Interaction of audition and vision for the perception of prosodic contrastive focus

Marion Dohen, Hélène Loevenbruck

To appear in Language and Speech, 52(2-3)

Interaction of audition and vision for the perception of prosodic contrastive focus

Running title: Auditory-visual perception of prosodic focus

Marion Dohen & Hélène Lœvenbruck

Speech and Cognition Department – GIPSA-lab

Grenoble, France

Corresponding author:

Marion Dohen

Speech and Cognition Department – GIPSA-lab

961 rue de la Houille Blanche

Domaine Universitaire – BP 46

38402 Saint Martin d'Hères Cedex

FRANCE

Tel: +33(0)4 76 57 48 50

Fax: +33(0)4 76 57 47 10

E-Mail: Marion.Dohen@gipsa-lab.inpg.fr

ABSTRACT

Prosodic contrastive focus is used to attract the listener's attention to a specific part of the utterance. Mostly conceived of as auditory/acoustic, it also has visible correlates which have been shown to be perceived. This study aimed at analyzing auditory-visual perception of prosodic focus by elaborating a paradigm enabling an auditory-visual advantage measurement (avoiding the ceiling effect) and by examining the interaction between audition and vision. A first experiment proved the efficiency of a whispered speech paradigm to measure an auditory-visual advantage for the perception of prosodic features. A second experiment used this paradigm to examine and characterize the auditory-visual perceptual processes. It combined performance assessment (focus detection score) to reaction time measurements and confirmed and extended the results from the first experiment. This study showed that adding vision to audition for perception of prosodic focus can not only improve focus detection but also reduce reaction times. A further analysis suggested that audition and vision are actually integrated for the perception of prosodic focus. Visual only perception appeared to be facilitated for whispered speech suggesting an enhancement of visual cues in whispering. Moreover, the potential influence of the presence of facial markers on perception is discussed.

KEYWORDS

speech, perception, auditory-visual perception, prosodic focus, whispered speech

INTRODUCTION

Prosodic information has 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. The study of language development has highlighted the bootstrapping role of prosody. It has been shown that newborns and infants are sensitive to prosodic information available in the ambient language and that prosodic cues are actually used in word segmentation by young infants (see *e.g.* Christophe *et al.*, 2003 for a review). On the production side, it has been shown that the control of fundamental frequency, intensity, lengthening and duration are acquired very early in the course of development (Konopczinsky, 1986; Kehoe *et al.*, 1995; Vihman, 1996).

Among the different aspects of prosody, prosodic focus deserves particular attention. It aims at highlighting 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 prosodic focus, contrastive informational focus, as described by Di Cristo (2000), 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 (capital letters signal focus):

– 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 *e.g.* for French: Séguinot, 1976; Touati, 1989; Morel &

Danon-Boileau, 1998; Rossi, 1999; Di Cristo, 2000; Touratier, 2000). Focus also has 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: Weintraub *et al.*, 1981; Baum *et al.*, 1982; Gussenhoven, 1983; Bryan, 1989; 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.

Although prosodic focus typically involves acoustic parameters, several works have suggested that articulatory – and more specifically visible lip and jaw motion – as well as other facial 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; Keating *et al.*, 2003; Dohen *et al.*, 2004; Cho, 2005; Beskow *et al.*, 2006). In previous studies (Dohen & Lœvenbruck, 2005; Dohen *et al.*, 2006), we conducted production analyses of the articulatory and other facial (eyebrow and head movements) correlates of prosodic contrastive focus in French. These studies showed that, in French, focus production is accompanied by the production of an articulatory contrast within the utterance. This contrast is produced using two main strategies (for details see Dohen *et al.*, 2006): either (a) lengthening and over-articulation of the focused constituent or (b) lengthening and over-articulation of the focused constituent plus under-articulation of the following constituent(s). Focus is also sometimes signalled by eyebrow raising and/or a head nod but the link is highly inter- and intra-speaker dependent. A number of studies also showed that there could be links between prosody in general and other facial movements such as eyebrow movements (Cavé *et al.*, 1996; Graf *et al.*, 2002; Keating *et al.*, 2003; Beskow *et al.*, 2006) or head movements (Hadar *et al.*, 1983; Graf *et al.*, 2002; Keating *et al.*, 2003; Munhall *et al.*, 2004, Beskow *et al.*, 2006). If such articulatory and facial visible 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). Parallel visual only perception studies (Dohen & Lœvenbruck, 2005) using a subset of the same data used for the production analyses described above, showed that the articulatory correlates which had been identified (over-articulation) were used for visual identification of prosodic focus in French. Other auditory-visual perception studies using animated talking heads (Dutch: Krahmer *et al.*, 2002a, 2002b; Krahmer & Swerts, 2006; Swedish: Granström & House, 2005) showed that facial cues such as eyebrow and/or head movements could play a role in the perception of prosodic focus. Swerts & Krahmer (2004, 2008) showed that the visual cues (articulatory and/or facial) produced by real speakers can influence the auditory-visual perception of prominence in Dutch. All these studies suggest that visual facial dynamics convey crucial prosodic information that may improve speech comprehension in conversational situations.

This paper investigates the potential contribution of visual information in the auditory-visual perception of prosodic contrastive focus in French. First, an experimental paradigm designed to test the contribution of visual information in prosodic focus detection is presented (whispered speech paradigm). Then two experiments are described with audiovisual data from several speakers producing prosodic focus in a natural conversational situation. The first one is a preliminary test of the whispered speech paradigm in a focus identification task on audiovisual whispered speech from two male speakers, with artificial facial markers on the eyes, lips and chin. Four focus conditions (neutral, subject, verb and object focus), two view points (front and profile) and three modalities (audio, visual and auditory-visual) were examined. The second experiment aimed at further testing how vision and audition interact in the perception of prosodic focus. Data from four speakers (two males and two females) were used as stimuli. In this second experiment, the speakers did not wear any facial markers. Two focus conditions, one view point (front), three modalities (audio, visual and auditory-visual) and two speech modes (whispered and normal speech) were studied. Focus detection scores and reaction times were measured. Results from the two experiments are presented and discussed in terms of the auditory-visual advantage relative to auditory alone perception for the detection of prosodic focus, in whispered speech. Interspeaker differences, the influence of artificial facial markers on visual

perception and the differences between normal and whispered speech are also discussed as well as the implications of these results for the better understanding of the auditory-visual perceptual processes involved.

MEASURING ENHANCEMENT OF PERCEPTION BY VISION FOR THE PERCEPTION OF PROSODIC FEATURES

The aim of this study was to test auditory-visual perception of French prosodic contrastive focus and to examine whether the visual modality can contribute to enhance perception *i.e.* if auditory-visual perception can be enhanced compared to auditory alone perception. Auditory perception of prosodic contrastive focus in French (as in other languages) is actually very good (close to 100% correct identifications of focus cases). This results in a ceiling effect and the impossibility to measure a potential significant improvement when the visual modality is additionally available. The solution therefore lies either in designing auditorily degraded prosodic stimuli or in designing audiovisual incongruent stimuli (the audio and visual streams carry contradictory information as in Swerts & Krahmer, 2004, 2008). For this study, we chose to use auditorily degraded stimuli for two reasons. The first one is that Swerts and Krahmer (2004, 2008) already conducted experiments using incrongruent audiovisual stimuli for the perception of prominence in Dutch and we found it interesting to use a different paradigm in order to compare the results. The second reason is that designing incongruent stimuli implies either manipulations of the audio and visual streams or asking speakers to produce very strange and difficult utterances which can result in very unnatural stimuli and therefore unecological experiments.

The classical paradigm used to put forward the role played by vision in speech perception when the audio signal is degraded is speech in noise (see *e.g.* Sumby & Pollack, 1954; Miller & Nicely, 1955; MacLeod & Summerfield, 1987). However even though adding noise to a signal reduces its lexical intelligibility, it does not greatly alter the global fundamental frequency (F0) contour. Miller and Nicely (1955) indeed showed that voicing was the most robust speech feature in noise. This is the reason why using a speech in noise paradigm to degrade auditory perception of prosodic features would probably be inefficient. Another possibility could have been to resynthesize the F0 contours in

order to "erase" the intonational cues to prosodic focus. However, in doing so, the naturalness of the stimuli is greatly altered (the listeners are under the impression that it is not a human being speaking). We therefore chose to use whispered speech for which, naturally, there is no intonational (F0) information since there is no vibration of the vocal folds. This does not mean that there is no prosody at all since the durational and intensity prosodic correlates are still present. In addition, as noted for instance by Higashikawa and Minifie (1999), listeners can perceive pitch during whispering, probably thanks to spectral cues to fundamental frequency. However, informal perceptual evaluation of prosodic focus produced in whispered speech showed that it is much more difficult to perceive prosodic focus auditorily from whispered speech than from normal speech. Moreover, whispering is used when one wants to be understood by the person he/she is speaking to but not overheard by others. It is possible that the speakers might then compensate the lack of auditory cues by emphasizing visual cues.

EXPERIMENT 1: TESTING THE WHISPERED SPEECH PARADIGM

Audiovisual recordings

Two male native speakers of French (thereafter referred to as SA and SB) were recorded. The recordings took place in a sound-attenuated room at the Speech & Cognition Department (GIPSA-lab). The speakers were seated in a chair. Their head was blocked in a helmet fixed to the wall behind them. They were filmed using two digital cameras (25 fps): one of them recording a front view of the speaker and the other a profile view. Spotlights (one above each camera) were used to optimize the quality of the video recordings. The field of view of the cameras was adjusted to the mid and lower parts of the face of the speakers (just above the eyes to below the chin: see Fig. 1). Sound was recorded using a microphone positioned in front of the speakers (about 50cm away) outside the field of view of the cameras. Fig. 1 provides an example of the images recorded. As annotated on the figure, the speakers wore several blue features (blue lipstick, eyeglasses with two blue markers, a blue ruler fixed on the glasses' side and a blue marker on the chin) in order to be able to use an automatic lip tracking device designed at Institut de la Communication Parlée (now Speech & Cognition Department, GIPSA-lab) (Lallouache, 1991; Audouy, 2000). This device is used to process the images

recorded as described above and automatically detect the lip contour and compute several articulatory parameters from both views (inter-lip area, protrusion, etc.). The results of these articulatory measurements will not be discussed here for the sake of clarity and conciseness.

