Exploring the contribution of prosody and gesture to the perception of focus using an animated agent*

Pilar Prieto,[2,1] *Cecilia Puglesi,*[1] *Joan Borràs-Comes,*[1] *Ernesto Arroyo,*[1] *and Josep Blat* [1]

[1] Universitat Pompeu Fabra, Barcelona, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
{pilar.prieto; joan.borras; ernesto.arroyo; josep.blat}@upf.edu; ceciliapuglesi@gmail.com

Corresponding author:

Pilar Prieto
ICREA- Universitat Pompeu Fabra
Departament de Traducció i Ciències del Llenguatge
Despatx 53.600
Campus de Comunicació – Poblenou
C/Roc Boronat, 138
08018 Barcelona

Tel. + 011 34 93 225 48 99
Fax. +011 34 93 542 16 17
Webpage: http://prosodia.upf.edu/home/en/index.php

Exploring the contribution of prosody and gesture to the perception of focus using an animated agent*

Abstract

Speech prosody has traditionally been analyzed in terms of acoustic features. Although visual features have been shown to help and enhance linguistic processing, the conventional view is that facial and body gesture information in oral (non-signed) languages tends to be redundant and has the role of helping the hearer recover the meaning of an utterance. We conducted two perception experiments with a 3D animated character showing non-matching auditory and visual information to investigate two related questions regarding the expression of the contrastive-corrective focus: (a) how important are facial cues with respect to auditory cues for the perception of contrastive focus? and (b) what is the relevance of the different gestural movements (i.e., head nodding, raised eyebrows) for the perception of contrastive focus? In the first experiment, participants were presented with a continuum of competing prosodic and gestural information for both information and contrastive focus statements, and had to decide whether they interpreted the utterance as having one meaning or the other. Results showed that the presence of either acoustic or gestural features of contrastive focus were key in guiding the listener towards this interpretation. The second experiment investigated the extent to which visual features (namely combinations of competing eyebrow and head movements) affect the perception of contrastive focus. Our results show that both types of cues (prosodic and gestural) contribute individually to the perception of contrastive focus, and that head nods are the most informative gestures for its detection. Overall, our findings lend support to the notion that the combination of prosodic and visual information is crucial in the interpretational component of language.

Index terms: contrastive focus, intonation, prosody, facial gestures, eyebrow movements, head movements, talking heads, audiovisual prosody, multisensory perception.

* A preliminary version of this article was presented at *Interspeech* 2011 (Florence, August 27-31). We are grateful to the audience at this conference, and especially to David House, for very helpful comments. We are particularly indebted to Javier Agenjo for his help with the NINOs platform. This research was supported by the following grants: FFI2009-07648/FILO, BFU2012-31995, CONSOLIDER-INGENIO 2010 'Bilingüismo y Neurociencia Cognitiva CSD2007-00012' and TIN2008-05995 Learn3 CICYT (awarded by the Spanish Ministry of Science and Innovation), 2009 SGR 701 (Generalitat de Catalunya), and iMP (European Union 7th Framework project). The second author also acknowledges a MAEC-AECID grant (Ministerio de Asuntos Exteriores y de Cooperación).

1. Introduction

The study of speech prosody has traditionally been based on acoustic information related to F0 pitch movements (information about the intonation contour of an utterance) and on information related to the duration and intensity of the sound chain. Yet it is well known that prosodic features have correlated visual features, such as head and eyebrow movements, or body and arm/hand movements, which can be very helpful in the processing of segmental and prosodic information during online speech communication. Classic experiments in the field of the prosody-gesture interface have shown that visual cues contribute to speech intelligibility and the detection of segmental information in noisy conditions (Sumby & Pollack 1954, Munhall, Jones, Callan, Kuratate & Vatikiotis-Bateson 2004). The study by Munhall et al. (2004) provided evidence that head movements can improve general speech comprehension. They used an animated character to assess speech-in-noise comprehension in the presence of normal head movements, exaggerated head movements, or no head movements, showing that performance comprehension increased in the presence of natural head movements. Al Moubayed & Beskow (2009) also showed that word recognition and intelligibility using animated talking heads increased when focally-accented (prominent) words were supplemented with head-nods or eyebrow raising gestures.

In general, little is known about the extent to which visual cues interact with acoustic information in the perception of sound and meaning. Previous work on how our perceptual system integrates auditory speech information and visual cues from a speaker's face has concentrated on effects at the segmental level. McGurk & MacDonald's (1976) classic study and Summerfield (1987) showed that perceptual confusions between consonants are different and complementary in the visual and auditory modalities, demonstrating that speech is multisensorally integrated. And our knowledge of audiovisual interactions in the presence of conflicting and non-conflicting information at the prosodic level is more limited (e.g., Dohen & Loevenbruck 2009, Swerts & Krahmer 2008). Swerts & Krahmer (2008) showed that conflicting visual and auditory prosodic information affect the location of emphatic words in a phrase. In their study, participants listened to recordings of a sentence with three prosodically prominent words whose auditory and visual prominence cues were manipulated. These cues were either congruent (i.e., occurring on the same word) or incongruent (i.e., occurring in such a fashion that the auditory and the visual cues were positioned on different words). Their results showed that participants could more easily determine prominence when the visual cue occurred on the same word as the auditory cue, while displaced visual cues hindered prominence perception. Dohen & Loevenbruck (2009) found that prosodic contrastive focus could be easily detected on the basis of either the audio or the visual modality alone, leading to a ceiling effect. In an experiment involving whispered speech, they reported that a combination of auditory-visual information constituted an advantage for the perception of prosodic features. Specifically, the study showed that adding vision to audition could improve focus detection and also reduce reaction times. They further suggested that audition and vision are integrated for the perception of prosodic focus.

In relation to the detection of prominence, some studies have shown that visual features can also be successfully used to identify prosodic information such as stress or determine which word in a sentence is emphasized (Bernstein, Eberhardt & Demorest 1989; Krahmer & Swerts 2007). Prominence is indeed one of the prosodic functions that has been studied in greatest detail and has been shown to be strongly correlated with facial features. It has commonly been reported that prominent words tend to be marked by means of facial gestures such as eyebrow movements or by more exaggerated movements of the articulators (Dohen, Loevenbruck, Cathiard, & Schwartz 2004;

Dohen & Loevenbruck 2009; Graf, Cosatto, Strom, & Huang 2002; Swerts & Krahmer 2008). Dohen, Loevenbruck & Hill (2006) tracked the movement of speaker's faces while producing prosodic contrastive focus in French and assessed head, eyebrow and cheek movements as well as the opening, spreading and protrusion of the lips, and chin movements. They found that contrastive focus in French is produced using lengthening and over-articulation of the focused constituent. With regard to articulatory correlates, the measure of lip protrusion had the largest correlation with speech stress, which suggests that this cue may be of particular importance, at least in French. They also found that focus is sometimes signalled by raised eyebrows and/or head nodding, but the link was highly inter- and intra-speaker dependent.

