Recognizing prosody from the lips

Marion Dohen, Hélène Loevenbruck, Harold Hill

Chapter proposal for *Visual Speech Recognition : Lip Segmentation and Mapping*
Alan Wee-Chung Liew & Silin Wang (Eds.)

Recognizing prosody from the lips: is it possible to extract prosodic focus from lip features?

Marion Dohen[1], Hélène Lœvenbruck[1] and Harold Hill[2,3]

[1] Speech & Cognition Department, GIPSA-lab, Grenoble, France
[2] ATR Cognitive Information Science Labs, Kyoto, Japan
[3] School of Psychology, University of Wollongong, Australia

Abstract

The aim of this chapter is to examine the possibility of extracting prosodic information from lip features. We used two measurement techniques enabling automatic lip feature extraction to evaluate the “lip pattern” of prosodic focus in French. Two corpora with Subject-Verb-Object (SVO) sentences were designed. Four focus conditions (S, V, O or neutral) were elicited in a natural dialogue situation. In a first set of experiments, we recorded two speakers of French with front and profile video cameras. The speakers wore blue make-up and facial markers. In a second set we recorded five speakers with a 3D optical tracker. An analysis of the lip features showed that visible articulatory lip correlates of focus exist for all speakers. Two types of patterns were observed: absolute and differential. A potential outcome of this study is to provide criteria for automatic visual detection of prosodic focus from lip data.

Introduction

For a spoken message to be understood (be it by a machine or a human being), the segmental information (phones, phonemes, syllables, words) needs to be extracted. Suprasegmental information, however, is also crucial. For instance, two utterances with exactly the same segmental content can have very different meanings if the suprasegmental information (conveyed by prosody) differs, as Lynne Truss (2003) nicely demonstrates:

A woman, without her man, is nothing.
A woman: Without her, man is nothing.

Prosodic information has indeed been shown to play a critical role in spoken communication. Prosodic cues are crucial in identifying speech acts and turn-taking, in segmenting the speech flow into structured units, in locating “important” words and phrases, in spotting and processing disfluencies, in identifying speakers and languages, or detecting speaker emotions and attitudes, for instance. The fact that listeners use prosodic cues in the processing of speech has led some researchers to try to draw information from prosodic features to enhance automatic speech recognition (see *e.g.* Waibel, 1988; Pagel, 1999; Yousfi & Meziane, 2006).

Prosodic information involves acoustic parameters, such as intensity, F0 pattern and duration. But prosodic information is not just acoustic, it is also articulatory, and in particular it involves visible lip features. Although prosodic focus typically involves acoustic parameters, several works have suggested that articulatory – and more specifically visible lip and jaw motion – modifications are also involved (e.g. Kelso et al., 1985; Summers, 1987; Vatikiotis-Bateson & Kelso, 1993; De Jong, 1995; Harrington et al., 1995; Lœvenbruck, 1999, 2000; Erickson et al., 2000; Erickson, 2002; Dohen et al., 2004; Cho, 2005; Dohen et al., 2006). If visual cues are associated with prosodic focus, then one can expect that prosodic focus should be detectable visually.

Prosodic phrasing and focus or stress for instance, are reflected in articulatory features associated with tongue, jaw and lip movements. More specifically, correlates of certain aspects of prosody have been reported on the lips, as will be explained below.

Despite these facts, the addition of dynamic lip information to improve automatic speech recognition robustness was limited to the segmental aspects of speech. Lip information is generally used to help phoneme categorization. Yet not only does visual information about the lips carry segmental information but also prosodic information. The question addressed in this chapter is whether prosodic information can successfully be extracted from visual facial cues, and more specifically from lip cues. If a visual speech recognition system is able to detect prosodic focus, it will better identify the information highlighted by the speaker, a function which can be crucial in a number of applications.

Background

A review of speech perception studies suggests that the extraction of prosodic information from visual lip dynamics might be possible. These studies have mostly examined the perception of “prosodic focus”, or “emphasis” the aim of which is to highlight a constituent in an utterance, without change to the segmental content. It consists for the speaker in putting forward the part of the utterance he/she wants to communicate as being the most informative (see *e.g.*, Halliday, 1967; Gussenhoven, 1983; Selkirk, 1984; Nølke, 1994; Birch & Clifton, 1995; Ladd, 1996). Focus attracts the listener’s attention to one particular constituent of the utterance and is very often used in speech communication. Among the different types of focus, contrastive focus is particularly interesting because it has clear acoustic consequences (for discussions on the distinction between different focus types, see *e.g.*, Touati, 1987; Pierrehumbert & Hirshberg, 1990; Bartels & Kingston, 1994; Di Cristo, 2000). Contrastive focus consists in selecting a constituent in the paradigmatic dimension. It is used to contrast a piece of information relative to another as in the answer to the question from the following example:

- Did Carol eat the apple?
- No, SARAH ate the apple.

Descriptions of prosodic focus in several languages have shown that the highlighted constituent bears a recognizable intonational contour (see Touati, 1989; Morel & Danon-Boileau, 1998; Rossi, 1999; Di Cristo, 2000; Touratier, 2000 for instance, for French). Focus has also durational correlates such as lengthening of the focused constituent. These cues (intonational and durational) are in fact well identified by listeners. Quite a number of studies have explored the auditory perception of prosodic contrastive focus in several languages (French: Dahan & Bernard, 1996; English: Baum *et al.*, 1982; Bryan, 1989; Gussenhoven, 1983; Weintraub *et al.*, 1981; Italian: D’Imperio, 2001; Swedish: Brådvik *et al.*, 1991). They have shown that, for all these languages, focus is very well perceived from the auditory modality.

As mentioned above, although prosodic focus typically involves acoustic parameters, several works have suggested that articulatory – and more specifically visible lip and jaw motion – modifications are also involved (e.g. Kelso *et al.*, 1985; Summers, 1987; Vatikiotis-Bateson & Kelso, 1993; De Jong, 1995; Harrington *et al*., 1995; Lœvenbruck, 1999, 2000; Erickson *et al*., 2000; Erickson, 2002; Cho, 2005). If visual cues are associated with prosodic focus, then one can expect that prosodic focus should be detectable visually.

Several studies on English, Swedish and reiterant French showed that visual detection of prosodic focus, even though not perfect, is possible (Thompson, 1934; Risberg & Agelfors, 1978; Risberg & Lubker, 1978; Bernstein *et al.,* 1989; Keating *et al.,* 2003; Dohen *et al.*, 2004). These studies suggest that visual and, typically, lip dynamics convey crucial prosodic information that might improve lip reading in conversational situations.

In order to examine the possibility of extracting prosodic information from visual lip features, we have used several measurement techniques enabling automatic lip feature extraction. We have voluntarily used very accurate measurement techniques, which have provided detailed measurements but which are unpractical for technical applications. The aim was to present what the “lip pattern” of prosodic focus consists of, taking into account inter-speaker variability. The findings presented here will provide criteria for automatic prosodic focus detection from lip data in French which can be implemented in automatic lip-feature extraction systems and which will complement the segmental information already used in most systems.

Main thrust: Methods for the extraction of prosodic information from lip features

1. *Experimental procedures*

1.1. Corpora

Two different corpora were used consisting of sentences with a subject-verb-object (S, V, O) structure and CV syllables. Sonorants were favoured in order to facilitate F0 tracking. Corpus 1 consisted of 8 sentences and corpus 2 of 13 sentences. Corpus 2 was designed as an improvement of corpus 1 after recording a first speaker. Below is an example of one of the sentences used (the reader may refer to appendices 1 & 2 for the detailed corpora).

