that a great part of the mismatching was toward a no focus condition. Participants more likely gave a no focus interpretation when there was in fact a focus than a focus interpretation when there was no focus. Nevertheless, there were also confusions between the no focus conditions and the subject focus utterances (7.8%). This was predictable since, as already mentioned, the articulatory data showed that the differences between the features corresponding to focus cases and no focus ones were the least important for the subject focus case. Therefore, a subject focus utterance is more easily confusable with a no focus case than the other focus cases. Another element of Figure 8 is interesting, namely the mismatching of the object focus stimulus with a verb focus utterance (10%). This can be explained by the fact that the object phrases in the corpus could be quite long ((1) and (2) have 5 syllable objects, (3) has a 2 syllable object and (4) has a 3 syllable object). We observed that focus is distributed over all the syllables of the object phrase. We recall that no indication as to how to produce focus was given to the speaker. We found that when the object is long, the focus pattern is quite uniform and not as hyper-articulated (the peaks of jaw opening and velocity are not much greater than the mean values over the whole utterance) as when the number of syllables is lower. An explanation for this could be related to articulatory effort or timing. It is probably easier to strongly increase jaw opening and duration in short phrases. In addition, 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, as explained above, 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 salient than the subsequent syllables and the participant may identify a verb focus.

Answer / Stimuli	S	V	O	NO
S	**90.9**	2.1	0.1	6.9
V	2.2	**81.3**	0.5	16
O	0.2	10.1	**81.5**	8.2
NO	7.8	2	1.4	**88.8**

Figure 8: Confusion matrix providing the percentages of each type of matching made by the participants. E.g. 90.9% of the subject focus stimuli were identified as focused on the subject. (S, V, O: subject, verb and object focus, NO: no focus).

3.5. Further analysis of the results for each stimulus

The results of the perceptual experiment show that there are visual cues for the perception of focus. However, the results do not prove that these cues are the specific articulatory correlates described in the articulatory analysis (see 2.3). The purpose here is to analyse the answers given for each of the 32 stimuli and to relate, if possible, the quality of the perception to the presence/salience of the articulatory correlates of 2.3. The percentages of correct answers were calculated for each stimulus. It appears that 29 stimuli corresponded to scores over 82% and only three stimuli corresponded to bad scores (around 35%). Eight stimuli corresponded to very high scores (over 95%). The articulatory visual correlates of these very well and poorly perceived stimuli were closely analysed. The purpose was to check whether the visual cues proposed in 2.3 were lacking in the three poorly perceived stimuli and particularly salient in the eight very well perceived ones.
Concerning the poorly perceived stimuli, the articulatory data showed that the prosodic correlates of 2.3 were not optimally produced. Average jaw opening magnitude over the focused phrase was lower than for the very well perceived stimuli. In addition, lip closure duration of the first segment was short.

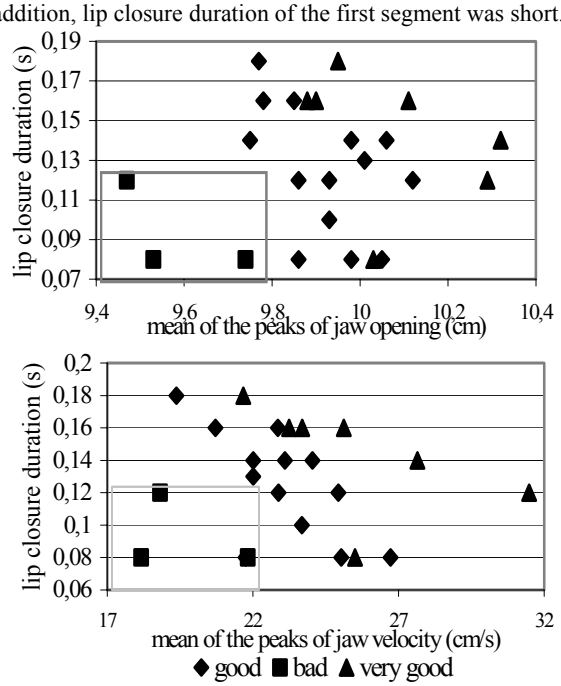

Figure 9: lip closure duration (s) of the first segment of the focused phrase as a function of the mean jaw opening over the focused phrase (cm) *(top)* and mean jaw opening velocity (cm/s) *(bottom)*. The gray rectangles contain the poorly perceived stimuli.