All in all, though the majority of studies suggest that visual facial dynamics convey important information that may improve the perception of focus in conversational situations, it is unclear how important facial cues are compared to auditory cues for the perception of focus. The first aim of this study will be to evaluate the relevance of prosodic and visual information in the perception of contrastive focus in an experiment which presents participants with conflicting prosodic and gestural stimuli. Experiment 1 will examine the potential influence of the presence of facial markers on the perception of contrastive focus when they are competing with auditory information. By using controlled manipulations of intonation (i.e., pitch range variation) and gestural variation through animated agents (i.e., variation in the strength of head nod and eye raising movements) we will be able to obtain a more focused analysis of the relative contribution of intonation and gesture to the conveyance of this type of prosodic meaning. The visual materials were implemented by means of a 3D animated character developed at the Technology Department of the Universitat Pompeu Fabra in Barcelona (Abadia et al. 2009), which crucially allowed us to assess in a precise way the contribution of varying degrees of these two visual cues.

Moreover, the extent to which specific facial cues influence linguistic perception of prominence and focus is far from well understood. In general, production studies have claimed that eyebrow movements and nodding are correlated with prominence marking in several languages (Ekman 1979, Dohen & Loevenbruck 2009, Swerts & Krahmer 2008, Dohen 2009), as well as being used as conversational signals (e.g. raised eyebrows to express interest on the part of the listener or nodding used to provide conversational encouragement; Ekman 1979).

In the perceptual domain, several studies have demonstrated the usefulness of eyebrow and head movements by synthetic talking heads to facilitate the audiovisual perception of speech prominence (Granström, House & Lundeberg 1999 and House, Beskow & Granström 2001, for Swedish; Krahmer, Ruttkay, Swerts & Wesselink 2002 and Krahmer & Swerts 2006, for Dutch; Massaro & Beskow 2002 and Srinivasan & Massaro 2003, for English). Recent work has also investigated which visual features are the most effective in perceptual terms for conveying prominence and focus in speech. Though results are partially contradictory, it seems that visual cues from the upper face together with head movements are powerful cues to prominence when synchronized with the stressed vowel of the prominent word. Swerts & Krahmer (2008) investigated which area of a speaker's face contains the strongest cues to prominence, using stimuli with either the entire face visible or only parts of it. Results showed that the upper facial area (i.e., eyebrow movements) had a stronger cue value for prominence detection than the lower area. Scarborough et al. (2009) assessed different facial measures during speech production and the extent to which these correlated with phrasal stress perception in English. While measures such as head and eyebrow movements played a role in perception performance, chin movement contributed most to correct perception independently of the other measures, thus suggesting that this is the most effective visual cue to stress. House et al. (2001) undertook a prominence detection experiment by systematically varying

the timing of both eyebrow and head movements of a talking face in a test sentence. Results indicated that both eyebrow and head movements were powerful visual cues for prominence and that perceptual sensitivity to timing misalignments was on the order of 100-200 ms.

Still, little is known about the cue value of visual gestures from the face (head/chin movement, eyebrow movement and lip movements) in the detection of prominence. The second aim of this study thus is to assess the role of the relative activation of two kinds of gestures, namely eyebrow movements and head/chin movements, in the perception of speech prominence. In particular, we are interested in analyzing the effects of head and eyebrow movements on the detection of contrastive focus. In Experiment 2, participants were presented with non-matching visual-only information combining four different activations of eyebrow raising and head nodding movement of a 3D animated avatar that ranged from the typical configuration of a statement to that of a contrastive focus statement. The task of the participants was to indicate for each stimulus whether they perceived the utterance as being a neutral statement or a 'correction' (i.e., contrastive focus) statement. The use of 3D animated characters allows us to finely control these sets of target facial movements and their combinations, something that is hard to achieve with human actors. By using controlled manipulations of these two types of gestural variation we were be able to obtain a more accurate assessment of the contribution of eyebrow raising and head nod movements to focus detection.

The article is organized as follows. Section 2 presents the methodology and results for Experiment 1 and section 3 the methodology and results for Experiment 2. Finally, we will conclude the article in section 4 with a general discussion of the implications of these results for an audiovisual model of focus perception.

2. Experiment 1

Experiment 1 investigates the relation between facial and intonational cues for the perception of information focus statements (IFS) and contrastive focus statements (CFS) in Catalan (see below for a definition of the types of foci). The experiment addresses this question by means of the presentation of conflicting audiovisual information using a 3D animated character.

2.1. Methods

For this experiment, participants were presented with a continuum of conflicting prosodic and gestural inputs for IFS and CFS, and had to decide whether they interpreted the utterance as having a contrastive or a non-contrastive statement meaning.

2.1.1. Participants

Eighteen Catalan-speaking participants from the Barcelona area (12 females; 6 males) took part in the experiment. They were university students aged 19–28 years (mean 23.1) with normal or corrected-to-normal vision and no symptoms of audiological disease. They were each paid 10 euros for their participation in the experiment.

2.1.2. Audiovisual recordings

First, high-definition audiovisual recordings of four subjects (two males and two females) were obtained as each of these subjects produced three natural renditions of the noun phrase *Marina* with either a IFS or a CFS meaning. By IFS, we refer to a neutral statement, i.e., a statement which carries new information in which there is a particular constituent that is focalized with respect to the background. On the other hand, a CFS refers to the marking of a constituent as "a direct rejection of an alternative" (Gussenhoven 2007). A CFS typically corrects "the value of the alternative assigning a different value" (Cruschina 2011). Therefore, the main difference between the two focus types is that while a CFS is dependent on a preceding assertion, which is denied/corrected by the new focalized item, an IFS is not. This denial or correction is often made explicit in the intonation and gestural planes of most intonational languages.

