‘Seeing tunes.’ The role of visual gestures in tune interpretation

JOAN BORRÀS-COMES[1] and PILAR PRIETO[2]

[1]*Universitat Pompeu Fabra*
[2]*ICREA – Universitat Pompeu Fabra*

Abstract

One of the unresolved questions in audiovisual prosody is the relative contribution of acoustic and visual cues to the expression of prosodic meaning. Though the majority of studies on audiovisual prosody have found a complementary mode of processing whereby sight provides relatively weak and redundant information in comparison with strong auditory cues, other work has found that sight provides information more efficiently than hearing. In Catalan, a pitch range contrast in a rising-falling nuclear configuration conveys a difference between a contrastive focus statement and an echo question. The main goal of this study is to investigate the relative contribution of visual cues in conveying this distinction. Twenty native speakers of Central Catalan participated in two identification tasks in which they had to decide between a focus statement and a question interpretation. Experiment 1 used a pitch range auditory continuum combined with two congruent and incongruent videotapes showing the facial gestures that are characteristic of the two pragmatic meanings. Experiment 2 used the same auditory continuum in combination with another continuum for facial gestures produced using a digital image-morphing technique. The responses and reaction times obtained in both experiments revealed a consistent reliance on visual cues in the listener’s decisions, but also a consistent effect of the auditory stimulus. We argue that although facial gestures are the most influential elements that Catalan listeners rely on to decide between contrastive focus and echo question interpretations, bimodal integration with the acoustic cues is necessary for perceptual processing to be accurate and fast. Finally, we discuss the implications of these results for models of audiovisual processing.

Keywords: *audiovisual prosody, facial gestures, pitch accent range, intonational contrasts, Catalan language*

JLP 2 (2011), 355–380
DOI 10.1515/LABPHON.2011.013

1836-6346/11/02–355

1. Introduction

The strong influence of visual cues upon speech perception in normal verbal communication has increasingly been recognized. Audiovisual speech studies have revealed that the visual component plays an important role in various aspects of communication typically associated with verbal prosody. The visual correlates of prominence and focus (movements such as eyebrow flashes, head nods, and beat gestures) boost the perception of these elements (Hadar et al. 1983; Cavé et al. 1996; Krahmer and Swerts 2007; Swerts and Krahmer 2008; Dohen and Lœvenbruck 2009). Similarly, audiovisual cues for prosodic functions such as face-to-face grounding (Nakano et al. 2003) and question intonation (Srinivasan and Massaro 2003) have been successfully investigated, as have the audiovisual expressions of affective meanings such as uncertainty (Krahmer and Swerts 2005) and frustration (Barkhuysen et al. 2005).

In the last few decades, an important research topic in this field has been the relative importance of facial cues with respect to auditory cues for signaling communicatively relevant information. A large number of studies on audiovisual prosody have described a correlated mode of processing, whereby vision partially duplicates acoustic information and helps in the decoding process. For example, it is well known that visual information provides a powerful assist in decoding speech in noisy environments, particularly for the hearing impaired (Sumby and Pollack 1954; Breeuwer and Plomp 1984; Massaro 1987; Summerfield 1992; Grant and Walden 1996; Grant et al. 1998; Assmann and Summerfield 2004). Another set of studies has found a weak visual effect relative to a robustly strong auditory effect. For example, it has been found that observers extract more cue value from auditory features when it comes to marking prominent information in an utterance (Scarborough et al. 2009). Krahmer et al. (2002) found that people pay much more attention to auditory than to the eyebrow information when they have to determine which word in an utterance represents new information, and other follow-up studies confirmed the relatively weak cue value of these visual features, yet at the same time provided evidence that visual cues do have some perceptual importance (given that a visual-cue-only identification task yielded 92.4% correct guesses; see Krahmer and Swerts 2004).

Srinivasan and Massaro (2003) showed for English that statements and questions are discriminated both auditorily (on the basis of the F0 contour, amplitude and duration) and visually (based on the eyebrow raise and head tilt), but they also found a much larger influence of the auditory cues than visual cues in this judgment. Their results were consistent with those reported by House (2002) for Swedish, who found that visual cues (consisting of a slow up-down head nod and eyebrow lowering for questions, and a smile throughout the whole utterance, a short up-down head nod and eye narrowing for statements) did not strongly signal interrogative meanings, compared to auditory information like pitch range and peak alignment differences. Dohen and Lœvenbruck (2009) showed that adding vision

to audition for perception of prosodic focus in French can both improve focus detection and reduce reaction times. When the experimental paradigm was applied to whispered speech, results showed an enhanced role for visual cues in this type of speech. However, when evaluating the auditory-visual perceptual processes involved in normal speech, they found that auditory-only perception was nearly perfect, which suggests a ceiling effect for visual information. These results were in line with those from Krahmer and Swerts (2004), which showed that prosodic prominence was very well perceived in auditory-only mode for normal speech in Dutch and Italian. In relation to this, fMRI studies have shown that when visual and audio channels share time-varying characteristics this results in a perceptual gain which is realized by subsequent amplification of the signal intensity in the relevant sensory-specific cortices (auditory and visual) (see Colin et al. 2002; Calvert and Campbell 2003).