[Lou]$_S$ [ramena]$_V$ [Manu.]$_O$ ('Lou gave a lift back to Manu.')

1.2. Prosodic focus elicitation

For all the recordings described below, four focus conditions were elicited: subject-, verb- and object-focus (narrow focus) and a neutral version (broad focus) thereafter respectively referred to as SF, VF, OF and BF. In order to trigger focus in a natural way, the speakers were asked to perform a correction task thereby focusing a phrase which had been mispronounced in an audio prompt. The recording went as follows (where capital letters signal focus):

Audio prompt: S1: Lou ramena Manu. ('Lou gave a lift back to Manu.')
S2: S1 a dit : Paul ramena Manu ? ('S1 said: Paul gave a lift back to Manu?')
Speaker utters: LOU ramena Manu. ('LOU gave a lift back to Manu.')

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

1.3. Visual lip feature selection

The typical lip parameters that characterize French vowels are lip opening (/ɛ, ɛ̃, ɔ, ɔ̃, œ, œ̃, ə, a, ɑ̃/ open *vs.* /i, y, u, e, o, ø/ closed), lip spreading (/i, e, ɛ, ɛ̃, a, ɑ̃/) and lip protrusion /y, ø, œ, œ̃, u, o, ɔ̃/). Together they satisfactorily describe French vowels (Straka, 1965; Carton, 1974). They were thus chosen as the lip features to be extracted from the articulatory data.

1.4. Data analysis

1.4.1. Preliminary acoustic validation

After the recordings, a first step consisted in acoustically validating the data *i.e.* checking whether focus had actually been produced acoustically. On the one hand, it was checked that the focused utterances displayed a typical focus intonation as described in Dohen & Lœvenbruck (2004) for example. On the other hand, an informal auditory perception test was conducted in order to check that focus was perceived auditorily.

1.4.2. Pre-shaping of the lip feature parameters

The area under the curve of variation of each parameter over time was automatically detected for each phrase (S, V and O) and then divided by the duration of the phrase. This parameter represents the mean amplitude of the feature considered over the phrase. The procedure is illustrated in Fig. 1. After this computation, three values per utterance and per feature were obtained.

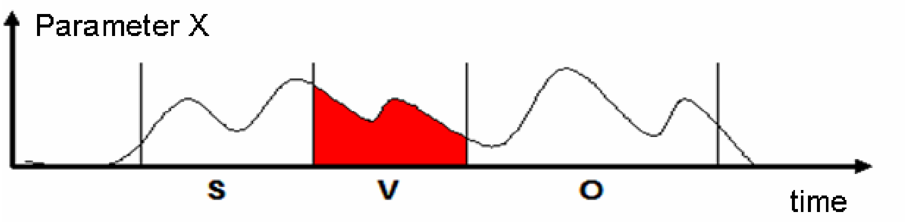

Figure 1 – *Illustration of the data pre-shaping: computation of the area under the curve corresponding to each constituent (here shown in red for the verb).*

1.4.3. Isolating supra-segmental variations from segmental varying material

Our aim was to be able to isolate and compare lip features reflecting supra-segmental variations (prosody). The problem was that, for the sake of naturalness and reproducibility, we used real speech (*vs*. reiterant speech) i.e. segmentally varying material. In order to isolate the lip features resulting solely from supra-segmental variations, and not from segmental variations (/a/ is produced with more open lips than /m/, for instance) we adopted a normalization technique. This first consisted in calculating, for each constituent (S, V, O), the mean of the areas for the neutral versions (BF) of the sentence *i.e.* two values for each constituent (two repetitions). Then all the area values corresponding to the same constituent in the same sentence uttered in the different focus conditions – *i.e.* 6 values: 3 focus conditions, 2 repetitions – were divided by this neutral mean. After this normalisation, a value of 1 corresponds to no variation of the considered parameter compared to a BF case, a value above 1 corresponds to an increase and a value below 1 to a decrease.

1.4.4. Complementary durational measurements

For all the experiments described below, complementary durational measurements were conducted since duration is an important aspect of prosodic focus (see *e.g.* Dohen & Lœvenbruck, 2004: focal syllables are lengthened and sometimes the pre-focal syllable is also lengthened as part of an anticipatory strategy) and can also be detected/processed visually. The durations of all the syllables were computed from acoustic labels assigned using Praat (Boersma & Weenink, 2005) and normalized according to the method described in 1.4.3 in order to isolate variability due to supra-segmental variations.

1.4.5. Presentation of the results

The results will always be presented using the same convention. Several graphs (such as those from Fig. 3) will be provided for each speaker, summarizing the results for all the features measured (durational and lip features). In these graphs, the means of a specific feature over all the utterances produced by the speaker are represented for three types of within utterance locations. The ‘foc’ item represents the mean of all the data corresponding to all the focused constituents (*i.e.* the focused phrase within the utterance, being either the subject, the verb or the object). The ‘pre-foc’ item represents the mean of the data corresponding to pre-focus constituents *i.e.* the subject in the case of verb focus or the subject + verb in the case of object focus. The ‘post-foc’ item represents the mean of all the data corresponding to post-focus constituents i.e. the verb + object in the subject focus case and the object in the verb focus case. In this representation, a value above 1 represents an increase compared to the neutral version of the same utterance.

1.4.6. Statistical analyses

For the sake of clarity and comparability, the same statistical analysis protocole was used for all the analyses described below. After the pre-shaping of the data described above, one value was available for each constituent (S, V, O) from each utterance. The statistical analyses were conducted for all the data corresponding to focus cases (SF, VF and OF) since the normalisation procedure (see section 1.4.3) included the neutral case as the basis for normalisation.

The first analysis aimed at testing intra-utterance contrasts *i.e.* contrasts within the utterance. The question was: *is there a significant difference between the focused constituent and the rest of the utterance?* This led to the analysis of two within-subject factors. The first one was a two-level factor called Congruency. The congruent cases correspond to S and subject focus (S&SF), V and verb focus (V&VF) and O and object focus (O&OF). The incongruent cases correspond to V and O for subject focus (V&SF, O&SF), S and O for verb focus (S&VF, O&VF) and S and V for object focus (S&OF, V&OF). The second within-subject factor was a three-level factor corresponding to focus type (SF, VF or OF). For each lip and durational feature (see the following sections for the feature definitions, depending on the measurement method), a two-way multivariate analysis of variance (ANOVA, see Howell, 2004) was conducted with the aforementioned within-subject factors (*i.e.* congruency and focus type). The sphericity of the data was tested using Mauchly’s sphericity test. When the test was significant we used a Huynh-Feldt correction on the degrees of freedom (the results presented below include these corrections but, for clarity, the “true” numbers of degrees of freedom are in fact reported).

The second analysis aimed at testing inter-utterance contrasts in order to answer the following question: *is there a significant difference between a constituent in the focused version of the utterance and in the neutral version of the utterance?* This was tested using t-tests (Howell, 2004) comparing the values corresponding to a specific constituent in the focused case to 1 (after normalisation the neutral case corresponds to 1). The following tests were conducted:

- test 1: comparison of the values corresponding to the focused constituents to 1
- test 2: comparison of the values corresponding to the pre-focus constituents (S in the VF and OF cases and V in the OF case) to 1

- test 3: comparison of the values corresponding to the post-focus constituents (V and O for SF cases and O for VF cases).

The results of all these tests are summarized in tables such as Table 1.