****

Insert Figure 1 about here

****

Corpus. The corpus consisted of 4 sentences extracted from a corpus used in a previous study (Dohen *et al.*, 2004). The sentences all had a subject (S) – verb (V) – object (O) structure and consisted of words composed of CV syllables. They are listed below:

Romain ranima la jolie maman. ('Romain revived the good-looking mother.')

Véroniqua mangeait les mauvais melons. ('Véroniqua was eating the bad melons.')

Mon mari veut ranimer Romain. ('My husband wants to revive Romain.')

Les loups suivaient Marilou. ('The wolves were following Marilou.')

Focus elicitation procedure. The speakers were not directly asked to produce focus. A correction task was used in order to trigger focus in the most natural way possible. The speakers listened to a prompt in which two speakers (S1 and S2) were talking. S1 first pronounced a sentence from the corpus which described a situation. S2 then repeated the sentence in a question mode because he was not sure to have understood correctly. The recorded speaker then had to correct S2 and thus potentially produce contrastive focus. The recording therefore went as follows (capital letters signal focus):

Audio prompt: *S1:* Romain ranima la jolie maman. ('Romain revived the good-looking mother.')

S2: S1 a dit : Denis ranima la jolie maman?

('S1 said: Denis revived the good-looking mother?')

Speaker uttered: ROMAIN ranima la jolie maman. ('ROMAIN revived the good-looking mother.')

No indication was given to the speakers on how to produce focus (*e.g.* which syllable(s) was(were) to be focused). When S2 had correctly understood (he produced the correct sentence in a question mode), the recorded speaker produced a neutral version (broad focus) of the sentence *i.e.* without focusing any particular constituent.

Conditions. The speakers were asked to produce all the utterances using whispered speech. They were instructed to whisper to a person positioned behind the front camera (*i.e.* about 2 meters away) as opposed to whispering in someone else's ear. For each sentence from the corpus, four focus conditions were elicited: subject (SF), verb (VF) and object (OF) focus (narrow focus) and a neutral version of the sentence (broad focus). Each of the 16 possible items (4 sentences × 4 focus conditions) was recorded twice resulting in 32 utterances produced by each speaker.

Experimental method

Materials. A preliminary informal auditory only perception test was conducted to check whether the detection of focus using the whispered speech audio signals was low enough to be possibly improved. For SB, it appeared that the perceptual performances (focus identification score) were indeed quite poor for whispered speech but for SA, they were still quite high. An acoustic analysis showed that SA seemed to have over-compensated the lack of F0 cues by dramatically boosting intensity cues in whispered speech. It was therefore decided to weigh the intensities of all the utterances recorded. This was done by bringing the main intensity of each constituent (subject, verb and object) of the focused utterance to the same intensity level as that of the same constituent in the neutral version of the same sentence (uttered by the same speaker). After this processing, the acoustic information provided by duration and part of the information provided by intensity were still present to signal focus to the listeners. A second informal auditory only perception test however showed that, after such an acoustic processing, auditory only perception of focus was significantly degraded.

Three films (one for each condition: AV, A and V) were elaborated from the video clips designed by combining the videos recorded and the degraded auditory signals. Each film consisted of two sequences each corresponding to a random combination of the 32 available stimuli in a given modality (64 stimuli per film). The films were video-taped onto three VHS tapes. One of the sequences was to be seen with the front view and the other with the profile view.

Experimental procedure. The tests took place in a quiet room in the lab in which the participants were isolated both from outside noise and from the experimenters. The videos were shown, via a VCR, on a video monitor placed approximately one meter away from the participants. The speaker's head on the

screen was approximately life size. The participants responded on an answer sheet (see below) placed in front of them (between them and the video monitor) with a highlighter pen.

The participants were told that the videos (resp. sounds) they would see (resp. hear) were extracted from a dialogue of the following type. A first speaker (S1) uttered a subject-verb-object sentence which described a situation. The second speaker (S2), believing he had misunderstood part of the sentence (S, V or O), then questioned S1 by repeating the sentence he had understood, in a question mode. S1 then repeated the first sentence he had uttered correcting what S2 had misunderstood (S, V or O). An example of such a dialogue situation is provided below:

S1: Romain ranima la jolie maman.

S2: Denis ranima la jolie maman?

S1: ROMAIN ranima la jolie maman.

The participants were told that there were four possible dialogue situations:

- Situation A: S2 misunderstood S and S1 corrected S (subject focus case);
- Situation B: S2 misunderstood V and S1 corrected V (verb focus case);
- Situation C: S2 misunderstood O and S1 corrected O (object focus case) ;
- Situation D: S2 understood well and S1 repeated the sentence in a neutral mode (neutral case).

Participants were told that they would only see (or hear or both hear and see) the second utterance produced by S1. Their task was then to infer the dialogue situation (A, B, C or D) by highlighting the potentially corrected constituent on an answer sheet such as the one presented below (empty column on the right for D-type situations):

Romain	ranima	la jolie maman.	

The purpose of such a procedure was to indirectly ask participants to detect and identify focus cases. They were never told about “prosodic contrastive focus” or about what the experiment aimed at testing.

Participants were randomly assigned to one of two separate tests (test a or test b) corresponding to exactly the same data set in different orders. For test a, participants were first tested audiovisually,

then auditorily and finally visually (AV – A – V). For test b, they were first tested auditorily, then audiovisually and finally visually (A – AV –V). The aim being to analyze the contribution of the visual modality, this paradigm allowed comparisons between the performances corresponding to the auditory only and auditory-visual conditions. For test a, for example, there could be a training effect during the auditory-visual session which would affect the performances during the auditory only session and vice versa. This is why two tests were necessary. The visual only perception condition represented a control. Before taking the test in the different conditions, the participants went through a short practice session. A total of 32 stimuli were thus evaluated by the subjects under three conditions (A, V and AV) and two different views (front and profile). This represents a total of 192 stimuli. The experiment lasted approximately 35 minutes.

Participants. A total of 13 native speakers of French took part in the test (8 men and 5 women). They were aged 30.7 on average (from 19 to 57 years old) and originated from various regions of France. All of them reported normal or corrected to normal vision and no auditory problems. After the test, each participant was asked to describe the task he/she had performed to check whether he/she had understood well. It appeared that all the participants had correctly understood the task.

Results

General analysis. Fig. 2 provides the percentages of correct answers for each participant and for each modality (A, V and AV). The means over all the participants for each modality are represented by big dark bars and are the following: AV: 74%, A: 64.9% and V: 61.8%. All the results are significantly above chance (25% in this case) meaning that whatever the modality considered, participants identified the focus situation above chance level. It was checked that there was no order effect.

****

Insert Figure 2 about here

****

This experiment has a $3 \times 2 \times 2 \times 4$ design with the following within subject factors: modality (3 levels: AV, A and V), speaker (2 levels), view point (2 levels: front and profile) and focus condition (4 levels: neutral, SF, VF and OF). A four-way repeated measures analysis of variance was conducted on

the percentages of correct answers with the above within subject factors as independent factors. The sphericity of the data was checked for using the Mauchly sphericity test. When the test was significant, the Huynh-Feldt correction was applied on the number of degrees of freedom and all the results presented below correspond to these corrected results (when necessary). For the sake of clarity, even when the results were corrected, they will be reported with the "true" numbers of degrees of freedom. The results of multiple pair-wise comparisons were corrected for using the Bonferroni correction. This statistical analysis revealed significant main effects of modality ($F(2,24)=7.232$, $p=.003$) and speaker ($F(1,12)=121.384$, $p<.001$). Neither the view nor the focus condition had a significant effect on focus identification (view: $F(1,12)=0.244$, $p=.63$; focus condition: $F(2,36)=11.231$, $p=.096$). For the effect of modality, analysis of the ANOVA contrasts showed that the AV performances were significantly higher than those corresponding to the other modalities ($p<.001$). The results corresponding to the auditory and visual only modalities did not differ significantly ($p=.451$). As for the effect of speaker, it is due to the fact that performances were globally better for SA's production (this effect will be analyzed in more detail below). This general analysis showed that focus identification was significantly better in the AV condition meaning that, for whispered speech, it is easier to detect focus and identify its location when both the auditory and visual modalities are available compared to both modalities taken separately.

Inter-speaker analysis. As mentioned before, there is a significant main effect of speaker. The ANOVA also puts forward a significant interaction between modality and speaker ($F(2,24)=11.231$, $p<.001$). Fig. 3 provides further illustration of this interaction by presenting the percentages of correct answers by modality for each of the two speakers.

****

Insert Figure 3 about here

****

Fig. 3 shows that the auditory-visual advantage (deviation between the scores for the AV and A modalities) is smaller for speaker A than for speaker B (SA: +3.9%; SB: +14.4%). Also note that, for SA, the performances are better in the A modality than in the V modality whereas the contrary can be

observed for SB. This explains the significant interaction between modality and speaker reported above. Actually, the auditory only performances measured for SA are very good (80.8% correct answers) and much better than those measured for SB (63.5%). It is possible that there is a ceiling effect for SA: the performances are too good in the auditory only condition to be improved to a significant extent. This is probably due to the fact that SA was a trained speaker. The acoustic cues he produced were very strong even after weighing the intensities. In addition, over-articulation (one of the correlates of prosodic focus as presented in the introduction) has visual as well as acoustic consequences: formant patterns are less reduced for instance, which may be heard (see Lœvenbruck, 1996 for the difference in auditory perception between hypo and hyper-articulated /iai/ sequences). This may have slightly biased the results and probably explains the small visual supplement measured for SA. SB on the contrary was a naïve speaker.