The abovementioned results could lead to the conclusion that visual information from the face is essentially redundant to auditory information, by using a set of audiovisual properties that can be found in most intonational languages. However, there are a few studies that have found that visual information is crucial in signaling certain types of attitudinal or emotional correlates. Studies like those of Swerts and Krahmer (2005), Dijkstra et al. (2006) and Mehrabian and Ferris (1967) have found that visual information is far more important for communicative purposes than acoustic information. In the first study, Dijkstra et al. (2006) studied speakers' signs of uncertainty about the correctness of their answer when answering factual questions. They noted the use of prosodic cues such as fillers ("uh"), rising intonation contours or marked facial expressions. Results showed that, while all three prosodic factors had a significant influence on the perception results, this effect was by far the largest for facial expressions. Similarly, Swerts and Krahmer (2005) showed that there are clear visual cues for a speaker's uncertainty and that listeners are more capable of estimating their feeling of an interlocutor's uncertainty on the basis of combined auditory and visual information than on the basis of auditory information alone. When visual expressions such as funny faces and eyebrow movements occurred, they seemed to offer a very strong cue for estimating uncertainty.[1] Mehrabian and Ferris (1967) analyzed how listeners got their information about a speaker's general attitude in situations where the facial expression, tone of voice and/or words were sending conflicting signals.[2] Three different speakers were instructed to say "maybe" with three different attitudes towards their listener (positive, neutral or negative). Next, photographs of the faces of three female models were taken as they attempted to convey the emotions of like, neutrality and dislike. Test groups were then instructed to listen to the various renditions of the word "maybe," with the pictures of the models, and were asked to rate the attitude of the speakers. Significant effects of facial expression and tone were found such that the study suggested that the combined effect of simultaneous verbal, vocal and facial attitude communications is a weighted sum of their independent effects with the coefficients of .07, .38 and .55, respectively. Nevertheless, these results do not

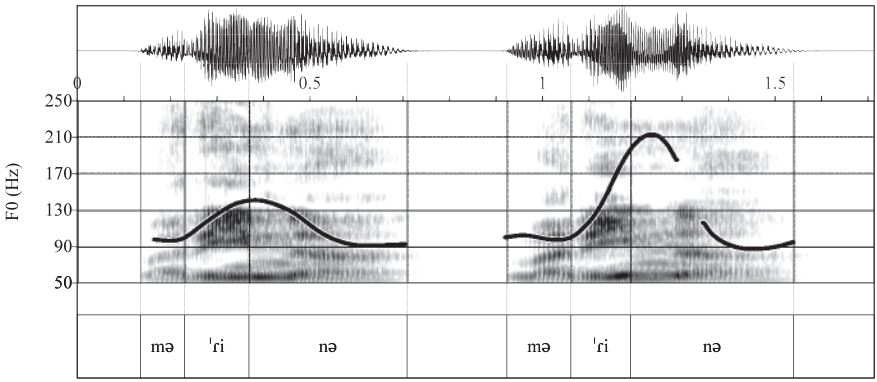

Figure 1. *Waveforms and F0 contours of the proper noun Marina 'person's name' produced with a contrastive focus statement meaning (left) and an echo question meaning (right).*

mean that the coefficients derived may not vary greatly depending upon a number of other factors, such as actions, context of the communication and how well the interpreting individual knew the other person (see also Lapakko 1997).

Thus, an overview of the literature reveals that visual cues are potentially useful as markers of prosodic information, yet it is still unclear how important they are compared to auditory cues. In the present study, we address this question by analyzing the patterns of prosodic perception of statements vs. echo questions in a group of Catalan speakers. The main goal of the study will be to investigate the relative contribution of visual and pitch accent cues in conveying this specific prosodic distinction in Catalan. In this language, a pitch range difference in a rising-falling nuclear configuration is the main intonational cue for the distinction between statements (both broad and contrastive focus statements) and echo questions (Borràs-Comes et al. 2010). Figure 1 shows the waveforms and F0 contours of the proper noun *Marina* produced with a contrastive focus statement meaning (left) and an echo question meaning (right). In line with this, Borràs-Comes et al. (2010) propose a L+H* L% nuclear configuration for the expression of contrastive focus and a L+¡H* L% nuclear configuration for an echo question (see the Cat_ToBI proposal in Aguilar et al. 2009, and Prieto [in press]).[3]

This article addresses two related questions regarding the perceptual processing of the audiovisual markers of echo question vs. contrastive focus meanings in Catalan. First, how important are facial gestural correlates to this distinction with respect to pitch accent cues? Second, are there differences in the relative weight of the acoustic information when facial cues are less prominent and thus more ambiguous? The advantage of using the Catalan distinction between focus statements and echo question meanings is that we will be assessing the relative perceptual importance of a well-known pitch accent contrast in the intonational phonology of Catalan (L+H* for contrastive focus and L+¡H* for echo question) in conjunction

with congruent and incongruent facial gesture information. To our knowledge, no previous studies have examined the bimodal perception of a prosodic contrast by using congruent and incongruent pitch accent and facial cue information. This methodology will allow us to create a very controlled situation where both pitch accent contrasts and visual information are carefully controlled for in a bimodal identification task.