2. Method 1: lip tracking from video data

2.1. Lip feature extraction

In the first method, we used a lip tracking device designed at the former Institut de la Communication Parlée (now Speech & Cognition Department, GIPSA-lab) (Lallouache 1991, Audouy 2000). This device consists in using blue make-up on the speaker's lips, a blue marker on his/her chin and front and profile blue references (front: on the eyeglasses; profile: vertical ruler fixed on the eyeglasses). The speaker is filmed using front and profile cameras (digital; 25 fps). Fig. 2 gives an example of the images recorded.

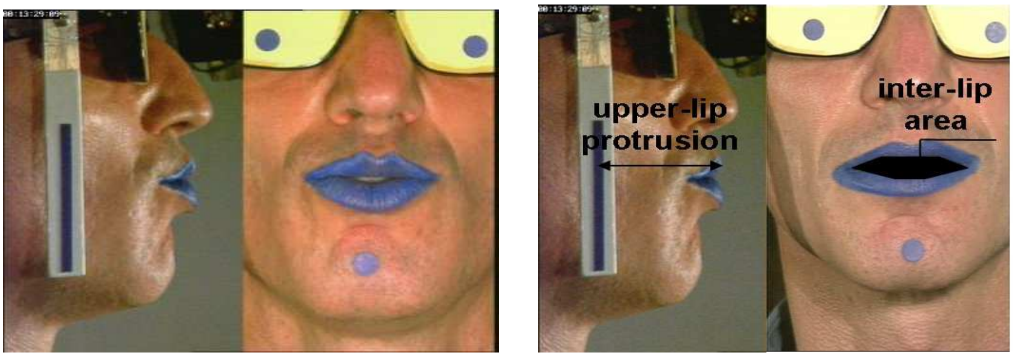

Figure 2 – *Lip tracking device: a. (left) example of a recorded image; b. (right) extracted features.*

A software program automatically extracts the lip contour from the sequence of digitalized video frames and derives parameters describing inter-lip area (LA) and upper-lip protrusion. Inter-lip area accounts for both lip opening and lip spreading. We analyzed these parameters as well as LA's first derivative using the procedures described in section 1.4.

2.2. Recordings

Corpus 1 was recorded for one native speaker of French (speaker A). Due to the fact that these data were recorded for parallel studies, the corpus was mainly designed to test lip opening and lip spreading and contained very few protruded vowels. Therefore for speaker A, only inter-lip area was extracted from the video. Corpus 2 was adapted to additionally make lip protrusion analysis possible and was recorded for another native speaker of French (speaker B). Therefore for speaker B, both inter-lip area and upper lip protrusion were extracted from the video.

2.3. Results

The results from the lip feature extraction are provided in Fig. 3 for both speakers. Table 1 provides the results of the statistical analyses conducted using the procedure described in section 1.4.6. A number of articulatory and durational correlates to prosodic focus can be extracted from these measurements for each speaker.

First, for the intra-utterance comparisons, table 1 shows that congruency has a significant effect for both speakers on duration, inter-lip area, inter-lip area's first derivative (SA only) and upper lip protrusion (SB only). This means when a constituent is focused, its duration, inter-lip area, inter-lip area's first derivative (SA only) and upper lip protrusion are significantly greater than those corresponding to the other constituents in the same utterance (intra-utterance contrast). Focus type has a significant effect on duration (SA only) and inter-lip area (SB only). The effect on duration for speaker A is due to the fact that all the syllables of the utterance were longer when the verb was focused, for this speaker. The effect on inter-lip area for speaker B is due to the fact that inter-lip area was always greater when the verb or the object were focused than when the subject was focused, for this speaker. There is a significant interaction between congruency and focus type for duration for SA only. This is due to the fact that intra-utterance contrast for duration was much greater for the focused verbs than for the focused subjects and objects. This is an artefact of the corpus for SA, in which there were many occurrences of monosyllabic verbs: when the focused constituent is mono-syllabic, the mean syllabic correlates of focus are increased.

Secondly, for the inter-utterance comparisons, table 1 shows that test 1 is significant for both speakers for duration, inter-lip area (SA only), inter-lip area's first derivative and upper lip protrusion (SB only). This shows that overall, when a constituent is focused, it is lengthened and hyper-articulated (larger and "faster" inter-lip area and upper-lip protrusion) compared with the same constituent in a neutral version of the utterance. Figure 3 illustrates this (values above 1). Test 2 is significant for all features for SA and for duration for SB. For speaker A, lip features were not only enhanced for the focused constituent but also for the pre-focal constituent (see Fig. 3: values above 1). This corresponds to an anticipatory strategy described in Dohen *et al.* (2004). For speaker B, the duration of the pre-focused constituent was significantly reduced compared with the neutral rendition (see Fig. 3: value below 1). Test 3 is significant for inter-lip area and upper lip protrusion for speaker B. This shows that for speaker B, inter-lip area and upper lip protrusion are decreased on the post-focal constituent compared to the same constituent in the neutral version (see Fig. 3: values below 1).

The strategies of both speakers are summarized below:
Speaker A – focal lengthening; focal increase of lip feature amplitudes (inter-lip area and its first derivative); largest contrast for inter-lip area features.
Speaker B – focal lengthening; focal increase of lip feature amplitudes (inter-lip area and its first derivative and upper-lip protrusion); largest contrast for upper-lip protrusion.

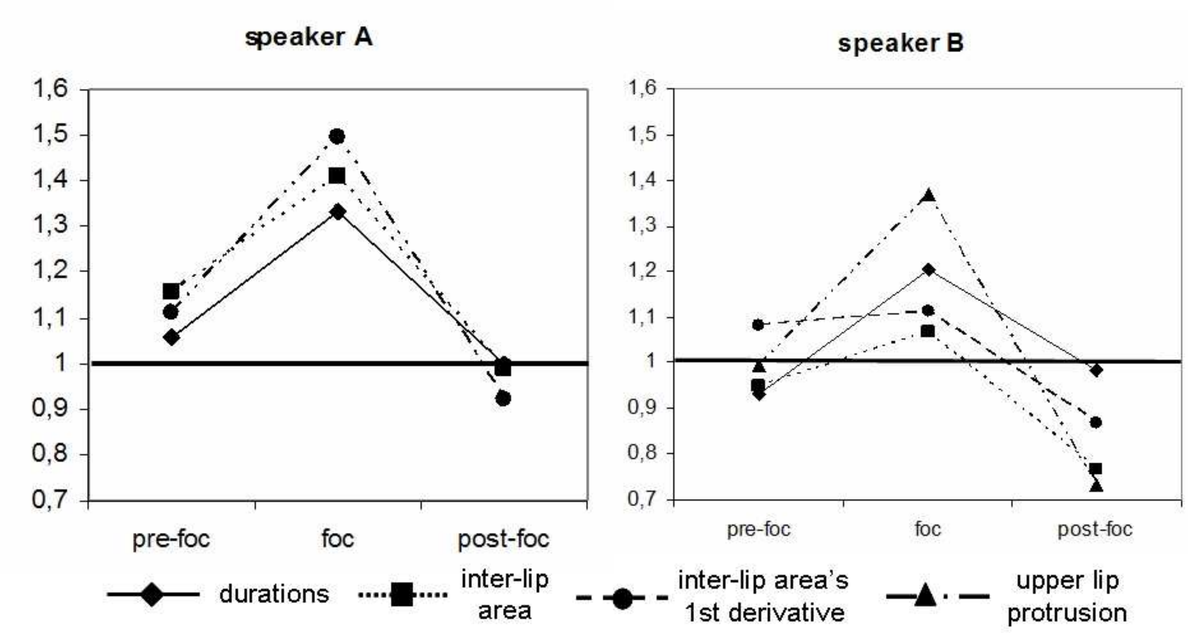

Figure 3 – *Lip tracking: durational measurements and lip features for speakers A and B: normalized values corresponding to the pre-focal, focal and post-focal sequences (the dark horizontal lines correspond to the neutral case).*