Summary & Discussion

This experiment showed that it is possible to measure an auditory-visual advantage for whispered speech. When the acoustic prosodic cues were degraded (no intonation and normalized intensity), visual information helped recover prosodic information. Adding vision indeed improved overall detection and identification of focus cases. The results also showed that even if auditory only perception of prosodic information was degraded for whispered speech, it was still above chance. This was predictable since, as explained before, the durational acoustic cues remain available in whispered speech. It is also possible that the intensity cues were not entirely "erased" and that other acoustic prosodic cues (*e.g.* spectral) remained. Other important observations are that the focus condition (neutral, subject focus, verb focus or object focus) and view of the speaker (front or profile) did not have a significant effect on the results. Note that, for this experiment, speakers wore facial markers for parallel production analyses (not reported here). This may have had an influence on the results in, for example, attracting the perceivers' attention to important visual correlates of prosodic focus such as over-articulation.

EXPERIMENT 2: TESTING HOW VISION AND AUDITION INTERACT FOR THE PERCEPTION OF PROSODY

Aim

The aim of this experiment was to use the whispered speech paradigm (validated in experiment 1) to explore the interactions between audition and vision for the perception of prosodic contrastive focus in French. Particularly, we included a normal (ordinary) speech condition in order to directly compare normal and whispered speech. In line with the results from experiment 1, the main prediction was that, for whispered speech, the performances in the auditory-visual modality would be better than those in both the auditory and visual modalities taken separately whereas, for normal speech, there would be no difference between the AV performances and the A ones (ceiling effect). Another purpose was to compare the performances in the V modality for normal and whispered speech. This experiment also aimed at evaluating the cognitive load (or more precisely the duration of the cognitive processes) involved for one modality *vs.* auditory-visual perceptions of prosodic focus: does adding the V modality facilitate cognitive processes? In order to assess this point we used parallel measurements of perceptual performances (percent correct focus detection, as in experiment 1) and reaction times. Since experiment 1 revealed a difference between the results for the two different speakers, the aim here was also to further analyze this potential inter-speaker variability. Lastly, as mentioned before, experiment 1 used stimuli in which the speakers wore several facial markers and we wanted to test if this may have had an influence.

Audiovisual recordings

Four native speakers of French (2 women and 2 men), hereafter referred to as SB, SC, SD and SE, were recorded for designing the stimuli for experiment 2. None of the speakers wore glasses and the male speakers were clean shaven. The recordings took place in a sound-attenuated room at the lab. The speakers were seated on a chair against a uniform background. The video was recorded using a digital camera positioned about 1.5 meters in front of them (25 fps). We did not record a profile view of the speaker since experiment 1 showed that the view point did not affect the results. A spotlight was

positioned above the camera in order to get optimal video recordings. The field of view was adjusted to film the speakers from just below the neck to just above the top of the head. Sound was recorded using a microphone positioned in front of the speakers (about 50cm away) outside the field of view of the camera. Fig. 4 shows an example of the images recorded.

****

Insert Figure 4 about here

****

Corpus. Considering the great number of conditions to test (see below) and the fact that the experiment should not last too long for the participants, the corpus consisted of only two sentences. These sentences had a subject (S) – verb (V) – object (O) structure and consisted of exactly the same number of syllables (S: 2-syllable first name; V: 2-syllable action verb; O: 1-syllable determiner followed by a 2-syllable common noun). The two sentences were the following:

Manon coupait le melon. ('Manon was cutting the melon.')

Ninon lançait la toupie. ('Ninon was throwing the spinning top.')

An extra sentence with exactly the same structure was recorded in all conditions as a practice item for the perception test.

Experimental conditions. The sentences were recorded for two speaking modes: normal speech and whispered speech. The task for the whispered speech condition was to whisper to a person located behind the camera (*i.e.* about 2 meters away) as opposed to whispering in someone's ear. A total of three focus conditions were elicited: broad focus (neutral, BF), subject focus (SF) and object focus (OF). The procedure for eliciting focus productions from the speakers was the same as the one described for the recordings corresponding to experiment 1 (see above). Each of the 12 possible items (2 sentences × 2 speech modes × 3 focus conditions) was recorded twice resulting in 24 utterances recorded for each speaker.

Experimental method

Materials. The auditory only (A) and visual only (V) streams were extracted from the audiovisual (AV) recordings. An informal perceptual validation was conducted on the A stimuli in order to check

whether focus had been produced correctly. On the basis of this validation, a selection was made between the two available versions of SF and OF utterances. Since the results from experiment 1 showed that focus location did not have a significant effect, the SF and OF cases were mixed into a single 'focus' condition (*vs.* 'no focus': BF). After this selection, a total of 64 stimuli per modality (A, V and AV) were available (4 speakers × 2 sentences × 2 speech modes × (2 BF + 1 SF + 1 OF) focus conditions).

Experimental procedure. The experiment took place in a quiet room in which participants were isolated both from outside noise and from the experimenters. The experimental procedure used was very similar to the one used for experiment 1. Participants were told beforehand that they would be viewing / hearing / viewing and hearing part of a dialogue and that their task would be to identify whether the speaker they had viewed / heard / viewed and heard had performed a correction or not. As explained above, unlike for experiment 1, we did not ask the participants to identify which part of the utterance had been corrected but only whether any part of the utterance had been corrected at all. The participants saw the video clips on a 1440×900 resolution PC screen and sound was played to them through stereo headphones. The stimuli were presented using Neurobehavioral Systems' Presentation software which enables precise measurement of reaction times. Participants were instructed to use the mouse's right and left buttons to respond and the middle button to go on to the next stimulus. Correspondence between button (right and left) and response type (correction and no correction) was maintained constant for each participant and randomly varied across participants. They were told to click on the button corresponding to their response as soon as they had made a decision and even if the speaker had not finished talking. The reaction times were measured relatively to the stimulus onset. The experiment was divided into 6 sessions, each one consisting of a combination of one speech mode and one modality (normal × A, normal × V, normal × AV, whispered × A, whispered × V, whispered × AV). The order in which participants were tested for each of these sessions was randomized across participants. Each session consisted of 32 stimuli which were presented in a random order. Before each stimulus, the sentence corresponding to the one pronounced in the stimulus was displayed on the screen so that the participants could read it. This was done so that they knew what was going to be

said and could concentrate on the task and not on the segmental content (especially in the visual only condition). Before each session, the participants briefly practised in the corresponding condition combination (4 stimuli corresponding to the practice sentence uttered by each one of the 4 speakers in varying focus conditions). During the practise sessions, no feedback (correctness) was given to the participants. After the test, participants were asked to fill-in a questionnaire in which they provided information on their perception of the difficulty of the task in the different conditions, on the compared intelligibility of the different speakers and on the cues (visual and auditory) they used to perform the task. Each participant evaluated a total of 192 stimuli (excluding practise) and the experiment lasted approximately 35 minutes (including questionnaire).

Participants. A total of 31 native speakers of French (16 women and 15 men) volunteered to take part in this experiment. Participants were aged 29.6 on average (from 21 to 52 years old). They were either right-handed or reported that they always use the mouse (response device) with their right hand. They all had normal or corrected to normal vision and no hearing problem.

Results – Perceptual performances: focus detection scores

This section will describe the results corresponding to the perceptual performances *i.e.* the percentages of correct focus detections.

General description. Table 1 provides the mean percentages of correct answers (focus condition identified correctly) for each speech mode, modality, focus condition and speaker. It turns out that whatever the condition, the mean percentages of correct answers were well above chance (50% in this case). This experiment has a $2 \times 3 \times 2 \times 4$ design with the following within subject factors: speech mode (2 levels: normal, whispered), modality (3 levels: AV, A, V), focus condition (2 levels: focus, no focus) and speaker (4 levels). A four-way repeated measures analysis of variance was conducted on the percentages of correct answers with the above within subject factors as independent factors. The sphericity of the data was checked for using the Mauchly sphericity test. When the test was significant, the Huynh-Feldt correction was applied on the number of degrees of freedom and all the results presented below correspond to these corrected results (when necessary). However, for the sake of clarity, the degrees of freedom reported are the "true" ones even when a correction was actually

applied. The results of multiple pair-wise comparisons were corrected for using the Bonferroni correction.

We found significant main effects of speech mode ($F(1,30)=29.922$, $p<.001$), modality ($F(2,60)=90.406$, $p<.001$), focus ($F(1,30)=9.376$, $p=.005$) and speaker ($F(3,90)=8.084$, $p<.001$). For the speech mode, analysis of ANOVA contrasts showed that the percentage of correct answers was significantly ($p<.001$) greater for normal speech than for whispered speech. This simply confirmed that, focus is better detected in the normal speech condition overall. For modality, analysis of contrasts showed that the percentage of correct answers was greater in the AV condition than in the A and V conditions ($p<.001$) and in the A condition compared to the V condition ($p<.001$). For focus, analysis of contrasts showed that the percentage of correct answers was significantly ($p=.005$) greater for the focus (*vs.* no focus) condition. This effect will be analyzed into detail in a following section as well as the differences observed in the perceptual performances corresponding to the productions of the different speakers (main effect of speaker). The significant interactions observed with the ANOVA will also be discussed in the following sections.

****

Insert Table 1 about here

****

Inter-modality analysis. Fig. 5 provides the mean percentages of correct answers over all the participants by speech mode (normal and whispered) and modality (AV, A and V): normal: AV: 97.4%, A: 95.9%, V: 79%; whispered: AV: 92.6%, A: 85.5%, V: 84.1%. Apart from the main effect of modality, the ANOVA also revealed a significant interaction between speech mode and modality ($F(2,60)=39.711$, $p<.001$). This interaction illustrates the fact, that for normal speech, the performances were not significantly different in the AV and A conditions ($p=.043$) whereas, for whispered speech, the percentages of correct answers were significantly higher in the AV condition than in both the A (+7.2%) and V (+8.6%) conditions ($p<.001$ for both comparisons). Note that there was a significant ($p<.001$) difference between the auditory only performances for normal speech and for whispered speech: 10.4% decrease in the focus detection score from normal to whispered speech.