In order to prompt the corresponding pragmatic meaning in a natural way, subjects were asked to act out by replying to the speaker's question in each of the two short dialogues reproduced in (1), with (1a) involving an IFS and (1b) exemplifying a CFS.

(1) a. ═"Com es diu, la seva filla?" ═ "Marina."
"What's their daughter's name?" ═ *"Marina."*
b. ═"Es diu Júlia, ella, no?" ═ "MARINA!"
"Her name's Júlia, isn't it?" ═ *"[No! It's] MARINA!"*

All of the recordings were obtained using a professional digital video camera in a sound-attenuating chamber at the Universitat Pompeu Fabra. The subjects were video-recorded against a uniform white background, facing the camera, and the recordings showed all of the head and upper part of the body. The video recordings were digitized at 50 frames per second, with a resolution of 1,920×1,080 pixels. The sample rate of the sound was 32,000 Hz using 16-bit quantization. A total of 24 target video files were obtained (4 subjects x 2 conditions (IFS, CFS) x 3 repetitions).

The 24 target utterances were prosodically and gesturally analyzed. With respect to prosody, the 24 renditions of the utterances revealed an important intonational difference between IFS and CFS, namely a pitch range difference in the rising-falling nuclear configuration, which is expanded in the case of CFS. This prosodic difference between IFS and CFS in Catalan has been examined in previous work (Borràs-Comes & Prieto 2011, Borràs-Comes 2012). The mean pitch range values in semitones were used as extreme F0 values for the end of the continua of the generated stimulus in each condition (IFS, CFS) (see below). With respect to the gestural analysis, the 24 renditions revealed that typically CFS are characterized by different activations of eyebrow raising movements together with a head nod movement. In Catalan, as in other languages like Dutch, IFS and CFS share the same gestural pattern but show different degrees of activation of eyebrow and head nod movements (see Swerts & Krahmer 2008 for Dutch).

2.1.3. Materials and stimuli creation

To prepare the target stimuli for Experiment 1, one video file representing a CFS and another video file representing the same subject producing an IFS were selected. We were very careful to select examples that reflected naturalistic and representative head and eyebrow movements for both

statement types. The audio tracks of these two video files were then used to prepare the experimental audio continua. This was done by manipulating the F0 peak height of the information focus rising-falling configuration using Praat (Boersma & Weenink, 2008). By modifying the F0 peak height in 4 steps (distance between each one = 1.5 semitones), we created a synthesized continuum ranging from the mean values of the IFS pitch contour (163.5 Hz) to the mean values of a CFS pitch contour (212 Hz), using the CFS recording as a reference. The total duration of the stimulus for each of the four stimuli was 535 ms. A schematic diagram of these manipulations is shown in Fig. 1.

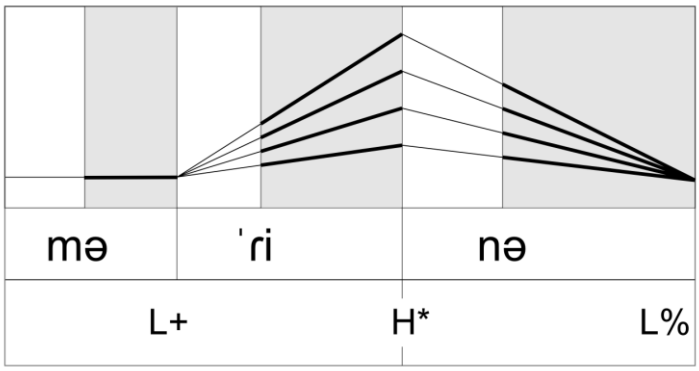

Figure 1. *Schematic diagram showing pitch manipulation.*

The visual materials for the experiment were created using the NINOs Platform on the basis of the two selected videos. NINOs is a platform created by a team at the Technology Department of the Universitat Pompeu Fabra that provides powerful software tools for the automatic production of audiovisual media, such as real-time animated virtual characters talking and acting in three dimensional virtual sets. The platform yields excellent visual performance as it is designed to work with the fastest 3D graphics techniques (Abadia et al. 2009). 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 order to create naturalistic synthesized visual continua for both statement and focus exemplars, the key frames for each gestural movement (the start and end points of every transition) for each visual variable (eyebrows and head) were obtained from the 24 target examples analyzed. The mean values of the key frames were then used to animate the 3D avatar. Figure 2 shows our gestural breakdown and analysis of a subject making a statement and the transfer of the relevant gestural features to images of the animated character. Smooth transitions between each of these images of the avatar were effected by the program, resulting in an animated video that lasted 1 second and contained 25 frames. This clip started and ended with the so-called Neutral Face FAP values (Facial Animation Parameters),[1]with the animation controlled by means of assigning different FAP

[1] The FAP values for the Neutral Face condition are as follows (see http://www.dsp.dist.unige.it, FAP Specifications Digital Signal Processing Lab. Department of Informatics Systems and Telecommunications (DIST), Faculty of Electronic Engineering, University of Geneva):

- the coordinate system is right handed; head axes are parallel to the world axes
- gaze is in direction of Z axis
- eyelids are tangent to the iris

specifications for eyebrow, head, and eyelid movements to the intervening frames (see colour lines below Fig. 2). Eyebrow and head movements went from a neutral state (Fig. 2, leftmost pictures) to a more intensely expressive one involving the raising of the eyebrows and a forward movement of the head or slight nod (Fig. 2, central pictures), and then back to a neutral state (Fig. 2, rightmost pictures). Finally, eyelids were animated as well, and lip movements were synchronized with the audio content, since they are essential for a naturalistic production of the sentence.