		Intra-utterance contrasts			*Inter-utterance contrasts*		
		Congruency	*Focus type*	*Interaction*	*Test 1*	*Test 2*	*Test 3*
Duration	SA	***F(1,15)=158.9 p<.001***	***F(2,30)=19.2 p<.001***	***F(2,30)=13.6 p<.001***	***t=8.6 p<.001***	***t=3 p=.004***	*t=-0.3 p=.731*
	SB	***F(1,25)=180.7 p<.001***	*F(2,50)=3.6 p=.036*	*F(2,50)=2.2 p=.117*	***t=8.2 p<.001***	***t=-4.8 p<.001***	*t=-1.1 p=.281*

Inter-lip area	SA	***F(1,15)=202.2*** ***p<.001***	*F(2,30)=4.3* *p=.023*	*F(2,30)=3.8* *p=.061*	***t=9.8*** ***p<.001***	***t=4.3*** ***p<.001***	*t=-0.8* *p=.447*
	SB	***F(1,25)=53.4*** ***p<.001***	***F(2,50)=8.6*** ***p=.001***	*F(2,50)=2.3* *p=.112*	*t=2.3* *p=.026*	*t=-2* *p=.047*	***t=-10.4*** ***p<.001***
LA's 1st derivative	SA	***F(1,15)=47.5*** ***p<.001***	*F(2,30)=1.9* *p=.17*	-	***t=6.6*** ***p<.001***	***t=2.8*** ***p=.007***	*t=-2.9* *p=.006*
	SB	**-**	-	-	***t=3.8*** ***p<.001***	-	*t=-2.6* *p=.011*
Upperl lip protrusion	SB	***F(1,25)=19.8*** ***p<.001***	*F(2,50)=2.7* *p=.076*	-	***t=3.3*** ***p=.001***	*t=-0.1* *p=.945*	***t=-6.2*** ***p<.001***

Table 1 – *Results of the statistical analyses for the lip-tracking data for speakers A and B, using the statistical analysis protocole described in section 1.4.6. The F values correspond to the F-test statistic. The t-values correspond to the t-test statistic. The p values correspond to the significance level. An effect was considered as significant when p < .01 (bold characters signal significant effects).*

These findings suggest that for speaker A, values of normalized duration, normalized inter-lip area and its first derivative, and normalized upper lip protrusion above 1.2 may characterize a focused constituent. For speaker B, the pattern is a little more complex: the graph suggests that a focused constituent may be detected when all parameters are above 1 for the given constituent and below one for the following constituent. This latter result is in line with the post-focal deaccenting phenomenon that has been described acoustically (see *e.g.* Dohen & Loevenbruck 2004).

3. *Method 2: Optotrak*

3.1. Recordings

Five native speakers of French (B, C, D, E and F) were recorded using corpus 2 (see section 1.1) and the procedure described in section 1.2. Speaker B was the same as the speaker B recorded using method 1. The recordings were made using a 3D optical tracking system: Optotrak. The system consists of three infrared (IR) cameras which track the positions of infrared emitting diodes (IREDs) glued to the speaker's face (thereafter referred to as markers). The 3D coordinates of each IRED were automatically detected over time. For this experiment, we used two Optotraks in order to compensate for missing data, corresponding to momentary hiding of markers when the speaker moves (head turns, for example). Data were corrected for head motion using 4 markers placed on a head rig as shown on Fig. 4 and Fig. 5 (markers 1-4). IRED positions were sampled at 60Hz and low-pass filtered. The acoustic signals were recorded simultaneously and sampled at 22kHz. Fig. 4 gives an idea of the experimental setup and Fig. 5 provides a schematic view of the marker locations. Only the measurements corresponding to the markers located on the lips of the speakers (see Fig. 5: markers 8-10 and 12-16) will be discussed here since the purpose of this analysis was to study lip features. The other facial and head movement markers were used for another study and for the sake of clarity and conciseness, they will not be discussed here. After the recordings, an acoustic validation was conducted using the procedure described in section 1.4.1. It showed that, from an acoustic point of view, all the speakers had correctly produced focus.

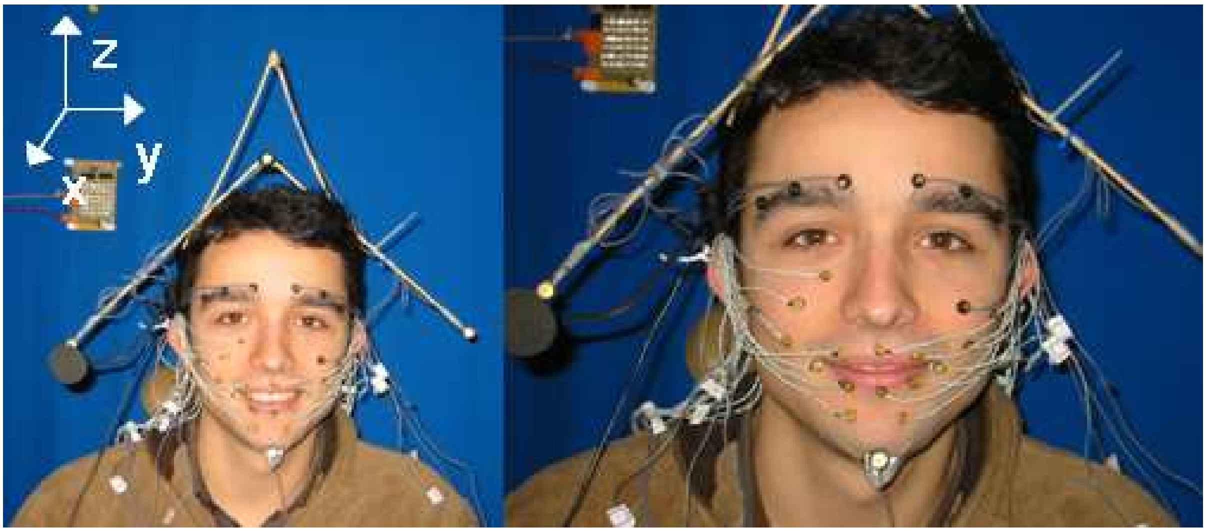

Figure 4 – *Optotrak measurement device: experimental setup.*

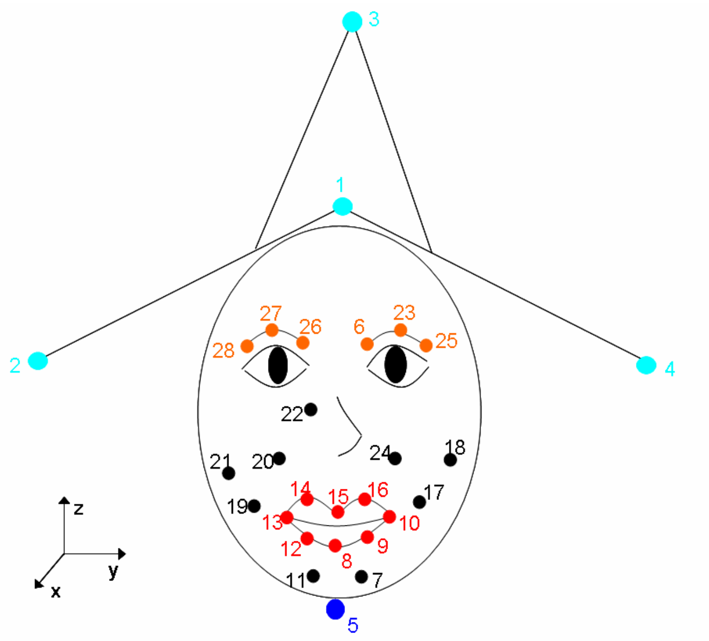

Figure 5 – *Map of the locations of the IRED diodes referred to as "markers".*

3.2. Lip feature extraction

Various lip features were computed from the 3D coordinates of the IREDs:

- Lip opening was computed as the difference between the z coordinates of the upper and lower middle lip markers (see Fig. 5: markers 8 and 15).
- Lip spreading corresponded to the difference between the y coordinates of the two lip corner markers (see Fig. 5: markers 13 and 10).
- Upper lip protrusion was assimilated to the x coordinate of the middle upper lip marker (see Fig. 5: marker 15).

In addition, vertical jaw movements (z coordinate of the chin marker, i.e. markers 5) were also analyzed but the results will not be discussed here, as they were intended for a different study.

3.3. Results

The results for all speakers are given in Fig. 6 and summarized thereafter. The jaw parameter was collected for a different study. Only results on the lips are reported here. For the sake of clarity, we will only give a general overview of the statistical results in the text. The aim is indeed to put forward trends which are consistent from one speaker to another. The detailed results of the statistical analyses are provided in Table 2, however.

First, for the intra-utterance comparisons, table 2 shows that, for all speakers, congruency has a significant effect on duration, lip opening and upper lip protrusion and no significant effect on lip spreading (except for SE). This shows that when a constituent is focused, it is significantly lengthened and hyper-articulated (larger lip opening, greater upper-lip protrusion) compared to the other constituents of the same utterance. There is a significant intra-utterance contrast between the visual lip features corresponding to the focused constituent and the visual lip features corresponding to the other constituents of the utterance. For duration, focus type also has a significant effect, illustrating the fact that when the verb is focused, all the syllables of the utterance are lengthened. There is also a significant interaction between congruency and focus type for duration for all speakers. This is due to the fact that when the verb is focused the intra-utterance contrast for duration is significantly stronger.

Secondly, for the inter-utterance comparisons, table 2 shows that test 1 is significant for all speakers for duration, lip opening and upper-lip protrusion. This shows that overall, when a constituent is focused, it is lengthened and hyper-articulated (larger lip opening and greater upper-lip protrusion) compared with the same constituent in a neutral version of the utterance. It is also the case for lip spreading for three of the five speakers. Figure 6 illustrates this (values above 1). For SC, SD, SE and SF, this corresponds to a significant lengthening of the pre-focal constituent compared to the same constituent in a neutral version of the utterance (see Fig. 6: values above 1). For SB, it corresponds to a significant reduction of the duration of this constituent (see Fig. 6: values below 1). Test 2 is also significant for SC and SD for lip opening (see Fig. 6: values above 1). This corresponds to an increase in lip opening for the pre-focal constituent. These results (for test 2) suggest that some speakers use an anticipatory strategy to signal focus by starting to lengthen and hyper-articulate before focus. Test 3 is significant for SB, SD and SE for lip opening, for SB for lip spreading and for SE and SF for upper-lip protrusion. In all these cases (except for SF for upper-lip protrusion), this corresponds to a decrease on the post-focal constituent compared to the same constituent in the neutral version (see Fig. 6: values below 1). This suggests that, after focus, some speakers tend to shorten and articulate less.

The strategies of all the speakers are summarized below:

Speaker B – focal lengthening; focal increase of lip feature amplitudes (except for lip spreading); post-focal decrease of lip feature amplitudes of all the parameters; largest contrast for protrusion and duration. Since speaker B was recorded using both methods (see section 2), the results can be compared. It appears that the trends are the same in the two methods with the same ranges except for protrusion. It is difficult to accurately measure lip protrusion, as it is very sensitive to the reference used. This could explain the range difference.

Speaker C – focal lengthening; focal increase of lip feature amplitudes; slight post-focal decrease of lip opening amplitudes; largest contrast for protrusion and duration.

Speaker D – focal lengthening; focal increase of lip feature amplitudes (except lip spreading); post-focal decrease of lip opening and protrusion amplitudes; largest contrast for protrusion; smallest contrast for lip spreading.

Speaker E – focal lengthening; focal increase of lip feature amplitudes; post-focal decrease of lip features amplitudes; largest contrast for protrusion; smallest contrast for lip opening and spreading.

Speaker F – focal lengthening; focal increase of lip feature amplitudes; largest contrast for protrusion; smallest contrast for lip opening.

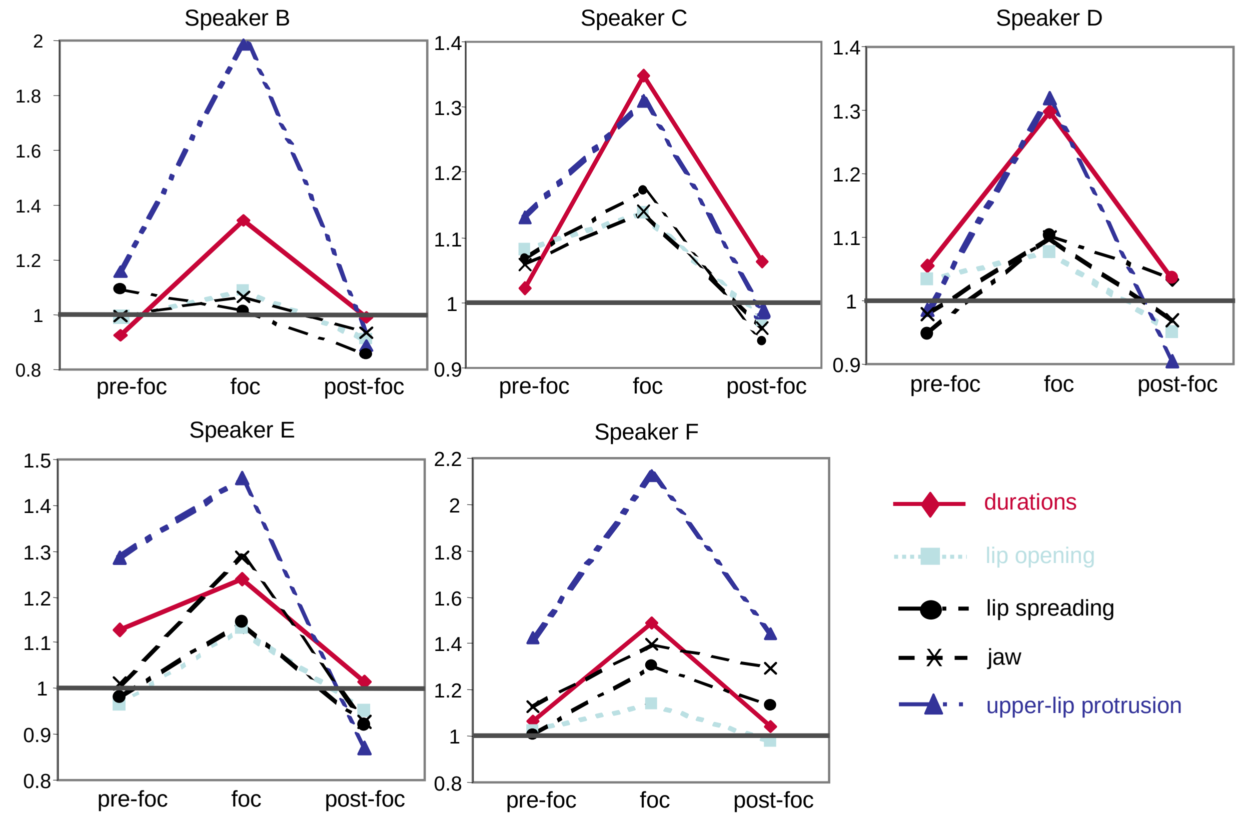

Figure 6 – *Optotrak: durational measurements and lip features for speakers B to F: normalized values corresponding to the pre-focal, focal and post-focal sequences (the horizontal line shows the neutral case).*