The auditory only perception of focus was thus more difficult for whispered speech. On the other hand, there was a significant (p=.005) increase (+5%) of the focus detection score from normal to whispered speech in the V condition. It was therefore easier to detect focus from the visual modality alone in whispered speech than in normal speech.

****

Insert Figure 5 about here

****

Inter-focus condition analysis. Apart from the main effect of focus, the ANOVA also showed a significant interaction between speech mode and focus condition ($F(1,30)=45.373$, $p<.001$). For normal speech, the two focus conditions did not differ significantly (p=.126). However, for whispered speech, the performances were significantly ($p<.001$) better in the focus condition than in the no-focus condition. Moreover, there was also a significant interaction between modality and focus condition ($F(2,60)=10.908$, $p<.001$). For both AV and A modalities, the performances were significantly better in the focus condition ($p<.001$ for both comparisons) whereas for the V modality there was no significant difference between the two focus conditions (p=.85). Table 2 provides the number of responses (either correct or not) corresponding to each focus condition by speech mode and modality. It shows that for conditions that lead to better performances in the focus condition, *i.e.* for whispered speech over all modalities and for AV and A modalities over all speech modes, the 'focus' responses were simply more frequent. In particular, the whispered speech condition might have induced more 'focus' responses, because the stimuli were more carefully pronounced, *i.e.* always slightly over-articulated. Since over-articulation is a visual cue to focus, this careful pronunciation may have induced wrong focus detections and resulted in a better response accuracy for the 'focus' condition. For normal speech, the number of responses was approximately the same for both focus conditions except in the V modality for which 'no focus' responses were more frequent. This is in line with the fact that focus appeared to be difficult to detect in the V modality. For whispered speech, 'focus' responses were always more frequent. This was especially true for the A modality probably due to the fact that whispered speech results in more intensity variations than normal speech (as specified for

Exp. 1 and as observed through informal measurements). These unusual intensity variations may have been mistaken for focus correlates.

****

Insert Table 2 about here

****

Inter-speaker analysis. In addition to the main effect of speaker, the ANOVA revealed a significant interaction between modality and speaker ($F(6,180)=6.831$, $p<.001$). Fig. 6 shows the mean percentages of correct answers over all the participants by modality and speech mode for each speaker. The interaction between speaker and modality results from the fact that, for SC, SD and SE, focus detection scores across speech modes were greater in the AV modality than in the A modality and in the A modality than in the V modality (AV>A>V) whereas, for SB, performances were the best in the A modality (A>AV>V). There was also a slightly significant interaction between speech mode and speaker ($F(3,90)=3.978$, $p=.01$). This is due to the fact that overall focus detection scores were higher for normal speech than for whispered speech except for SC (for whom we observed approximately the same mean percent correct in both speech modes). The two latter interactions are tightly linked to the significant three-way interaction between speech mode, modality and speaker ($F(6,180)=11.866$, $p<.001$). For normal speech and for all speakers, the performances in the AV and A conditions were equivalent and better than in the V modality (AV≈A>V). For whispered speech, for SC, SD and SE, the pattern was approximately the same: the performances were the best for the AV modality and slightly better or equivalent in the V modality compared to the A modality (AV>V≻A). Note that for SD, the results in the V modality were much better than those in the A modality. However, for SB, a completely different pattern was observed for whispered speech: the pattern observed for normal speech (AV≈A>V) was actually replicated. The ANOVA also revealed a significant interaction between speaker and focus condition ($F(3,90)=10.479$, $p<.001$). This illustrates the fact that, for speakers SB, SC and SD, the performances were better in the focus than in the no-focus condition (same as the general trend described above). For speaker SE, the reverse pattern was observed since the percentage of correct answers was higher in the no-focus condition. This is reflected by the fact

that SE was the only speaker for whom more 'no focus' responses were provided than 'focus' responses ('no focus': 759; 'focus': 729). Following this observation, a significant three-way interaction was also measured between speech mode, focus condition and speaker ($F(3,90)=5.579$, $p=.001$). Actually, the only exceptions to the general pattern (focus > no focus) are observed for speakers SD and SE in the normal speech mode (no focus > focus). Also note that for all speakers except SB, performances in the V only condition improved from normal speech to whispered speech (same as the general pattern). Looking at the results into more details showed that it appeared to be easier to visually detect focus from speakers SD and SE.

****

Insert Figure 6 about here

****

Inter-stimulus analysis. A further analysis was conducted aiming at comparing the perceptual results corresponding to the different stimuli. The mean percentage of correct answers was computed for each stimulus across participants by modality. The aim was to analyze each stimulus according to its characteristics in terms of acoustic and visual cue value. What is reported here are the results of a qualitative analysis. Four main cases were identified:

- case 1: the one modality scores (A and V) were both excellent (more than 80%): this was the case for 56.3% of the stimuli;
- case 2: the one modality scores ranged from 50% to 80%: 7.8% of the stimuli;
- case 3: one of the one modality score was very good (more than 80%) and the other one modality score ranged from 50% to 80%: 26.6% of the stimuli;
- case 4: one of the one modality scores was poor (less than 50%): 9.4% of the stimuli.

For case 1., the corresponding auditory-visual score observed was also very good (more than 80%) and not systematically better than the one modality scores. For case 2., the auditory-visual score was very good (more than 80%) for normal as well as whispered speech. This suggests that the combination of the information from both modalities resulted in a perceptual enhancement. For case 3., the auditory-visual score approximately corresponded to the best one modality score. This suggests

that the information provided by the modality for which the detection score ranged from 50% to 80% did not enhance perception which is probably simply linked to a ceiling effect (nothing to be improved). For case 4., the pattern was different for normal speech than for whispered speech. For normal speech, the corresponding auditory-visual score was very high (more than 80%). For whispered speech, the auditory-visual score was lower than the highest one modality score. This case (case 4) corresponded to a case for which one of the one modality scores was below chance level (50%) meaning that the corresponding information actually appeared to be misleading. A score of 50% (chance level) indeed corresponds to an absence of information from the considered modality but a score below chance reflects a systematic trend towards a misinterpretation. This suggests that some cues were present that led the participants towards an interpretation that is contradictory with the information the speaker intended to transmit.

Questionnaire analysis. The answers to the questionnaire indicated that the test was considered as easiest in the auditory-visual modality by 68% of the participants, followed by the auditory modality and the visual modality (AV > A > V). Only 16% of the participants found the test to be easier in the visual modality compared to the auditory one (AV > V > A) and 13% preferred the auditory modality overall (A > AV > V). In addition, to the question "Did you find the test really difficult for one condition (AV, A or V)?", 95% of the participants responded that they found the test really difficult in the visual only condition . These general tendencies appear to be in line with the focus detection scores analyzed above (perceptual performances). Moreover, the test was considered as easier for normal speech than for whispered speech by 82% of the participants, which is also consistent with the perceptual results described previously. The participants were also asked to rank the speakers along who was easier to understand. The most frequent order of preference was: SB > SD > SE > SC even though many variations across participants were observed for this classification. This approximately corresponds to the overall results for each speaker.

Results – Reaction Times

This section describes the analysis of the reaction time (RT) measurements. A preliminary screening of the data was performed in order to detect and exclude potential outliers. These were identified

separately for each stimulus (1 sentence, 1 focus condition, 1 speaker) since different segmental contents and focus conditions as well as productions from different speakers are expected to result in different reaction times. For each stimulus, data corresponding to reaction times at a distance of at least 2 standard deviations from the overall mean for the specific stimulus considered were identified as outliers and replaced with the overall mean. In order to "filter" variation in reaction times due to variation of segmental content (from one sentence to another) or of production (by the same or different speakers) and since our interest was to analyze differences in reaction times between different modalities, we pre-processed the data as follows. Reaction time differences between modalities were computed for each stimulus corresponding to exactly the same segmental content (same sentence) and exactly the same production (same speaker, same focus condition). This resulted in the computation of three differences for each stimulus: RT_{AV}-RT_A, RT_{AV}-RT_V and RT_V-RT_A. In particular, the first two differences should reveal potential speeding up of reaction times from one modality conditions (A and V) to the auditory-visual condition (AV). The absolute RT values are not provided since their interpretation is difficult. As explained above, variations in absolute RT originate from multiple causes: change in segmental content, change in production style…

General description. Table 3 provides the detailed differential RT measurements: mean RT differences over all the participants by speech mode (normal and whispered), modality difference (AV - A, AV - V and V - A), speaker (SB, SC, SD and SE) and focus condition (focus and no focus). As already detailed in the performance analysis (see 'perceptual performances' section), the experimental design is a 2 × 3 × 2 × 4 design with the following within subject factors: speech mode (2 levels: normal, whispered), modality (3 levels: AV - A, AV - V, V - A), focus condition (2 levels: focus, no focus) and speaker (4 levels). The same repeated measures analysis of variance was conducted as that conducted on the percentages of correct answers with the same factors and performing the same types of corrections (Mauchly sphericity test and Huynh-Feldt correction on the degrees of freedom; Bonferroni correction for multiple pair-wise comparisons). The ANOVA revealed a significant effect of modality difference ($F(2,60)=31.271$, $p<.001$). None of the other factors had significant effects (speech mode: $F(1,30)=.055$, $p=.816$; speaker: $F(3,90)=.725$, $p=.54$; focus

condition: F(1,30)=.091, p=.765). For the modality difference, analysis of the ANOVA contrasts showed that the RT_{AV}-RT_V differences were the greatest compared to the other RT differences (RT_{AV}-RT_V > RT_A-RT_V >RT_{AV}-RT_A).