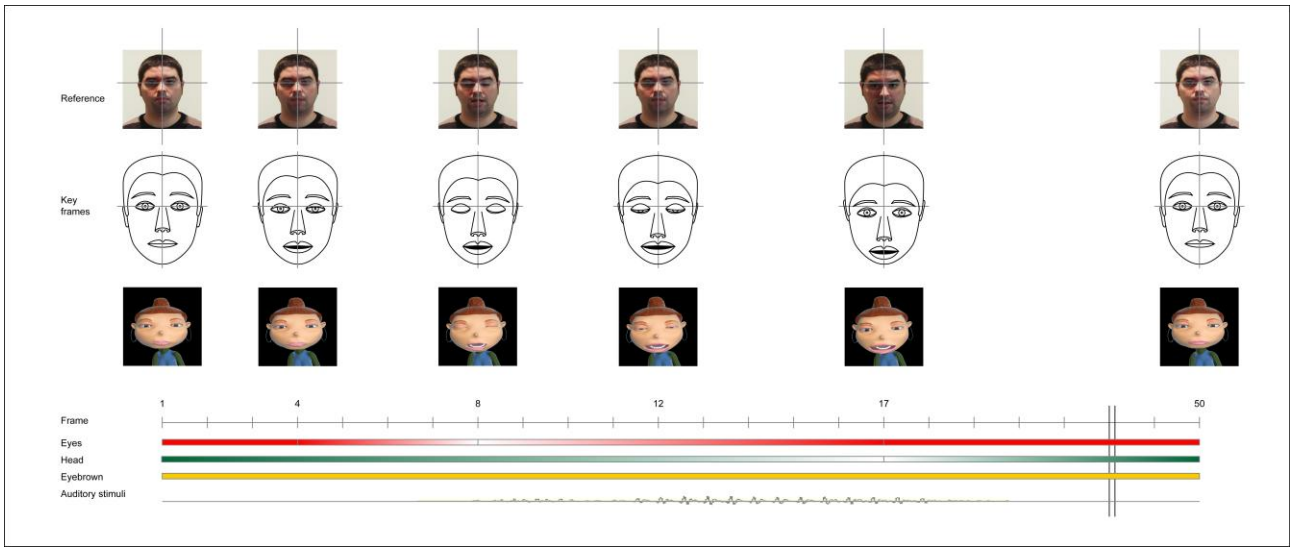

Figure 2. Representative stills of one of the recorded subjects for a statement utterance (top), gestural feature sketches (middle) and images of the animated character displaying those features (bottom).

After creating the facial movements for the IFS, a three-step visual continuum transformed it into a CFS utterance by modifying the degrees of activation of the eyebrows and the head. As mentioned above, the range values for the CFS head and eyebrow movements correspond to the mean values obtained in the analysis of the production data (12 recordings of live subjects). Figure 3 shows the transition in four steps from the IFS facial configuration (first row) to the CFS facial configuration (fourth row). From left to right, the stills correspond to the five key frames in the video sequence, namely initial position (first column), eye blink (second column), eyebrows completely raised, i.e., apex of the eyebrow raising gesture (third column), apex of the head nod gesture (fourth column) and final position (fifth column). Thus, a total of four video files were created (one for each row), with the initial and final frames in all of them depicting the neutral facial configuration.

• the pupil diameter is one third of the iris diameter
• lips are in contact; the line of the inner lips is horizontal and at the same height of lip corners

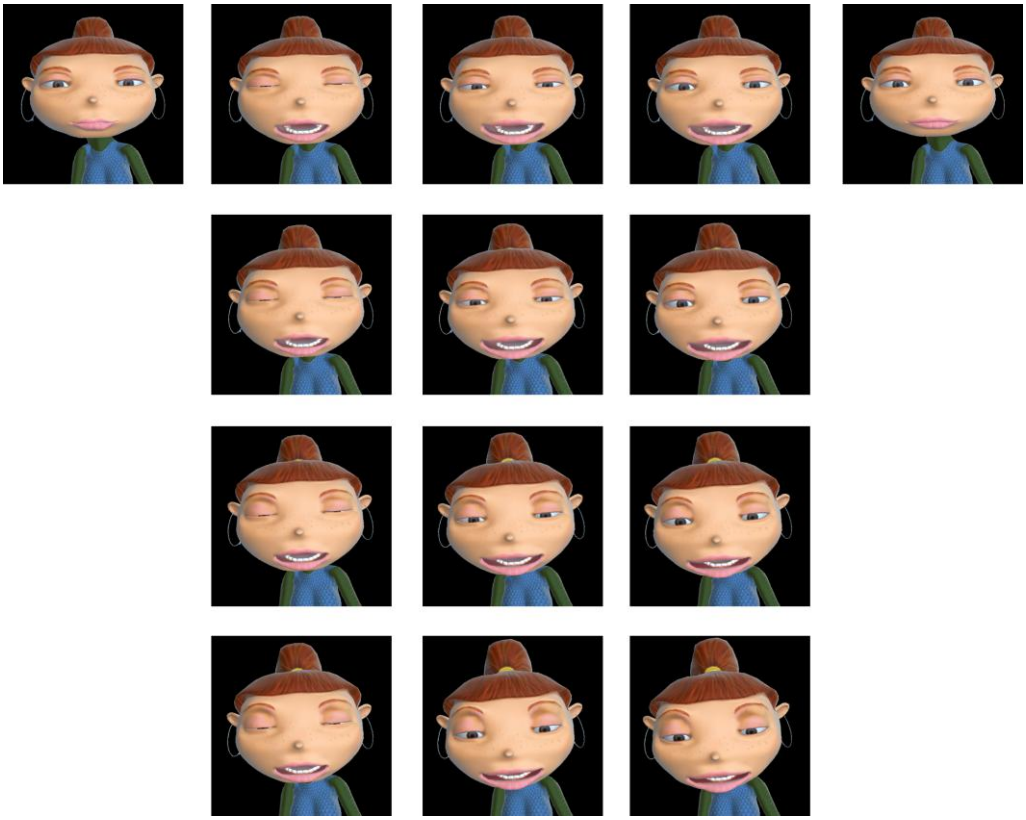

Figure 3. *Facial gesture transition from IFS configuration (first row) to CFS (fourth row) in four steps. The stills correspond to the five key frames in the video sequence: initial position (first column), with eyes completely closed (second column), with eyebrows fully raised (third column), with head completely down (fourth column) and in final position (fifth column). The initial and final frames are the same for all four sequences.*

A total of 16 audiovisual stimulus clips were created by combining the 4-step acoustic continuum of the rising-falling intonation contour (see Fig. 1) with the 4-step continuum of visual materials created with the animated character (Fig. 3). The combinations of audio and video stimuli were created with Adobe Premier CS3. The video recordings were rendered at 25 frames per second, with a resolution of 720×576 pixels. The sample rate of the sound was 32,000 Hz using 16-bit quantization.

2.1.4. Procedure

For this experiment, participants were presented with the AV combinations and were asked to indicate which interpretation — IFS or CFS — was more likely for each stimulus by pressing the corresponding computer key, namely 'A' for IFS ('statement' in Catalan is *afirmació*) and 'C' for CFS (in Catalan, *correcció*). The experiment was set up using the 16 stimulus video files in E-prime version 2.0 (Psychology Software Tools Inc., 2009). Reaction time measures were also recorded.