		Intra-utterance contrasts			*Inter-utterance contrasts*		
		Congruency	*Focus type*	*Interaction*	*Test 1*	*Test 2*	*Test 3*
Duration	SB	**F(1,25)=198.7 p<.001**	**F(2,50)=15.6 p<.001**	**F(2,50)=5.6 p=.006**	**t=9.8 p<.001**	**t=-2.8 p=.007**	t=-0.8 p=.402
	SC	**F(1,25)=323.9 p<.001**	**F(2,50)=13.9 p<.001**	**F(2,50)=6.6 p=.003**	**t=13.9 p<.001**	**t=4 p<.001**	**t=4.3 p<.001**
	SD	**F(1,25)=109.4 p<.001**	**F(2,50)=15.5 p<.001**	**F(2,50)=10.2 p<.001**	**t=9,5 p<.001**	**t=3.2 p=.003**	t=2.2 p=.033
	SE	**F(1,25)=50 p<.001**	**F(2,50)=5.8 p=.005**	**F(2,50)=7.3 p=.002**	**t=9 p<.001**	**t=5.3 p<.001**	t=0.8 p=.434
	SF	**F(1,25)=239.6 p<.001**	**F(2,50)=10.7 p=.001**	**F(2,50)= 7.9 p=.003**	**t=11.7 p<.001**	**t=3.8 p<.001**	t=2.1 p=.041
Lip opening	SB	**F(1,25)=11.1 p<.001**	**F(2,50)=11.7 p<.001**	F(2,50)=0.907 p=.41	**t=5.3 p<.001**	t=-0.1 p=.89	**t=-10 p<.001**
	SC	**F(1,25)=149.1 p<.001**	F(2,50)=0.1 p=.880	-	**t=10 p<.001**	**t=6 p<.001**	t=-2.4 p=.02
	SD	**F(1,25)=49.2 p<.001**	F(2,50)=0.6 p=.557	-	**t=4.2 p<.001**	**t=3.3 p=.002**	**t=-3.5 p=.001**
	SE	**F(1,25)=111.4 p<.001**	F(2,50)=2.1 p=.137	-	**t=9.1 p<.001**	t=-2.4 p=.018	t=-3.7 p<.001

	SF	**F(1,25)=97.9 p<.001**	F(2,50)=0.6 p=.562	-	**t=7.3 p<.001**	t=1 p=.335	t=-1 p=.332
Lip spreading	SB	F(1,25)=0.6 p=.462	F(1,725.50)=2.2 p=.134	-	t=0.2 p=.831	t=0.6 p=.567	**t=-5.7 p<.001**
	SC	**F(1,25)=11.8 p=.002**	F(2,50)=3.7 p=.033	F(2,50)=0.5 p=.59	**t=3.6 p=.001**	t=0.7 p=.459	t=-1.5 p=.130
	SD	F(1,25)=3.8 p=.063	F(1,421.50)=1.2 p=.298	-	t=1.7 p=.092	t=0.1 p=.943	t=0.8 p=.436
	SE	**F(1,25)=47.1 p<.001**	F(2,50)=4 p=.024	F(2,50)=1.4 p=.250	**t=3.5 p=.001**	t=0.7 p=.514	t=-2.4 p=.021
	SF	**F(1,25)=11.6 p=.002**	F(2,50)=3.7 p=.033	F(2,50)=1.6 p=.218	**t=3.6 p=.001**	t=-0.6 p=.585	t=1.5 p=.139
Lip protrusion	SB	**F(1,25)=72 p<.001**	**F(2,50)=10.5 p<.001**	**F(2,50)=7.8 p<.001**	**t=8.3 p<.001**	t=2.7 p=.01	t=-2 p=.048
	SC	**F(1,25)=38.2 p<.001**	**F(2,50)=7.4 p<.001**	F(2,50)=0.5 p=.628	**t=6.5 p<.001**	t=1.9 p=.065	t=-0.3 p=.803
	SD	**F(1,25)=5.5 p<.001**	F(2,50)=3.2 p=.05	F(2,50)=0.5 p=.592	**t=4 p<.001**	t=0.6 p=.574	t=-2.5 p=.014
	SE	**F(1,25)=5.7 p<.001**	F(2,50)=0.2 p=.860	-	**t=5.1 p<.001**	t=2 p=.046	**t=-2.5 p=.002**
	SF	**F(1,25)=17 p<.001**	F(2,50)=6.1 p=.099	-	**t=5.7 p<.001**	t=2.7 p=.011	**t=3.2 p=.002**

Table 2 – *Results of the statistical analyses for the Optotrak data for speakers B to F using the statistical analysis protocole described in section 1.4.6. The F values correspond to the F-test statistic. The t-values correspond to the t-test statistic. The p values correspond to the significance level. An effect was considered as significant when p <.01 (bold characters signal significant effects).*

The results suggest that when the normalized values of duration, lip opening, lip spreading and upper-lip protrusion are above 1 for a given constituent and decrease for the following constituent, the first constituent might bear focus. Furthermore, when the normalized values are only slightly above 1 for a given constituent, the fact that the values for the next constituent are below one is a further indication that the first constituent was focused.

Conclusion: Lip feature criteria for the detection of prosodic focus

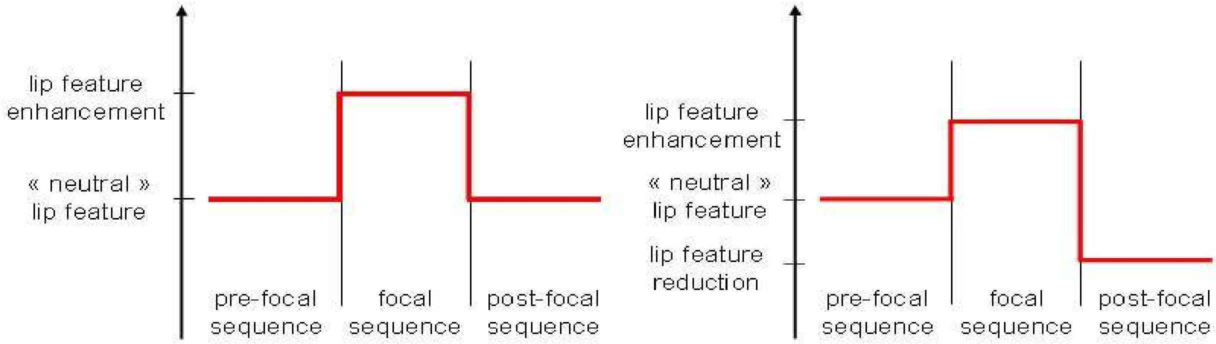

Figure 7 – *Schematic representation of a. (left) the absolute contrast pattern and b. (right) the differential contrast pattern.*

The results described above suggest that it is possible to extract information from lip features about an important phenomenon in conversational situations, namely prosodic focus. One of the main conclusions that can be drawn is the fact that focus affects the lip features of the whole utterance and not only that of the specific focused constituent. Another important observation is that there is inter-speaker variability. However, after examining the productions of six different speakers, two main lip feature patterns can be extracted corresponding to prosodic focus production:

Absolute contrast pattern: the focal constituent is lengthened and the features describing lip shape (inter-lip area, lip opening, lip spreading, upper-lip protrusion) are increased to a large extent. The peak velocities of the evolution of these features over time are also increased. This pattern is illustrated in Fig. 7.a.