****

Insert Table 3 about here

****

Since the reaction time measurements were primarily conducted to test focus detection speed, in the following analyses, we will mainly discuss the results corresponding to the focus condition. The no focus condition provides less information on detection speed since, when participants make a 'no focus' judgment, it appears that they wait until the end of the utterance probably to be sure that there will actually be no focus. Moreover, the basic point being to study auditory-visual perception compared to one modality perception, we will mainly discuss the AV - A and AV - V differences.

Detailed analysis. The ANOVA revealed a significant interaction between speech mode and modality difference (F(2,60)=5.75, p=.005) as well as between speaker and modality difference (F(6,180)=3.377, p=.004) and between focus condition and modality difference (F(2,60)=12.352, p<.001). Fig. 7 provides the mean reaction time differences corresponding to the focus condition for the AV - V and AV - A modality differences and for both speech modes. It illustrates the interaction between speech mode and modality. For normal speech, in the focus condition, the reaction times significantly decreased (comparison to 0) from the visual only to the auditory-visual condition (normal: RT_{AV} -RT_V <0, p<.001) but not from the auditory only to the auditory-visual condition (RT_{AV} -RT_A not different from 0, p=.284). For whispered speech, responses were significantly faster for the auditory only compared to the auditory-visual condition (whispered: RT_{AV} -RT_A<0, p=.013) as well as for the visual only compared to the auditory-visual condition (whispered: RT_{AV} -RT_V<0, p<.001). What is clearly visible on Fig. 7 is that, on the one hand, the reaction time advantage from A to AV increases in the whispered speech condition whereas, on the other hand, this advantage decreases from V to AV in the whispered speech condition. Participants respond faster in the AV condition compared

to the A and V conditions for whispered speech. For normal speech, there is no AV advantage compared to A but for V, there is a larger one than for whispered speech.

****

Insert Figure 7 about here

****

Fig. 8 provides the same information as Fig. 7 but for each speaker individually. The individual patterns correspond to the general pattern described previously except for speaker B. For this speaker there was no difference between the AV advantage for reaction time for normal and whispered speech. Note that, for focus detection scores, there was no difference across speech modes for this speaker. What is actually surprising is that even though the performance results show that there was no AV advantage (compared to A) for this speaker, there was an AV advantage as far as reaction times are concerned: responses were quicker in the AV condition than in the A condition. The fact that there was no true difference between the AV advantage (relative to V) in normal and whispered speech confirms that, unlike for other speakers, visual only perception was not advantaged for whispered speech compared to normal speech. For speaker E, it appears that, for whispered speech, the AV advantage is the same (and not very strong) relative to the A and V modalities. This suggests that, unlike for the other speakers, the cognitive advantage when adding audition to vision was not very strong for this speaker. The reasons for this observation remain unclear.

****

Insert Figure 8 about here

****

Summary & Discussion

This experiment replicated and extended the results from Experiment 1. It additionally enabled the comparison of normal speech and whispered speech, the analysis of inter-speaker variability and the joint analysis of focus detection scores and reaction times. The results confirmed that it is possible to measure an auditory-visual advantage for the perception of prosodic features using a whispered speech paradigm. The AV performances were significantly better than both the A and the V performances for

whispered speech. This experiment showed that such a measurement is not possible for normal speech because of a ceiling effect (A only performances: 96% correct answers). The results in the A only condition for whispered speech (85% correct answers) confirmed that there are still some acoustic cues to prosodic focus in whispered speech since the mean percentage of correct answers was well above chance (50%). Moreover it appeared that it is easier to perceive focus auditorily for normal speech than for whispered speech. Unlike for normal speech, however, the performances in the A and V conditions were not significantly different. The results in the V only condition showed that it is easier to perceive focus visually for whispered speech than for normal speech. The analysis of inter-speaker differences revealed that the global pattern described above approximately corresponded to the individual patterns except for SB. The extent to which prosody can be perceived from whispered speech both auditorily and visually however seems to depend on the speaker. This is also the case for normal speech especially for the V modality. The analysis of the feedback provided by the participants via the questionnaires showed that their impressions concerning the difficulty of the task for the different speech modes and modalities were consistent with the actual focus detection scores measured.

The perceptual performances evaluated via the percentages of correct focus detections were also related to reaction time measurements. This analysis showed that there was no significant difference measured for reaction times from the A to the AV modalities for normal speech showing no gain in the duration of cognitive operations when the V modality was added to the A one for normal speech. This is consistent with the fact that adding the V modality for normal speech did not lead to an AV advantage as far as perceptual performances were concerned. This observation however has to be qualified since, for SB, an advantage was measured for reaction times even though no performance advantage was measured. This shows that, at least in some cases, even when no perceptual advantage was measured, adding the V modality to the A one can speed up responses suggesting a reduction of the cognitive load. For whispered speech, not only were the focus detection scores significantly higher when the V modality was added to the A one, it also appeared that adding the V modality reduced the duration of cognitive processes (responses were faster: RT_{AV}-RT_A<0). Another important indication was provided by the measurements of RT_{AV} -RT_V for whispered speech compared to normal speech.

For both speech modes, a quite large advantage in reaction times was measured from the V to the AV modality. This is quite trivial since evaluating speech from the V modality alone is not a natural task and is quite difficult. The interesting point is that this advantage was smaller for whispered speech (RT_{AV} -RT_{V} normal > RT_{AV} -RT_{V} whispered). In line with focus detection scores (better for V alone in whispered speech), this suggests that it is easier to perceive focus from the visual modality alone from whispered speech than from normal speech.

GENERAL DISCUSSION

This article presented two experiments aiming at evaluating the auditory-visual perceptual processes involved in the perception of contrastive prosodic focus in French. Previous research has already shown that there are prosodic visual cues and that these are perceived visually (see introduction for more details). It has also been shown that different visual features may have different cue values (Swerts & Krahmer, 2008). However, it still remains unclear how the auditory and the visual modalities interact in auditory-visual perception of prosodic features. The experiments presented here aimed at measuring how vision can enhance auditory perception of prosodic features and at evaluating the process (combination or integration) at stake. The first challenge was to design a paradigm which would make it possible to measure a potential auditory-visual advantage *i.e.* an increase in perceptual abilities when both the auditory and the visual modalities are available rather than only one of them. There is indeed a ceiling effect on auditory only perception of prosodic focus. In order to do so, we decided to use whispered speech for which part of the prosodic information is absent (namely intonation: no fundamental frequency). Experiment 1 aimed at testing the paradigm and evaluating the effect of different focus conditions and different views of the speakers. It concerned only the lower face (mainly articulation of the lips and jaw) which has been shown to be significantly visibly influenced by prosodic focus production (Dohen *et al.*, 2006). In this experiment, the speakers wore facial markers (for production studies not reported here) and their head was maintained still (no head movements possible). Experiment 2 aimed at using the paradigm to perform comparisons of the perception of normal and whispered speech for different speakers. The performance measurements (percentage of correct answers) were combined to reaction time measurements in order to further

study and characterize the processes involved. In this second experiment, the speaker's entire face was visible and the speakers wore no facial markers and were free to move their heads. For both of these experiments, we used a special focus elicitation procedure (conversational situation) in order for the stimuli to be as natural as possible.

Experiment 1 validated the whispered speech paradigm for measuring an AV advantage for the perception of prosodic features. It showed that the auditory only perception of prosodic focus for whispered speech was still possible (identification results above chance) but degraded. This degradation made it possible to measure a potential increase in the performances when the visual modality was additionally available. Experiment 1 showed that such an improvement exists and is significant. It also showed that perceptual performances are independent of the focus condition: results do not differ significantly when focus is on the subject, the verb or the object of the sentence. Moreover auditory-visual perception is independent of the viewpoint from which the speakers are observed: results did not differ when the speakers were viewed from profile and when they were viewed facewise. The fact that vision facilitated overall perception is in line with studies showing that speech comprehension is enhanced by vision for the perception of speech in noise (*e.g.* Sumby & Pollack, 1954; Miller & Nicely, 1955; MacLeod & Summerfield, 1987) or semantically complex speech (Reisberg *et al.*, 1987). However, what this study additionally showed is that vision does not help only for the perception of the segmental content but also plays a part in the perception of prosodic features such as contrastive prosodic focus.

Experiment 2 confirmed these results and provided more detailed information on the perceptual cognitive processes involved. First, it confirmed the fact that auditory only perception of prosodic focus is nearly perfect for normal speech leading to a ceiling effect and is degraded for whispered speech. These results are also in line with those from Swerts & Krahmer (2004) which showed that prosodic prominence was very well perceived auditorily for normal speech, rendering AV improvement measurements impossible. Experiment 2 also confirmed the fact that an AV advantage can be measured for whispered speech. Reaction time measurements additionally showed that when acoustic cues are not sufficient (whispered speech) adding vision significantly speeds up responses.

Not only is focus detection more accurate when vision is available, but it is also quicker probably reflecting a reduction of the cognitive load involved in processing prosodic information when vision is added to audition. In some cases (for speaker B), it appeared that adding vision speeded responses even when focus detection was not affected (due to a ceiling effect). This suggests that even when accuracy cannot be improved, the duration of underlying cognitive operations can be reduced when vision is added to audition.