The experiment was administered in a quiet room at the Universitat Pompeu Fabra using a laptop computer equipped with professional headphones. A total of 1,440 responses were obtained (4 audio × 4 video × 5 blocks × 18 subjects). The experiment lasted a total of 10 minutes.

2.2. Results

The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 1440 data points, 92 cases were treated as outliers, i.e., cases where the reaction times were at a distance of at least three standard deviations from the overall mean (in this instance, 2790 ms.). These cases were excluded from the analysis. A Generalized Linear Mixed Model (GLMM) analysis was then conducted, with identification rate as the dependent variable, intonation, gesture (4 levels) and sound (4 levels) as fixed factors, and subject and block as crossed random factors. A main effect of intonation was found ($F_{3, 1331} = 42.451$, $p < .001$) and also a main effect of gesture ($F_{3, 1331} = 64.034$, $p < .001$), with no interaction between the two ($F_{9, 1331} = 1.099$, $p = .874$). Figure 4 shows the identification functions obtained. The y-axis represents the proportion of CFS identifications as a function of the continuum of visual information (line codes: 1 = IFS gestures and 4 = CFS gestures) and auditory information (x-axis: 1 = IFS intonation and 4 = CFS intonation). Crucially, clear gestural cues (3 and 4 line types) combined with appropriate acoustic cues (4 on the x axis) led to accurate identification responses more than 80% of the time, while more conflicting gestural and acoustic cues led to chance-level scores.

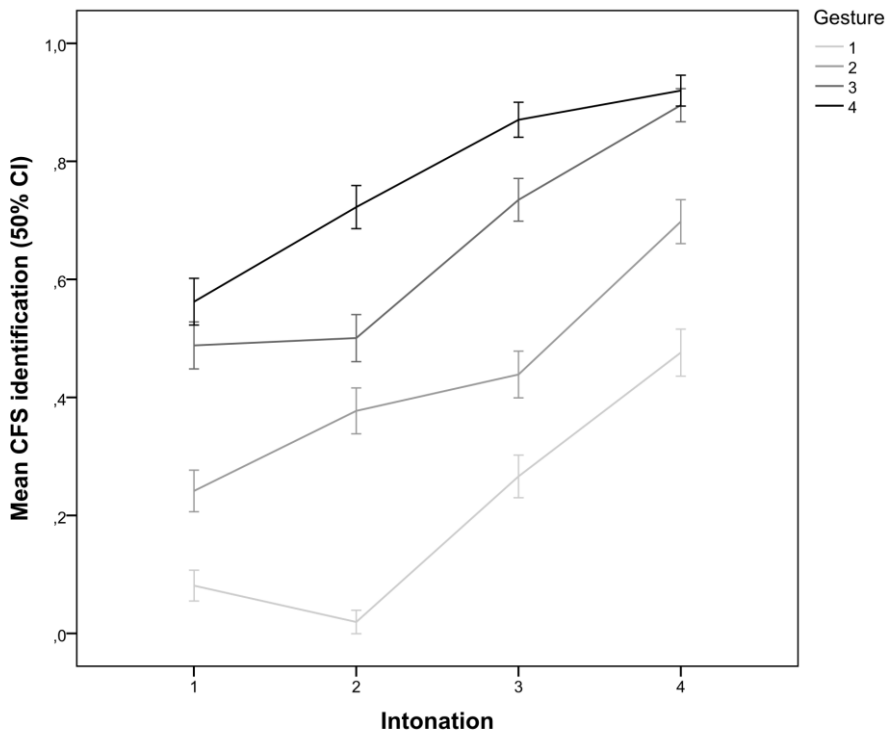

Figure 4. *Mean CFS identification (y-axis) as a function of the continuum of visual information (line codes: 1 = IFS gestures, and 4 = CFS gestures) and auditory information (x-axis: 1 = IFS intonation, and 4 = CFS intonation).*

Another GLMM analysis was conducted with reaction time as the dependent variable and the same fixed and random factors mentioned above. None of the fixed factors was found to be significant, neither intonation ($F_{3, 1331} = 2.316$, $p = .074$), nor gesture ($F_{3, 1331} = 2.197$, $p = .087$), nor the interaction between intonation and gesture ($F_{9, 1331} = 0.879$, $p = .544$), revealing that different combinations of gesture and intonation contours in these cases did not lead to significant differences in reaction times. The fact that the processing time was not significantly reduced in any of the conditions suggests that listeners were not especially puzzled by non-matching visual combinations

and indeed did not interpret them as incongruent AV combinations. This makes sense when one reflects that in the real world there are speakers that can use expressively prosodic patterns in their speech and at the same time use fewer gestural cues, or vice versa.

3. Experiment 2

Experiment 2 investigated the extent to which different activations of the gestural cues (namely combinations of competing eyebrow and head movements) would be responsible for the perception of contrastive focus. In contrast to Experiment 1, this was a visual-only experiment in which participants were presented with combinations of gestural cues but were not exposed to audio information.

3.1. Methodology

3.1.1. Participants

After they had performed the AV experiment (Experiment 1), the same 18 Catalan participants took part in the visual-only (VO) experiment (Experiment 2). In the interval between the two experiments, they watched a 2-minute video documentary about nature that was not narrated.

3.1.2. Materials

For this experiment, each of the two gestural cues (raised eyebrows and head nod) was presented in a continuum of 4 degrees of activation, from less pronounced to more pronounced gestures, with all the possible combinations between them. Figure 5 shows stills of the peak gesture combinations depicted in the 16 stimulus videos used in the experiment. They were obtained by combining 4 degrees of activation of the raised eyebrows with 4 degrees of head nod activation. The 4 activation levels of raised eyebrows are represented from left to right (the leftmost column corresponds to the minimum activation and the rightmost column to the maximum activation), and the 4 activations of head nod are represented from top to bottom (the first row represents the minimum head nod activation and the last row the maximum activation).