Differential contrast pattern: in this case, the focal constituent is also lengthened and the features describing lip shape are also increased but to a smaller extent. Additionally, the lip features corresponding to the post-focal sequence are decreased. An important contrast is thus created inside the utterance: the focal increase is not made very strong but is reinforced by the post-focal decrease. Fig. 7.b illustrates this pattern.

Therefore although inter-speaker variability exists, consistent strategies can be described. Futhermore, the differential contrast strategy seems to be the most used (4 speakers out of 6). This strategy seems the most economical in terms of articulatory effort while preserving a good contrast within the utterance and allowing correct focus detection. These production strategies provide good criteria for focus detection. An absolute contrast or a differential contrast on a given constituent in the utterance seems to be a good criterion for detecting the presence of focus.

We found that whatever the pattern observed, the lip feature with the highest variations under focus was upper-lip protrusion. This is consistent with the finding that lip protrusion is the most visible lip feature (Benoît *et al.*, 1994).

We note also that the results obtained with the second method are consistent with those found with the first method. Interlip area was used in the first method, where internal lip contour was easily derivable from video data. Lip opening + lip spreading were used in the second method, where the positions of specific markers were easily obtainable and could provide these distances. The consistency in the results suggest that any of these two parameter sets (interlip area vs. lip spreading + lip opening) can be used to detect prosodic focus.

These findings therefore enabled us to sketch a model for the production of visual features corresponding to prosodic focus in French obtained with very accurate and detailed measurement techniques. This model covers the different strategies used by different speakers. It could now be used on visual data extracted using other (more practical) methods that extract lip parameters such as lip protrusion and lip opening or lip area.

4. Future research directions: visual speech recognition

The two studies described in this chapter suggest that it is possible to extract prosodic information from lip information. The measurements and analyses described enabled us to design a model characterizing lip features typically associated with prosodic focus in French. The lip features concerned are vertical lip opening, horizontal lip spreading and upper lip protrusion. Interlip area is a good summary of lip opening and lip spreading and can be used instead. What this model mainly shows is that prosodic focus results in a marked enhancement of the lip features corresponding to the focused constituent compared to that of the other constituents of the same utterance. These findings can potentially be used for the detection of prosodic focus in automatic visual speech recognition in the following way: the contrast criteria described above can be applied to the pattern of lip features automatically extracted from the utterance.

In the studies described here, we used two lip feature extraction devices which cannot easily be used for commercial applications because of heavy and sophisticated setups, both from the equipment and the speaker point of view. We used these devices because of their very good accuracy, since precision was important to establish a reliable model. We used two different devices in order to evaluate many different lip parameters and test whether the observations were the same from one device to another. However, now that the model is established, it seems feasible to use other more “portable” lip feature extraction devices which could potentially be integrated into commercial applications. We suggest that crucial prosodic information, that might improve lip reading in conversational situations, can potentially be detected using our model.

Acknowledgements

We thank Guillaume Rolland for designing corpus 1 and recording speaker A. For the lip tracking data, we thank Christophe Savariaux and Alain Arnal for their technical help. For the optotrak recordings, we thank Ishi-san, Guillaume Vignali and the ATR Research Institute (Japan) for their help. We are grateful to all the speakers who participated. We also thank Jean-Luc Schwartz for his comments on our work, as well as three anonymous reviewers who provided very helpful comments on a previous version of this chapter.

References

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

Bartels, C., & Kingston, J. (1994). Salient Pitch Cues in the Perception of Contrastive Focus. In P. Bosch & R. Van Der Sandt (Eds.), Focus & Natural Language Processing, *Proceedings of Journal of Semantics Conference on Focus*, IBM Working Papers, TR-80 (pp. 94-106).

Baum, S. R., Kelsch Daniloff, J., Daniloff, R., & Lewis, J. (1982). Sentence Comprehension by Broca's aphasics: effects of some suprasegmental variables. *Brain and Language,* 17, 261-271.

Benoît C., Mohamadi T., & Kandel S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *J. Speech and Hearing Research*, 37, 1195-1203.

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(1), 397-405.

Birch, S., & Clifton, Jr. C. (1995). Focus, accent, and argument structure: effects on language comprehension. *Language and speech*, 38, 365-391.

Boersma, P., & Weenink, D. (2005). PRAAT: Doing phonetics by computer (version 4.3) [Computer program]. Retrieved from http://www.praat.org.

Brådvik, B., Dravins, C., Holtås, S., Rosén, I., Ryding, E., & Ingvar, D. (1991). Disturbances of Speech Prosody Following Right Hemisphere Infarcts. *Acta Neurologica Scandinavica*, 84, 114-126.

Bryan, K. (1989). Language Prosody and the Right Hemisphere. *Aphasiology,* 3, 285-299.

Carton, F. (1974). *Introduction à la phonétique du français*, Bordas, Paris.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *Journal of the Acoustical Society of America*, 117(6), 3867-3878.

De Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 97(1), 491-504.

Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français (deuxième partie). *Journal of French Language Studies*, 10, 27-44.

Dahan, D., & Bernard, J.-M. (1996). Interspeaker Variability in Emphatic Accent Production in French. *Language and Speech,* 39(4), 341-374.

D'Imperio, M. (2001). Focus and Tonal Structure in Neapolitan Italian. *Speech Communication*, 33(4), 339-356.

Dohen, M., & Lœvenbruck, H. (2004). Pre-focal Rephrasing, Focal Enhancement and Post-focal Deaccentuation in French. *Proceedings of ICSLP 2004* (pp. 1313-1316).

Dohen, M., Loevenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech, *Speech Communication,* 44, 155-172.

Erickson, D., Maekawa, K., Hashi, M., & Dang, J. (2000). Some articulatory and acoustic changes associated with emphasis in spoken English. Proceedings of the *ICSLP 2000 conference, Beijing, China,* (vol. 3, pp. 247-250).

Erickson, D. (2002). Articulation of Extreme Formant Patterns for Emphasized Vowels. *Phonetica*, 59, 134-149.

Gussenhoven, C. (1983). Testing the Reality of Focus Domains. *Language and Speech*, 26(1), 61-80.

Gussenhoven, C. (1984). *On the grammar and semantic of sentence accents*. Dordrecht: Foris.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

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

Howell, D. C. (2004). *Fundamental statistics for the behavioral sciences (5th edition)*. Belmont, CA: Brooks/Cole.

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 the ICPhS 2003 conference, Barcelona, Spain* (pp. 2071-2074).

Kelso, J.A. S., Vatikiotis-Bateson, E., Saltzman, E., & Kay, B. A. (1985). A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, 77(1), 266-280.

Ladd, R. D. (1996). *Intonational phonology*. Cambridge studies in Linguistics.