In the light of the results from experiments 1 and 2, we would also like to discuss how the auditory and the visual cues "interact" in the perception process. For speech comprehension in general (perception of the segmental content of utterances), many studies show that vision is not simply superimposed on (or added to) audition for speech perception but rather that the two modalities are integrated. The clearest and most famous example of this is the Mc Gurk effect (McGurk & MacDonald, 1976) in which an auditory /ba/ combined with a visual /ga/ results in a /da/ percept. However, it is unclear whether the same processes are involved for the perception of prosodic information. Swerts and Krahmer (2004, 2008) designed a paradigm to test this for the perception of prosodic prominence. They designed stimuli for which the visual and acoustic cues to prominence were incongruent (for example: auditory "accent" on one word but visual cues to "accent" on another word). Their results tend to show that acoustic cues have a stronger "value" than visual cues as far as prosodic information is concerned even if visual cues can also have an influence: the acoustic "accent" attracts perception most of the time but if perception is wrong it is most often towards the word bearing the visual "accent". Results from our experiment 2 (see inter-stimulus analysis in particular) are not really in line with these observations. We found that when visual only perception is really poor (below chance level), meaning that the visual cues are misleading (and not simply absent), auditory-visual perception can become lower than auditory only perception. If auditory cues were stronger, this should not happen and auditory-visual perception should be equivalent to auditory only perception (the misleading visual cues should be ignored). Of course it must be born in mind that not many stimuli correspond to this particular case in our experiment so we cannot draw general conclusions. These cases however suggest that audition and vision are actually integrated for the perception of

prosodic features. Moreover, the fact that the duration of cognitive operations was reduced when vision was added to audition also suggests that the two modalities are rather integrated. If this was not the case and if perceivers mainly relied on acoustic cues, one would expect that either the responses would not be quicker when the visual modality was added or that the reaction times may even be increased because of the added time necessary for processing an extra modality and combining the information to make a decision. The fact that our results suggest a different pattern than that observed by Swerts and Krahmer (2004, 2008) which suggested that the auditory modality was predominant and that the visual modality was only added to it as extra information, may however be explained in the following manner. The paradigm used by Swerts and Krahmer (2004, 2008) cannot be strictly identified to the Mc Gurk effect paradigm and may therefore not be suitable for evaluating auditory-visual integration processes. In their paradigm, the auditory-visual discrepancies were clearly identifiable by the listeners/viewers which is not the case for the Mc Gurk effect (most of the time participants have no idea that the stimuli presented are actually incongruent). One can assume that in cases for which the incongruency is clearly identifiable, participants may indeed rely more on audition because speech is more often perceived through the auditory modality alone than through the visual modality alone rendering the auditory modality more natural for perceiving speech. In our sense this does not mean that in a "natural" case (no discrepancy) "auditory cues are stronger than visual cues" for the perception of prosody. In line with this analysis we suggest that, as well as for the perception of the segmental content of speech, the auditory and visual modalities are integrated to perceive prosodic information.

Detailed comparisons of performances for normal and whispered speech also show that visual only perception is significantly better for whispered speech as well as faster (see reaction time advantage). Since this observation cannot be explained by an experimental bias, it suggests either that the visual cues are clearer for whispered speech or that there are more visual cues in whispered speech. In both cases, it appears that, for production, speakers compensate the lack of acoustic cues by reinforcing or adding visual cues. It is more likely that visual cues are reinforced in whispered speech rather than there be additional visual cues. We recall here that whispered speech was elicited in a way that may

have induced more carefully produced visual cues, since the speakers were seated 2m away from the experimenter whom they were instructed to speak to. Their task was not to whisper into someone's ear.

These experiments also revealed inter-speaker variabilities. The most important one was that, for one speaker, no AV advantage could be measured for whispered speech because the auditory only perception of prosodic focus was too high also for whispered speech (ceiling effect even for whispered speech). Another important point is the relatively great variations between the scores in the V modality for different speakers (greater than that observed for auditory only perception). This reflects the fact that there are variations in the productions of visual cues (articulatory as well as facial movements such as eyebrow movements or head movements). Some speakers therefore appear to be more visible than others.

Another interesting result is that of the influence of visual markers on the perception of speech. It seems quite legitimate to think that the use of blue lipstick for example may enhance the perception of lip articulation by attracting the viewer's attention to the speaker's lips. This consideration would predict that the visual only perceptual results should be better for experiment 1, in which speakers wore blue features (lipstick and markers), than for experiment 2, in which speakers wore no facial markers at all. Comparisons between the results of experiment 1 and experiment 2, show that this is not the case. The visual only performance results were in fact better for experiment 2 than for experiment 1. This result should be confirmed by further studies since the task was not exactly the same in the two experiments (experiment 1: focus condition identification, experiment 2: focus detection) which may have had an influence on the overall perceptual performances. However, it does suggest that, contrary to intuition, the use of facial markers (in order to conduct parallel production measurements on the stimuli for example) does not enhance perception. Another possible interpretation is that, in line with studies from Swerts and Krahmer (2008), the visual cue value of the upper face (eyebrows…) and head as a whole (head movements) is quite important. In experiment 1 only the lower part of the face was visible and the head was held still whereas, in experiment 2, the entire face was visible and the speaker was free to move his/her head.

In future studies, we would like to evaluate more clearly the auditory-visual integration processes for the perception of prosodic features by varying the relative strength and presence/absence of the different acoustic and visual cues. For doing so, we will use a talking head developed at the Speech and Cognition Departement – GIPSA-lab (Bailly et *al.*, 2007). We will then have the possibility to entirely control the different cues to prosodic focus. The use of talking heads to evaluate auditory-visual perception of prosodic features has already been validated by other researchers (Krahmer *et al.*, 2002a, 2002b, 2006; Granström & House, 2005). In addition, we would like to test the auditory-visual perception of other prosodic features in order to examine the possibility of generalizing the results presented here.

ACKNOWLEDGMENTS

We thank Christophe Savariaux and Alain Arnal for their technical help with the audiovisual recordings. We also thank Jean-Luc Schwartz and Marie-Agnès Cathiard for their advice. We are grateful to our five patient speakers as well as to all the participants to the perceptual tests and those who helped us recruit them. We warmly thank Robert D. Ladd and Björn Granström for their insightful and helpful reviews on an earlier version of this paper.

REFERENCES

AUDOUY, M. (2000). *Traitement d'images vidéo pour la capture des mouvements labiaux*. Engineering Master Thesis, Institut National Polytechnique de Grenoble, France.

BAILLY, G., ELISEI, F., & RAIDT, S., (2007). Virtual talking heads and ambiant face-to-face communication. In A. Esposito, E. Keller, M. Marinaro and M. Bratanic (Eds.), *The fundamentals of verbal and non-verbal communication and the biometrical issue* (pp. 302-316). Amsterdam The Netherlands: IOS Press BV.

BARTELS, C., & KINGSTON, J. (1994). Salient Pitch Cues in the Perception of Contrastive Focus. In P. Bosch & R. Van Der Sandt (Eds.), *Proceedings of the Conference on Focus and Natural Language Processing. Working Papers of the Inctitute for Logic and Linguistics,* (pp. 94-106). Heidelberg: IBM.

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.

BERNSTEIN, L. E, EBERHARDT, S. P., & DEMOREST, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *The Journal of the Acoustical Society of America*, **85(1)**, 397-405.

BESKOW, J., GRANSTRÖM, B., & HOUSE, D. (2006). Visual correlates to prominence in several expressive modes. *Proceedings of Interspeech 2006 – ICSLP*, Pittsburgh, USA (Pennsylvania), pp. 1272-1275.

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

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.

CAVÉ, C., GUAÏTELLA, I., BERTRAND, R., SANTI, S., HARLAY, F., & ESPESSER, R. (1996). About the Relationship between Eyebrow movements and F0 Variations. *Proceedings of the ICSLP 1996*, Philadelphia, USA (Pennsylvania), Vol. 4, pp. 2175-2179.

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

CHRISTOPHE, A., GOUT, A., PEPERKAMP, S., & MORGAN, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, **31**, 585-598.

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

DE JONG, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The 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.

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

DOHEN, M., LŒVENBRUCK, H., CATHIARD, M.-A., & SCHWARTZ, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, **44**, 155-172.

DOHEN, M., & LŒVENBRUCK, H. (2005). Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study. *Proceedings of Interspeech 2005 - EUROSPEECH*, Lisbon, Portugal, pp. 2413-2416.

DOHEN, M., LŒVENBRUCK, H., & HILL, H. (2006). Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variabilities. *Proceedings of Speech Prosody 2006*, Dresden, Germany, vol. I, pp. 221-224.

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*, Beijing, China, vol. 3, pp. 247-250.

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

GRAF, H. P., COSATTO, E., STROM, V., & HUANG, F. J. (2002). Visual Prosody: Facial Movements Accompanying Speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 396-401.

GRANSTRÖM, B., & HOUSE, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, **46**, 473-484.

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

HADAR, U., STEINER, T. J., GRANT, E. C., & ROSE, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, **26**, 117–129.

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

HIGASHIKAWA, M., & MINIFIE, F . D. (1999). Acoustical perceptual correlates of "whispered pitch" in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research*, **42**, 583-591.

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*, Barcelona, Spain, pp. 2071-2074.

KEHOE, M., STOEL-GAMMON, C., & BUDER, E. H. (1995). Acoustic Correlates of Stress in Young Children's Speech. *Journal of Speech, Language, and Hearing Research*, **38**, 338-350.

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. *The Journal of the Acoustical Society of America*, **77(1)**, 266-280.

KONOPSCYNSKY, G. (1986). *Du prélangage au langage : acquisition de la structuration Prosodique*. PhD Thesis, Strasbourg University.

KRAHMER, E., RUTTKAY, Z., SWERTS, M., & WESSELINK, W. (2002a). Perceptual Evaluation of Audiovisual cues for prominence. *Proceedings of the ICSLP 2002,* Denver, USA, pp. 1933-1936.

KRAHMER, E., RUTTKAY, Z., SWERTS, M., & WESSELINK, W. (2002b). Pitch, Eyebrows and the Perception of Focus. *Proceedings of Speech Prosody 2002,* Aix-en-Provence, France, pp. 443-446.

KRAHMER, E., & SWERTS, M. (2006). Perceiving focus. In C.-M. LEE (Ed.), *Topic and focus: a cross-linguistic perspective* (pp. 121-137). Dordrecht The Netherlands: Kluwer Academic Publishers.