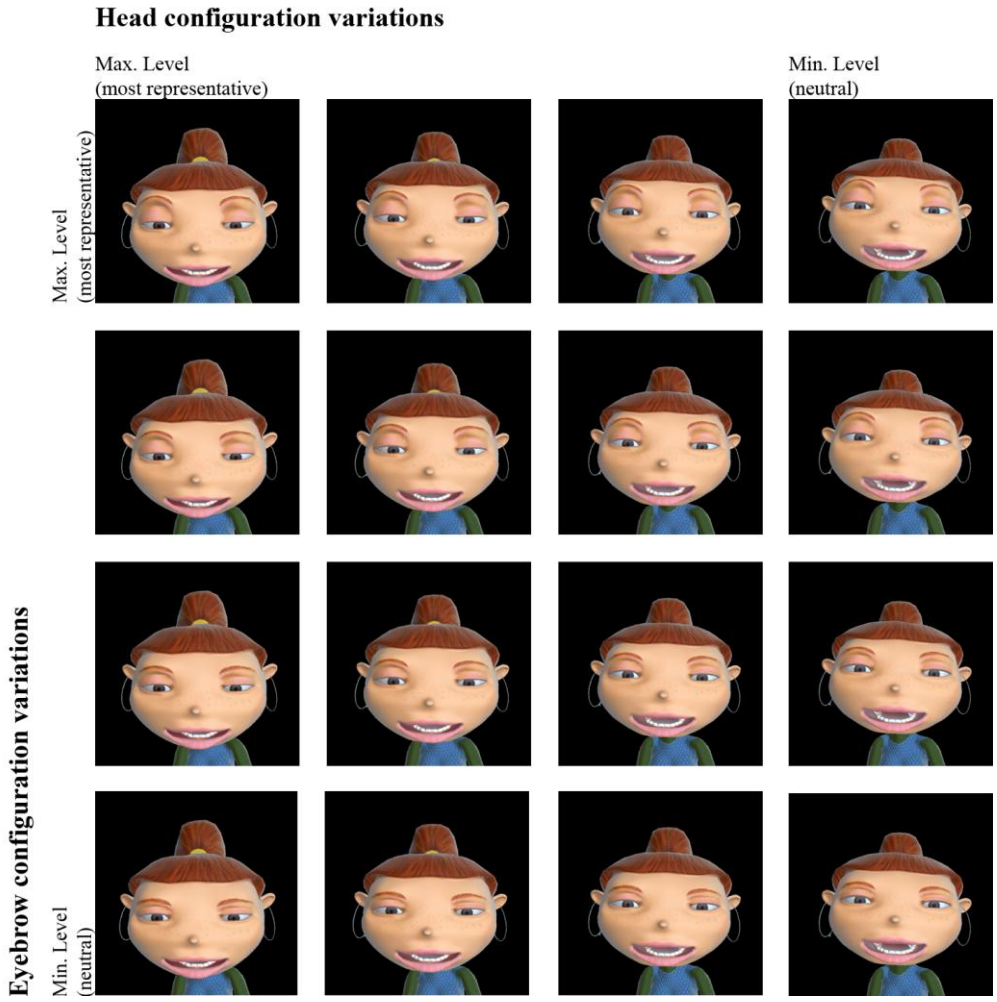

Figure 5. *Stills showing peak gesture combinations in the 16 videos used in Experiment 2. The four activations of eyebrow raising are represented from left to right (from lower to higher degrees of intensity) and the four activations of head nod movement are represented from top to bottom (again from lower to higher degrees of gesture prominence).*

3.1.3. Procedure

The procedure and instructions were the same as in Experiment 1, with the difference that subjects had to judge each video as portraying either an IFS or a CFS without hearing any audio track. A total of 1440 responses were again obtained for this experiment (4 head × 4 eyebrow × 5 blocks × 18 subjects). The experiment lasted a total of 10 minutes.

3.2. Results

Based on reaction time results, of a total of 1440 data points, 84 cases were treated as outliers because reaction times were at least three standard deviations from the overall mean (in our case, 2649 ms). A Generalized Linear Mixed Model (GLMM) analysis was conducted, with identification rate as the dependent variable, eyebrow (4 levels) and head (4 levels) gestures as fixed factors, and subject and block as crossed random factors. Main effects were found for eyebrow ($F_{3, 1340} = 8.480$, $p < .001$) and head ($F_{3, 1340} = 124.038$, $p < .001$) gesture, and their interaction was not statistically significant ($F_{9, 1340} = 1.832$, $p = .058$). Figure 6 shows the mean CFS identification rate (y-axis) as a function of eyebrow activation (x-axis) and head nod activation (lines). In general, the effect of head

movement was much stronger than that of eyebrow movements, especially when head nods were visually very pronounced (i.e., stimuli 3 and 4). The *F* coefficients obtained by the two fixed factors in the GLMM analysis confirm that this is indeed the case (i.e., *F*(head factor = 124.038 vs. *F*(eyebrow factor) = 8.480). From these results we can conclude that the head nod gesture was especially informative in the conveyance of contrastive focus.

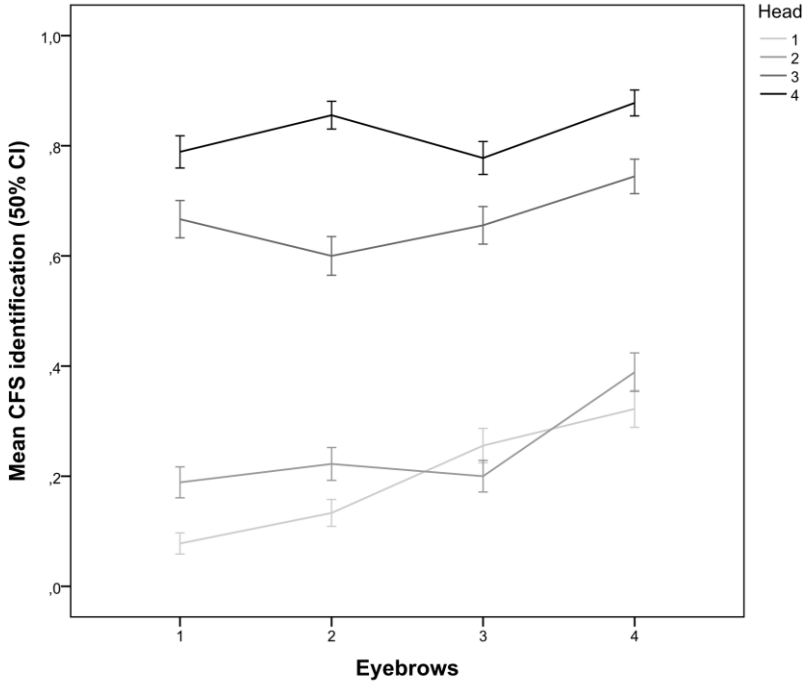

Figure 6. *Mean CFS identification (y-axis) as a function of head nod movement (lines) and eyebrow raising (x-axis).*

Another GLMM analysis was conducted with reaction time as the dependent variable and the same fixed and random factors as mentioned above. A main effect was found for head inclination ($F_{3, 1340}$ = 11.996, $p < .001$), but not for eyebrow raising ($F_{3, 1340}$ = 1.811, p = .143), and no significant interaction was found between eyebrow and head gestures ($F_{9, 1340}$ = 0.623, p = .778).