A global analysis was made for the 8 very well perceived stimuli. They all showed a very typical articulatory pattern of focus. The mean jaw opening over the focused phrase was higher than elsewhere in the utterance. Moreover, the maximum of jaw opening and velocity was always on a focused syllable and was of high magnitude compared to that of the other syllables. The lip closure duration of the first segment of the focused phrase was quite long. All the visual cues found after our articulatory analysis (see 2.3) were thus very marked for these utterances.
Figure 9 displays the lip closure duration of the first segment of the focused phrase as a function of the mean of the peaks of jaw opening and of the mean of the peaks of jaw velocity over the focused phrase. It shows that all the poorly perceived stimuli correspond to a lip closure duration less than 0.12 s, a mean jaw opening less than 9.8 cm and a mean of the peaks of jaw velocity less than 22 cm/s. If only one of the visual cues is lacking (e.g. small lip closure duration but large mean jaw opening and velocity) i.e. if the stimulus remains outside the gray rectangles (Figure 9), perception is still good.

We can conclude that the error rate was high when the correlates were not very clear and on the contrary very low when the visual articulatory correlates were clearly marked. This enables us to assume that the visual cues to perception of contrastive focus are probably the correlates described in 2.3.

4. Conclusion

The purpose of this study was to show that the visual prosodic cues identified through an articulatory analysis were actually perceived and helped listeners collect information on contrastive focus. The corpus under analysis consisted of reiterant speech under 4 focus conditions (subject, verb, object focus and no focus).

The first part of the study aimed at checking that the expected tonal correlates of focus were present in the corpus and at examining whether there were also visual prosodic correlates. The results showed that the speaker had pronounced the focused phrases with a typical intonation pattern. Moreover, a large jaw opening associated with a high opening velocity and a long lip closure appeared to be possible visual cues to the perception of contrastive focus on reiterated [ma] sequences.

The second part of the work consisted of testing whether the articulatory correlates were cues for the perception of contrastive focus. A perceptual experiment was carried out in which participants were presented with purely visual stimuli and had to identify the focus condition. The results showed that the subjects successfully perceived the focus through the visual modality alone (86% of correct answers on average). It appeared that, apart from a necessary trial of adaptation, training did not improve the performances. Subject focus was found to be significantly easier to detect than any other focus condition. Participants most often mismatched focus condition with no focus rather than the contrary. Subject focus seemed easy to mismatch with no focus, a finding supported by the fact that differences between articulatory features from focus condition to no focus condition were the least significant for the subject focus case. Taken together, these results enable us to assert that the visual modality is relevant for the perception of contrastive focus in French. Moreover, it was noticed that the stimuli with high error rates corresponded to cases where the articulatory correlates were not very salient. Those with low error rates corresponded to utterances for which the correlates were very clear and prominent. Thus, it can be assumed that there are visual cues to the perception of focus and that these cues may correspond to those described in our articulatory sketch of the production of focus. In the future, we will study perception of focus for normal non-reiterant speech.

5. Acknowledgments

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

6. References

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

[2] 't Hart J. Collier R. & Cohen A. *A perceptual study of intonation - an experimental phonetic approach to speech melody*. Cambridge University Press, 1990.

[3] House D., Hermes D. & Beaugendre F. 'Perception of Tonal Rises and Falls for Accentuation and Phrasing in Swedish'. *Proceedings of ICSLP 1998*, pp. 2799-2802.

[4] House D. 'Differential Perception of Tonal Contours through the Syllable'. *Proceedings of ICSLP 1996*, pp. 2048-2051.

[5] D'Imperio M. 'The Role of Perception in Defining Tonal Targets and their Alignment'. *PhD of The Ohio State University*, 2000.

[6] Rolland G. & Loevenbruck H. 'Characteristics of the Accentual Phrase in French: an Acoustic, Articulatory and Perceptual study'. *Speech Prosody 2002*, Aix-en-Provence, April 11-13th 2002, p 611-614.

[7] Yehia H., Kuratate T. & Vatiokis-Bateson E. 'Facial Animation and Head Motion Driven by Speech Acoustics'. *5th seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster seeon, Germany, pp. 265-268.

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

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

[10] Fougeron C & Keating P. 'Variations in velic and lingual articulation depending on prosodic position: results for 2 French speakers', UCLA *Working Papers in Phonetics* 92: pp. 88-96, 1997.

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

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

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

[14] 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*, pp. 9-24, 1993.

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

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

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

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

[19] Jun S.-A. & Fougeron C. 'Realizations of Accentual Phrases in French Intonation', Probus (14), pp. 147-172.

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

[21] 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 84, pp. 157-233.

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

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

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

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

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