LADD, R. D. (1996). *Intonational phonology*. Cambridge UK: Cambridge University Press (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, France.

LŒVENBRUCK, H. (1996). *Pistes pour le contrôle d'un robot parlant capable de réduction vocalique*. PhD thesis, Institut National Polytechnique de Grenoble, France.

LŒVENBRUCK, H. (1999). An investigation of articulatory correlates of the Accentual Phrase in French. *Proceedings of the 14th ICPhS*, 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 XXIIIth JEP*, Aussois, France, pp. 165-168.

MACLEOD, A., & SUMMERFIELD, A. Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131-141.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MILLER, G. A., & NICELY P. (1955). An Analysis of Perceptual Confusions among some English Consonants. *The Journal of the Acoustical Society of America*, **27(2)**, 338-352.

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.

MUNHALL, K. G., JONES, J. A., CALLAN, D.E., KURATATE, T., & VATIKIOTIS-BATESON, E. (2004). Visual Prosody and Speech Intelligibility – Head Movement Improves Auditory Speech Perception. *Psychological Science*, **15(2)**, 133-137.

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

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.

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-114). Hillsdale USA: Lawrence Erlbaum Associates.

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

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

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

SÉGUINOT, A. (1976). L'accent d'insistance en français standard. *Studia Phonetica*, **12**, 1-58.

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). Cambridge MA USA: The MIT Press.

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

SUMBY, W. H., & POLLACK, I. (1954). Visual contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, **26(2)**, 212-215.

SWERTS, M., & KRAHMER, E. (2004). Congruent and Incongruent Audiovisual Cues to Prominence. *Proceedings of Speech Prosody 2004*, Nara, Japan, pp. 69-72.

SWERTS, M., & KRAHMER, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics,* **36(2)**, 219-238.

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

TOUATI, P. (1987). *Structures prosodiques du suédois et du français*. Sweden: Lund University Press.

TOUATI, P. (1989). De la prosodie française du dialogue. Rapport du projet KIPROS. *Working Papers in Linguistics,* **35**, 203-214. Lund Sweden: University Press.

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.

VIHMAN, M. M. (1996). *Phonological Development: The Origins of Language in the Child*. Oxford UK: Blackwell Publishers.

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

TABLES

Table 1

Experiment 2: mean percentages of correct answers (focus condition identified correctly) and standard errors for each speech mode (normal and whispered), modality (AV, A and V), focus condition (focus: F and no focus: NF) and speaker (SB, SC, SD and SE). Chance level: 50%.

speaker	focus condition	normal speech						whispered speech					
		AV		A		V		AV		A		V	
		mean	std error	mean	std error	mean	std error	mean	std error	mean	std error	mean	std error
SB	F	98.4	0.2	97.6	0.2	91.9	0.5	94.4	0.3	95.2	0.4	89.5	0.5
	NF	97.6	0.2	100	0	79.8	0.6	92.7	0.4	96	0.3	59.7	0.9
SC	F	97.6	0.2	97.6	0.2	71	0.8	98.4	0.2	94.4	0.4	87.9	0.5
	NF	93.5	0.4	82.3	0.7	72.6	0.8	87.9	0.7	69.3	0.7	79.8	0.6
SD	F	98.4	0.3	97.6	0.2	68.5	0.7	100	0	98.4	0.2	86.3	0.7
	NF	98.4	0.2	95.2	0.4	88.7	0.5	82.3	0.6	60.5	0.6	91.9	0.5
SE	F	95.2	0.4	98.4	0.2	66.1	0.7	95.2	0.4	91.9	0.5	89.5	0.5
	NF	100	0	98.4	0.2	93.5	0.4	90.3	0.5	78.2	0.7	87.9	0.5

Table 2

Experiment 2: number of responses corresponding to each speech mode (normal, whispered), modality (AV, A and V) and focus condition ('focus' *vs* no 'focus').

	normal speech				whispered speech			
	AV	A	V	total	AV	A	V	total
'focus' responses	496	515	450	1461	539	590	539	1668

'no focus' responses	496	477	542	1515	453	402	453	1308
total	992	992	992	2976	992	992	992	2976

Table 3

Experiment 2: mean reaction times differences (in ms) and standard errors (std error) by speech mode (normal speech and whispered speech), modality comparison (AV - A, AV - V and A - V), speaker (SB, SC, SD and SE) and focus condition (F: focus, NF: no focus).

speaker	focus condition	normal speech						whispered speech					
		AV - A		AV - V		A - V		AV - A		AV - V		A - V	
		mean	std error	mean	std error	mean	std error	mean	std error	mean	std error	mean	std error
SB	F	-184	11.1	-304.5	13.2	-101.1	16.1	-175.6	16.9	-413.1	32.4	-237.4	34
	NF	-79.2	8.4	-609.1	34.7	-82.9	36.9	-49.4	16.1	-270.1	34.6	-220.6	35.3
SC	F	-24.8	12.1	-745.5	26.9	-119.8	29.5	-134.5	13.5	-316.1	19	-181.6	16.5
	NF	-5.9	16.6	-149.6	17.5	14.3	17	-63	21.3	-23.4	18.2	39.6	19
SD	F	-11.7	14.4	-428	13.6	-17.7	16.6	-81.2	15.3	-427.1	14.4	-345.9	13
	NF	-148.2	11	-135.8	16.2	3.9	18.9	2.7	17.5	96.1	21.4	93.4	11.9
SE	F	46.4	18.9	-405.4	18.7	-18.2	14.8	-63.8	19	-67.5	20.8	-3.7	18.5
	NF	-197.5	10	-192.7	10.7	24.1	13.2	-185.6	20.1	-15.9	20	169.7	14

FIGURE CAPTIONS

Figure 1

Example of an image from the recordings used for experiment 1.

Figure 2

Experiment 1: mean percentages of correct answers for each participant (light grey bars) and mean percentage over all participants (big dark bars) for each modality (chance level: 25%).

Figure 3

Experiment 1: mean percentages of correct answers and standard errors by modality for each speaker (chance level: 25%).

Figure 4

Example of an image from the recordings used for experiment 2.

Figure 5

Experiment 2: mean percentages of correct answers over all the participants and standard errors by speech mode (normal, whispered) and modality (AV, A, V). Chance level: 50%.

Figure 6

Experiment 2: mean percentages of correct answers over all the participants and standard errors by speech mode (normal, whispered) and modality (AV, A, V) for the four different speakers (SB, SC, SD and SE). Chance level: 50%.

Figure 7

Experiment 2: mean reaction time differences (in ms) and standard errors corresponding to the focus condition for the AV - A and AV - V modality differences and for both speech modes (normal speech and whispered speech).

Figure 8

Experiment 2: mean reaction time differences (in ms) and standard errors corresponding to the focus condition for the different modality differences (AV - A, AV - V, A - V), for both speech modes (normal speech and whispered speech) and for the different speakers (SB, SC, SD and SE).