The visual-only perception experiment thus shows that visual perception of contrastive focus was significantly above chance when the head inclination was at its maximum or near maximum, with eyebrow raising being less accurately perceived for the detection of focus (see x-axis).

4. Discussion

This study has presented the results of two perception experiments with animated 3D avatars which sought to measure the importance of prosody and gestures (or visual markers) in the perception of contrastive focus interpretations in Catalan. To analyze the contribution of audio and visual cues, the strategy used in both experiments was to present subjects with combinations of different degrees of activation of either combined audiovisual (Experiment 1) or visual cues only (Experiment 2) and then ask them to detect the presence of contrastive focus. The main findings of these two experiments are now discussed in the context of the previous literature on the respective roles of visual and auditory cues in the detection of prominence.

Experiment 1 analyzed the relevance of facial and auditory cues in the detection of CFS as compared to IFS. A pilot study revealed that the differences in the actual production of IFS vs. CFS can be found in both a gradient expansion of pitch range and variation in the activation level of two specific gestures, namely head nodding movement and eyebrow raising. Listeners were presented with an acoustic continuum with intonation ranging from a typical IFS to a typical CFS realization. The audio continuum was combined with a visual continuum of facial gestures that ranged from the gestural sequence characteristic of IFS (slight head nod) to that characteristic of a CFS (more pronounced head nod). The results showed a main effect of gesture ($F = 64$) and, less powerfully, a main effect of intonation ($F = 42$) (in both cases, $p < .001$) in the perception of CFS. Listeners were more accurate in their detection of contrastive foci when clear gestures involving eyebrow raising and head nod were combined with intonation patterns involving the highest increase in pitch range. Because both modalities showed a gradient and equally salient distinction concerning the linguistic contrasts studied (IFS vs. CFS), a balanced use of auditory and visual cues was found in the participants' identification of both categories (with head movement being a clearer correlate of CFS marking than eyebrow raising in terms of gestural correlates). Still, head nod information was slightly more powerful than prosodic information, as stimuli with the strongest activation of head nod movements (even when combined with the IFS prosody) were more often rated as CFS than stimuli with the most expanded pitch range (even when combined with IFS gestural properties).

The analysis of reaction time patterns revealed significant effects of neither intonation nor gesture, nor the interaction between intonation and gesture, thus indicating that conflicting combinations of gesture and intonation contours do not lead to significant differences in reaction times. The fact that reaction times were not increased in the non-matching audiovisual conditions revealed that listeners did not perceive those combinations as incongruent, and that they were able to use acoustic and visual information independently of each other. One possible explanation for this perceptual behaviour relates to the findings in the production of focus which indicate that the visual marking of focus through eyebrow raising and/or a head nod is subject to a high degree of inter- and intra-speaker variability (see Dohen & Lovenbruck 2009, Dohen et al. 2006 for French; Scarborough et al. 2009 for English). If gestural marking of focus (instantiated through eyebrow raising and head inclination) is largely optional, then it is to be expected that its activation or lack thereof will not be perceived by listeners as incongruent with the presence of acoustic cues of focus.

This result contrasts with the very strong effect of bimodal audiovisual congruity reported in Borràs-Comes & Prieto's (2010) study of audiovisual discrimination between contrastive focus and counter-expectational questions. In that experiment, stimuli were identified as a counter-expectational question more quickly and more accurately when question-based visual stimuli was presented with a congruent audio stimulus. By contrast, identification became slower and less accurate when the visual stimuli occurred with exemplars of the non-matching nuclear pitch configuration (i.e., when Catalan participants saw a question-based visual stimulus but heard a non-matching low-pitched auditory stimulus or vice versa, a significant time delay appeared in their response). In our view, the strong effects of congruity/incongruity in reaction time measures found in Borràs-Comes & Prieto (2010) might be due to the coocurrence of different marked gestures for the two meanings (furrowing of the brows for counter-expectational questions vs. raising of the brows for contrastive focus) which were juxtaposed with greater auditory differences between the two meanings (normal pitch range for contrastive focus, very expanded pitch range for questions).

The main goal of Experiment 2 was to assess which head gestures are the most effective in perceptual terms for conveying contrastive focus in speech. The experiment analyzed the contribution of two visual cues to contrastive focus (namely, eyebrow raising and head nod) by

varying the relative strength of these two cues in a visual-only experiment. The results obtained showed that the perceptual impact of head movement is significantly greater than that of eyebrow movement. All in all, head movements acted as stronger perceptual cues of contrastive focus than eyebrow raising, possibly due to the stronger visual perceptibility of this gesture.

The strong perceptual value of head nod cues has not been previously reported in the literature. Scarborough et al. (2009) found that chin movement contributed the most to correct perception of phrasal stress in English, independently of the other measures, thus suggesting that this is the most effective visual cue to stress. Compared to chin and lip movements, head and eyebrow movements are optional cues to focus marking and this is probably why their relative weight has been played down. Swerts & Krahmer (2008) found that the upper facial area (eyebrow movements) had stronger cue value for prominence detection than the lower facial area. House et al. (2001) found that both head nod and eyebrow raising movements were important for prominence detection. However, the effects of head movement were not controlled for in their study. Since in our study we directly controlled the activation of eyebrow raising and head inclination, our results are informative of the fact that head movements are stronger indicators of focus than eyebrow raising. Yet it remains an open question whether chin and lip movements would cause potential confounding effects in focus detection for Catalan speakers.