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

Lœvenbruck, H. (1999). An investigation of articulatory correlates of the Accentual Phrase in French. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, USA, (Vol. 1, pp. 667-670).

Lœvenbruck, H. (2000). Effets articulatoires de l'emphase contrastive sur la Phrase Accentuelle en français. *Proceedings of the Journées d'Etude de la Parole, Aussois, France* (pp. 165-168).

Morel, M.-A., & Danon-Boileau, L. (1998). *Grammaire de l'intonation. L'exemple du français oral*. Paris-Gap, France: Ophrys, Bibliothèque de Faits de Langues.

Nølke, H. (1994). *Linguistique modulaire : de la forme au sens*. Louvain-Paris, Peeters.

Pagel, V. (1999). *De l'utilisation d'informations acoustiques suprasegmentales en reconnaissance de la parole continue*. Unpublished doctoral dissertation, Université Henri Poincaré - Nancy 1, France.

Pierrehumbert, J., & Hirshberg, J. (1990). The meaning of intonational contours in discourse. In P. Cohen, J. Morgan; M. Pollack (Eds.), *Intentions in Communication* (pp. 271-311). Cambridge, MA, USA: The MIT Press.

Risberg, A., & Agelfors E. (1978). On the identification of intonation contours by hearing impaired listeners. *Speech Transmission Laboratory Quarterly Progress Report and Status Repor*t, 19(2-3), 51-61.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report and Status Repor*t, 19(4), 1-16.

Rossi, M. (1999). La focalisation. *L'intonation, le système du français: description et modélisation*, (Chap. II-6, pp. 116-128). Ophrys.

Selkirk, E. O. (1984). The grammar of intonation. In E. O. Selkirk (Ed.), *Phonology and syntax: the relation between sound and structure* (pp. 197-296). The MIT Press.

Straka, G. (1965). *Album phonétique*, Les Presses de l'Université de Laval, Québec.

Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, 82(3), 847-863.

Thompson, D. M. (1934). On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues. *Journal of General Psycholog*y, 11, 160-172.

Touati, P. (1987). Structures prosodiques du suédois et du français. *Lund Working Papers*, 21, Lund University Press.

Touati, P. (1989). De la prosodie française du dialogue. Rapport du projet KIPROS. *Working Papers, Lund University*, 35, 203-214.

Touratier, C. (2000). *La sémantique*. Paris, France: Armand Collin.

Vatikiotis-Bateson, E., & Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Journal of Phonetics*, 21, 231-265.

Waibel, A. (1988). *Prosody and speech recognition*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Weintraub, S., Mesulam, M.-M., & Krahmer, L. (1981). Disturbances in Prosody: A Right-hemisphere Contribution to Language. *Archives of Neurology*, 38, 742-744.

Yousfi, A., & Meziane, A. (2006). The Centisecond Two Levels Hidden Semi Markov Model (CTLHSMM). *International Symposium on Parallel Computing in Electrical Engineering* (pp. 101-104).

Additional Reading

Abry, C., & Boë, L.-J., (1986). Laws for Lips. Speech Communication, 5, 97-104.

Abry, C., Boë, L.-J.,, Corsi, P., Descout, R., Gentil, M., & Graillot, P. (1980). *Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales.* Publications de l'Université des Langues et Lettres de Grenoble Grenoble.

Baum S. R., & Pell M. D. (1999). The neural bases of prosody: insights from lesion studies and neuroimaging. *Aphasiology*, 13 (8).

Beckman, M. E. (1986). *Stress and Non Stress Accent.* Dordrecht: Foris, The Netherlands.

Bolinger, D. (1989). *Intonation and its uses. Melody in grammar and discourse*. Stanford University Press, CA.

Collier, R., & 't Hart, J. (1975). The role of intonation in speech perception. *Structure and Process in Speech*, Cohen and Noteboom (eds.), 107-23, Berlin, Springer Verlag.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Cutler, A. (1984). Stress and accent in language production and understanding. In D. Gibbon & H. Richter (Eds.), *Intonation accent and rhythm : studies in discourse phonology* (pp. 77-90).

Cutler A., Dahan D., & van Donsellar W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and speech*, 40(2).

Delatttre, P. (1967). La nuance du sens par l'intonation. French Review, 41, 3, 326-339.

Dohen, M. (2005). *Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français*. Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble, France.

Firth, J. R. (1948). Sound and prosodies. *Transactions of the Philological Society*, 127-152.

Fónagy I. (1981). Fonction prédictive de l'intonation. *Problèmes de prosodie* II, Expérimentations , modèles et fonctions. Léon P. & Rossi M. (eds.), 113-120.

Jun, S.-A., & Fougeron, C. (2000). A Phonological Model of French Intonation. In A. Botinis (Ed.), *Intonation: Analysis, modelling and technology* (pp. 209-242). Dordrecht: Kluwer Academic Publishers.

Hirst, D.,. & Di Cristo, A. (1998). Intonation systems: a survey of twenty languages. Cambridge University Press.

House, D., Bruce, G., Ericksson, L., & Lacerda, F. (1990). Recognition of prosodic categories in Swedish: Rule implementation. *Working Papers*, Lund University, 34, 62-66.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge Mass: MIT Press.

Mandel, D. R., Jusczyk, P. W., & Kemler Nelson, D.G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, 53 (2).

Monrad-Krohn, G. H. (1947). Dysprosody or altered "melody of language." *Brain*, 70, 405–415.

Searle, J. R. (1969*). Speech acts*. Cambridge: Cambridge University Press.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25.

Tabain, M. (2003). Effects of prosodic boundary on /aC/ sequences: Articulatory results. *Journal of the Acoustical Society of America*, 113 (5).

Terken, J. (1993). Issues in the perception of prosody. *Proceedings of the ESCA Workshop on Prosody*, Lund.

Truss, L. (2003). *Eats, shoots and leaves. The zero tolerance approach to punctuation*. Profile Books Ltd, London, UK.

Vaissière, J. (1989). The use of prosodic parameters in automatic speech recognition. In Nieman, Lang & Sagerer (Eds.), *Recent advances in speech understanding and dialog systems*. Nato Asi Series, Springer Verlag.

Welby, P. (2003). French intonational rises and their role in speech seg mentation. *Proceedings of Eurospeech: The 8th Annual Conference on Speech Communication and Technology*, Geneva, Switzerland (pp. 2125–2128).

Appendix 1 – Corpus AV1

The number next to S/V/O is the number of syllables of the constituent.

(s1) [Jean]$_{S1}$ [veut ménager]$_{V4}$ [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]$_{O7}$

'Mélanie saw the unhappy bad wolves.'

(s4) [Véroniqua]$_{S4}$ [mangeait]$_{V2}$ [les mauvais melons]$_{O5}$

'Véroniqua was eating the bad melons.'

(s5) [Les mauvais loups]$_{S4}$ [mangeront]$_{V3}$ [Jean]$_{O1}$

'The bad wolves will eat Jean.'

(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]$_{V1}$ [Véroniqua]$_{O4}$

‘The good-looking sailor saw Véroniqua.’

Appendix 2 – Corpus AV2

The four first sentences of corpus AV2 are (s2), (s4), (s6) and (s7) from corpus AV1. The nine other sentences are given below (the number next to S/V/O is the number of syllables of the constituent):

(s9) [La nounou]$_{S3}$ [mariera]$_{V3}$ [Li]$_{O1}$

‘The nanny will marry Li.’

(s10) [La 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 nanny 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 lit the wheels again.’