FIGURES

Figure 1

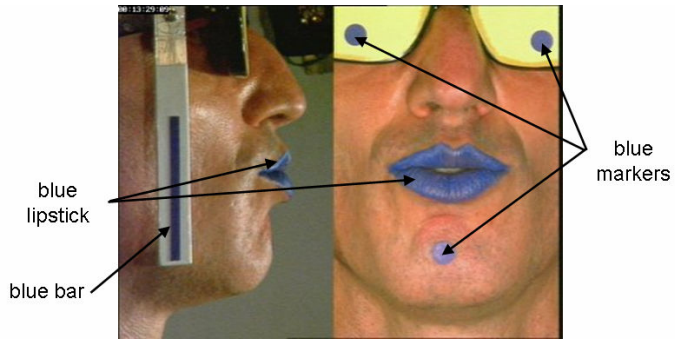


Figure 2

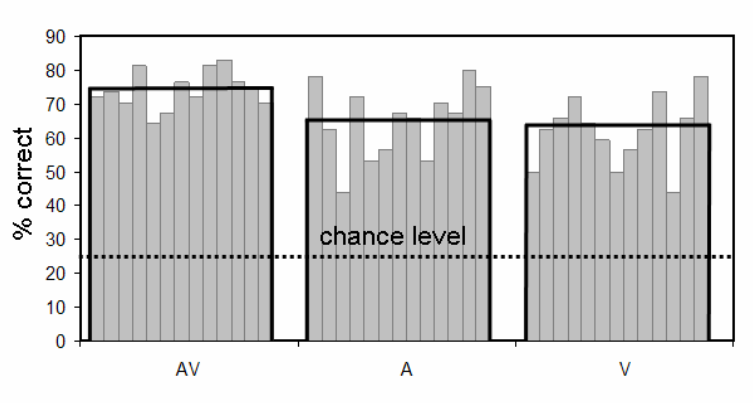


Figure 3

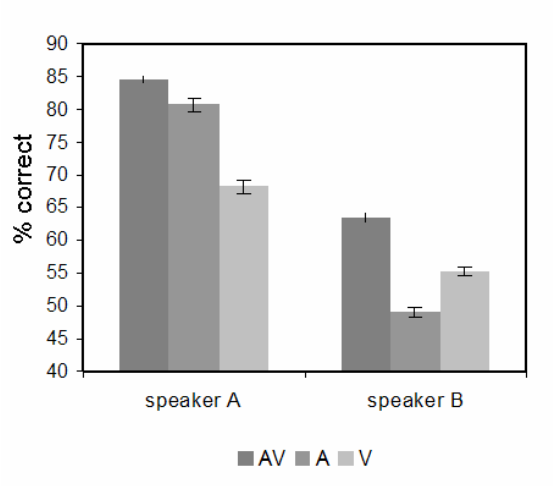


Figure 4

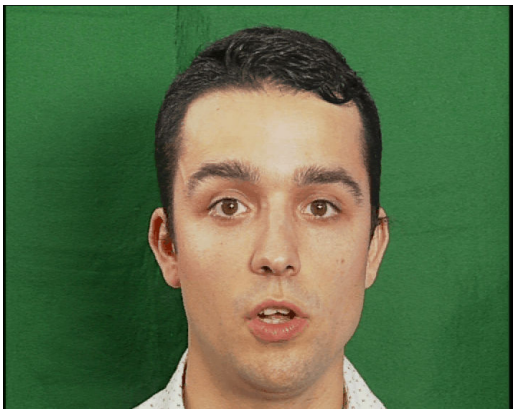


Figure 5

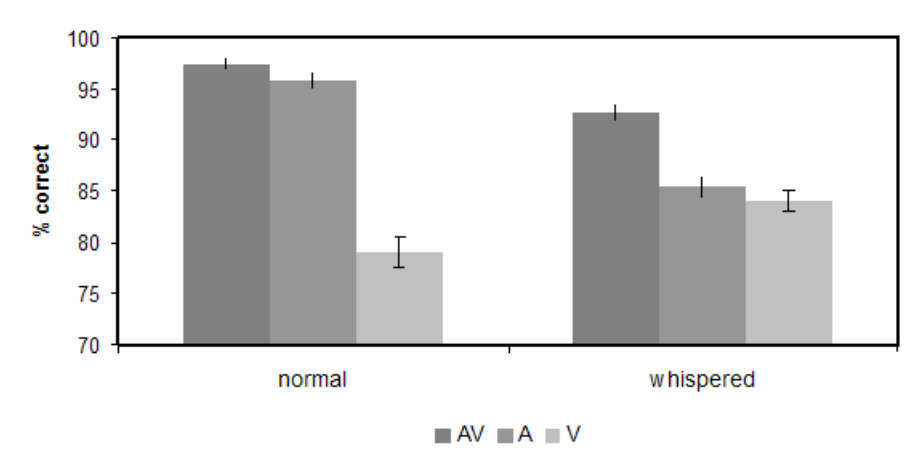


Figure 6

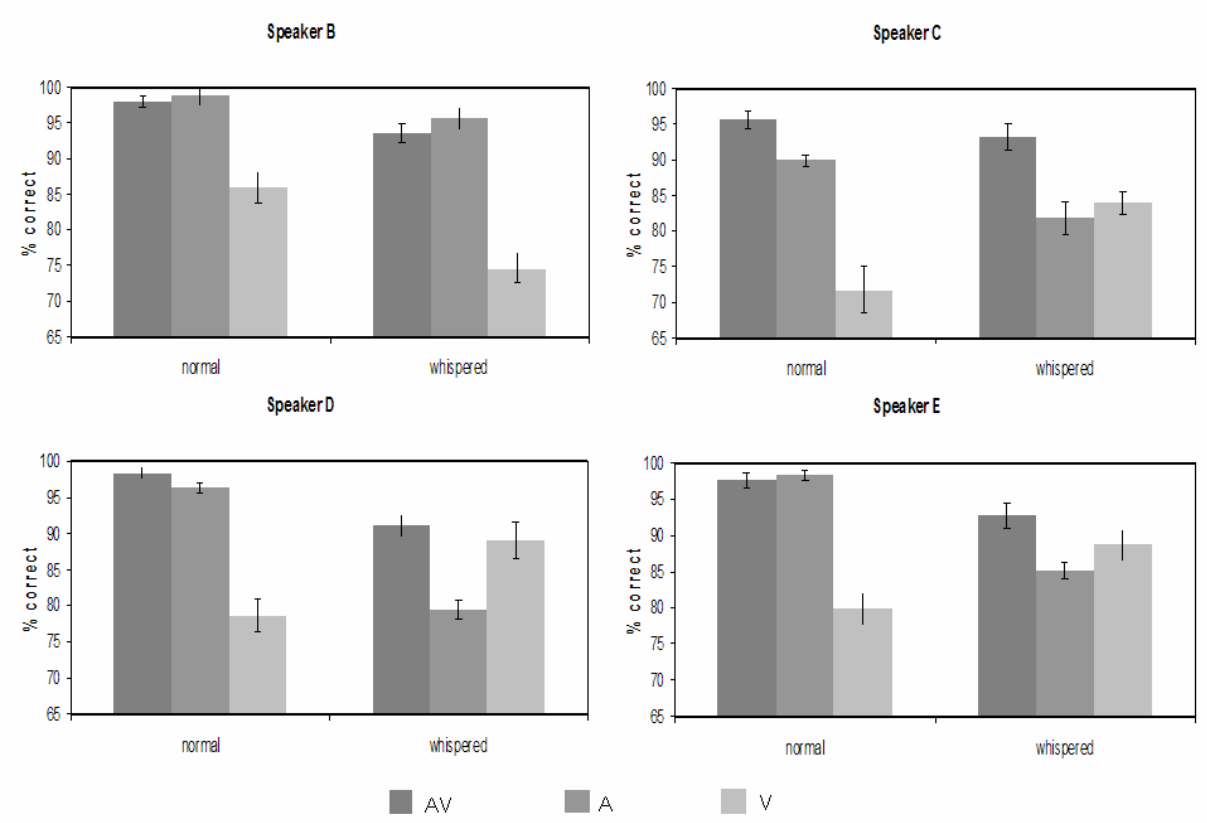


Figure 7

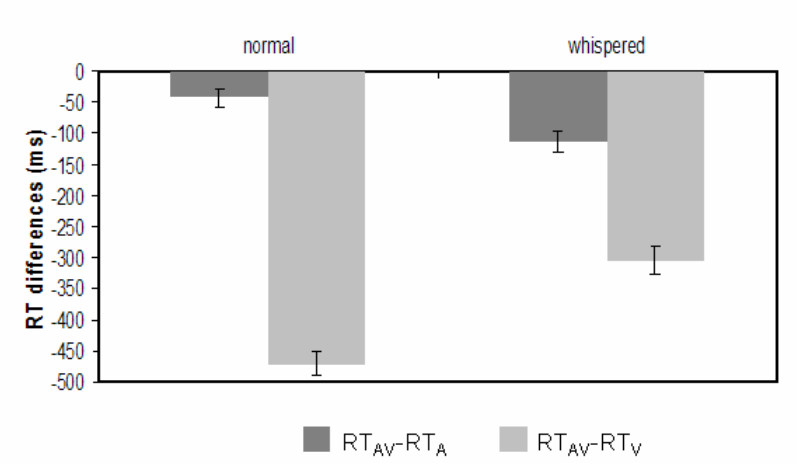
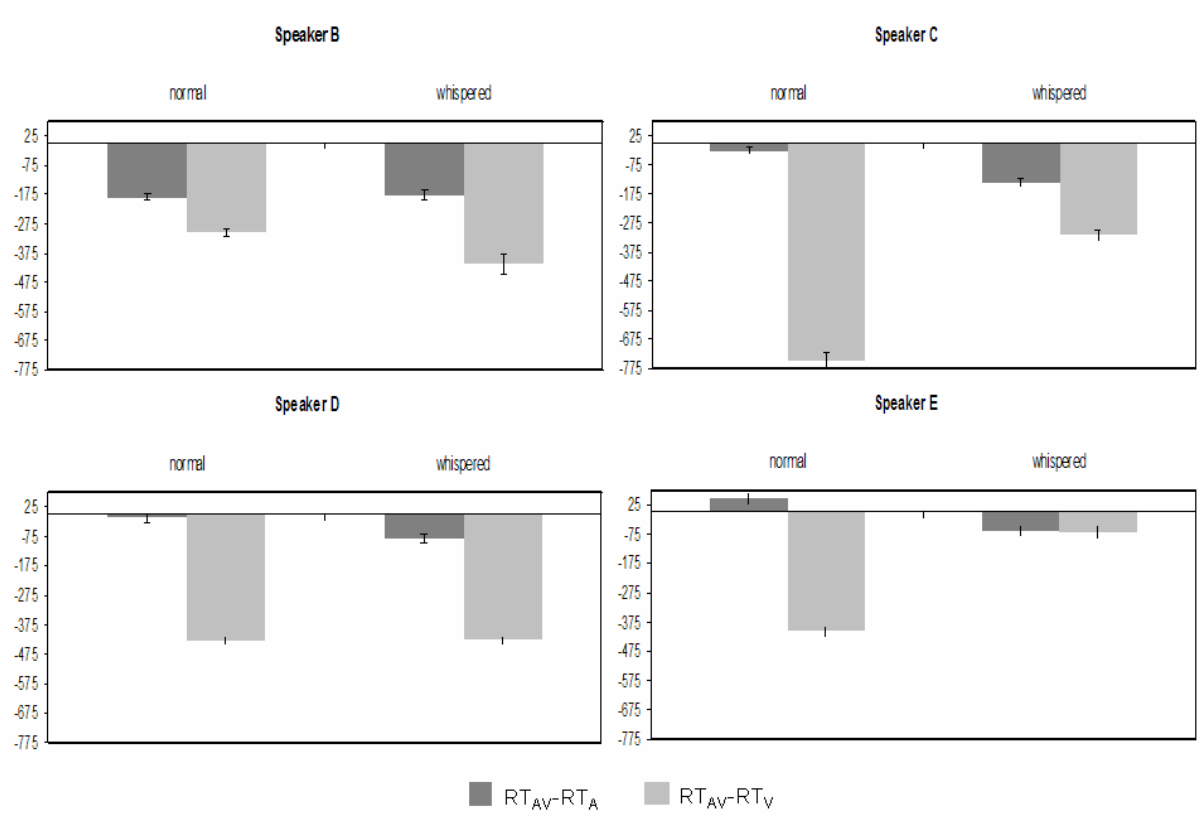


Figure 8


Speaker B
normal
whispered
Speaker C
normal
whispered
Speaker D
normal
whispered
Speaker E
normal
whispered
25
-75
-175
-275
-375
-475
-575
-675
-775
RTAV-RTA
RTAV-RTV