What is the basis for the relative perceptual robustness of the head movement cues? A potential explanation for this effect is the fact that the physical displacement of the head is much more visibly obvious than the physical displacement of eyebrow raising, simply because the head is a larger object. In our digitized materials, the magnitude of eyebrow and head displacements were measured by using a measure related to the number of total visual segments contained in the talking head (20 segments). While brow displacement between its most extreme positions was just half a segment in the case of CFS, head nod displacement represented one segment. Thus, as expected, the overall magnitude of head displacement was higher than the overall magnitude of eyebrow displacement

In general, this study demonstrates the usefulness of using animated talking heads in the study of the importance of gestures and speech in communication. Notice that by using an animated character that is clearly non-realistic avoid the classic “uncanny valley” effect noted by the roboticist Masahiro Mori, whereby as robots become more human-like, their appeal increases, but only up to the point at which their human likeness became too realistic (albeit still not perfect), when their appeal falls dramatically. At the same time, such animated avatars allow for a precise control of gestural variation, which is necessary for this kind of focused analysis of the role of gesture in language understanding. Indeed, we believe that the use of animated agents for research on the relevance of visual cues in language understanding could be extended to the study of other potentially useful bodily markers such as hand and arm movements, which have also been shown to serve as beat gestures temporally aligned with prominent syllables (Krahmer & Swerts 2007).

Overall, the results of both experiments confirm the perceptual relevance of gestural cues in focus detection. Even in the visual-only experiment (Experiment 2), focus was accurately detected on the basis of gestural cues alone. Our results show that in Catalan, like in French, Dutch and English (see Dohen & Loevenbruck 2009 for French, Swerts & Krahmer 2008 for Dutch, and Scarborough et al. 2009 for English), visual cues are important in conveying focus. Although traditionally researchers dealing with correlates of contrastive focus have exclusively analyzed prosodic cues in speech-only stimuli, it is becoming increasingly clear that visual features interact with auditory features and need to be integrated in models of language understanding. Speech is multimodal and is in fact produced not just with the mouth, but also the face, hands and other parts

of the body, hence it is perceived not only with our ears but also with our eyes. In conclusion, our findings lend support to the view that the visual component does not merely accompany acoustic prosodic information but rather constitutes a crucial component in the semantic interpretation of utterances.

5. References

Abadia, J., Evans, A., Gonzales, E., Gonzales, S., Soto, D., Fort, S., Romeo, M., & Blat, J. (2009). Assisted animated production creation and programme generation. *Proceedings of the International Conference on Advances in Computer Entertainment Technology* (pp. 207-214). New York, NY: ACM.

Al Moubayed, S. & Beskow, J. (2009). "Effects of visual prominence cues on speech intelligibility", In *Proceedings of Auditory-Visual Speech Processing*, 43-46.

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.

Boersma, P., & Weenink, D. (2008). Praat: doing phonetics by computer (version 5.0.09). Computer Program. On line < http://www.fon.hum.uva.nl/praat/>

Borràs-Comes, J. & Prieto, P. (2011). 'Seeing tunes'. The role of visual gestures in tune interpretation. *Journal of Laboratory Phonology 2(2),* 335-380.

Borràs-Comes, J. (2012). *The role of pitch range in establishing intonational contrasts in Catalan*. Ph.D. Dissertation, Universitat Pompeu Fabra.

Borràs-Comes, J.; Costa-Faidella, J.; Prieto, P. & Escera, C. (2012). "Specific neural traces for intonational discourse categories as revealed by human evoked potentials". *Journal of Cognitive Neuroscience, 24.4*, 843-853.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In Bunnell, H.T. and W. Idsardi (Eds.), *Proceedings of the 4th International Conference on Spoken Language Processing* (pp. 2175-2178). Philadelphia, PA, USA.

Cruschina, S. (2011). Focalization and word order in Old Italo-Romance. *Catalan Journal of Linguistics*, *10*, 92-135.

Dohen, M. (2009). Speech through the ear, the eye, the mouth and the hand. In A. Esposito, A. Hussain, and M. Marinaro (Eds.). *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 24-39). Berlin/Heidelberg: Springer.

Dohen, M., Lœvenbruck, H., & Hill, H. (2006). Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variabilities. *Proceedings of the Third International Conference on Speech Prosody* (pp. 221-224), Dresden.

Dohen, M., & Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus", *Language and Speech, 52(2/3),* 177-206.

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.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepinies & D. Ploog (Eds.), 169-248. Cambridge: Cambridge University Press.

Foxton, J.M., Riviere, L.-D. & Barone, P. (2010). Cross-modal facilitation in speech prosody, *Cognition* 115: 71-78.

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition* (FGR 2002), Washington, 396-401.

Granström, B., & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46, 473–484.

Gussenhoven, C. (2007). Types of Focus in English. In Lee, C., Gordon, M., and Büring, D. (Eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation* (pp. 83-100). Heidelberg/New York/London: Springer.

House, D., Beskow, J., Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of Eurospeech* 2001 (pp. 387-390). Aalborg, Denmark.

Kendon, A. (2002) Some uses of the head shake. *Gesture, 2(2),* 147-182.

Kendon, A. (2004) Gesture: Visible Action as Utterance. Cambridge: Cambridge UP.

Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002a). Perceptual evaluation of audiovisual cues for prominence. Proceedings of the ICSLP 2002 (pp.1933–1936). Denver, CO.

Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002b). Pitch, eyebrows and the perception of focus. In *Proceedings of Speech Prosody 2002* (pp.443–446). Aixen-Provence, France.

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.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception, *Journal of Memory and Language, 57(3)*: 396-414.

Massaro, D. W., & Beskow, J. (2002).Multimodal speech perception: A paradigm for speech science. In B. Granstrom, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp.45 –71). Dordrecht, The Netherlands, Kluwer Academic Publishers.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 22,* 1777-1786.

Massaro, D. (1998) Perceiving talking faces: from speech perception to a behavioral principle. Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices: A new illusion, *Nature, 264,* 746-748.

Mori, M. (1970). The Uncanny Valley, *Energy,* 7(4), pp. 33-35

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.

Psychology Software Tools Inc. (2009). *E-Prime* (version 2.0). Computer Program.

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English, *Language and Speech, 52*: 135-175.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, 26: 212-215.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In V. Bruce, A. Cowey, A. W. Ellis and D. I. Perrett [Eds.], *Processing the facial image*. Oxford: Oxford University Press, 71-78.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving from the face and voice: Distinguishing statements from counter-expectational questions in English. *Language and Speech, 46(1),* 1-22.

Swerts, M., & Krahmer, E. (2008). Facial expressions and prosodic prominence: Comparing modalities and facial areas, *Journal of Phonetics, 36(2),* 219-238.