The following sections describe the two experiments that were conducted to address these questions. Experiment 1 tackled the relative contribution of visual and auditory information to the target prosodic contrast by means of an identification experiment. For this task, subjects were presented with two video clips of a person's face as they spoke the word *Petita(?)* 'small' with their expression conveying one or the other of the two target meanings. The visual material was coupled with an audio track selected from a continuum of varying degrees of pitch range for the rising-falling configuration (the main acoustic cue to the distinction between the two meanings). Subjects were thus presented with either congruent or incongruent audio and visual target stimuli. Experiment 2 also investigated the role of auditory and visual information using the same stimuli but this time the continuum of audio cues was combined with a continuum of facial expressions created using a digital image-morphing technique. The task of the participants was again to identify the intended meaning (contrastive focus or echo question), for each combined audio + visual stimulus.

This article is organized as follows. Section 2 presents the audiovisual recordings that were used as a basis for creating the stimulus materials for both experiments. Section 3 presents the methodology for the two experiments, and section 4 shows the results of the two experiments. Finally, section 5 discusses the implications of these results for understanding the relative role of visual information and pitch accent contrasts in prosody perception.

2. Audiovisual recordings

Little research has been undertaken on the description of gestural patterns in Catalan. Most of the studies have been devoted to the description of Catalan emblems, i.e. specific hand/arm gestures which convey standard meanings that are used as substitutes for words (for example, holding up the hand with all fingers closed except the index and middle finger, which are extended and spread apart, can mean 'V for victory' or 'peace').[4]

There has been no previous research dealing specifically with the facial gestures that characterize echo questions and focus meanings in Catalan. Thus in order to decide which gestural patterns would be used as target facial expressions in our visual materials, ten native speakers of Catalan between the ages of 20 and 47 were videotaped pronouncing both possible interpretations of the utterance. Two of the

ten speakers were the authors, and the other eight were graduate students and professors, with no previous experience in audiovisual research. In order to prompt the corresponding answer, subjects were asked to read in an expressive way the two dialogues in (1), with dialogue (1a) involving contrastive focus statement and dialogue (1b) exemplifying an echo question. As is well known, in echo questions the listener repeats information that s/he has just heard, and these questions are sometimes marked by a nuance of surprise or incredulity. Subjects were given no instructions as to how to express these pragmatic meanings in audiovisual prosody. The audiovisual recordings of all ten speakers were carried out in quiet research rooms at the Universitat Autònoma de Barcelona and the Universitat Pompeu Fabra. Speakers were seated on a chair in front of a digital camera that recorded their upper body and face at 25 frames per second.

(1)	a.	— Volies una cullera gran, no?	*You wanted a big spoon, didn't you?*
		— **Petita**, [la vull, i no gran].	*[I want a] little [one, not a big one].*
	b.	— Jo la vull petita, la cullera.	*I want a little spoon.*
		— **Petita**? [N'estàs segur?]	*[A] little [one]? [Are you sure?]*

From these twenty visual tokens (ten for each pragmatic meaning), the authors assessed qualitatively the facial gesture correlates that were most effective and representative for each pragmatic meaning. One of the facial expressions that correlate most clearly with the perception of contrastive focus is the upward eyebrow movement and forward head movement. For an echo question conveying incredulity, the facial expression is characterized by a furrowing of the brows and a squinting of the eyes, often accompanied by a head shake. Figure 2 shows two representative stills of the facial expression as one of our speakers spoke a contrastive focus (left panel) and an echo question (right panel). For describing the facial gestures, we have used the Facial Action Coding System (FACS), developed by Paul Ekman and his colleagues, which allows coding of all visually distinguishable facial expressions (Ekman and Friesen 1978; Ekman et al. 2002). FACS groups muscle

Figure 2. *Representative stills of a facial expression of one of our speakers while producing a contrastive focus statement (left panel) and an echo question (right panel).*

activity into so-called Action Units (AUs) that bundle uniquely identifiable facial movements; the articulatory basis of these movements can thus be the activity of one or multiple muscles. Three AUs are relevant in the production of eyebrow movements (see also de Vos et al. 2009): AU 1, the Inner Brow Raiser; AU 2, the Outer Brow Raiser; and AU 4, the Brow Lowerer. For focus interpretations, the most common facial expression consisted of a combination of action units AU1+2 (Inner and Outer Brow Raisers) and M57 (Head Forward). For echo question interpretation, the most common pattern was a combination of AU4 (Brow Lowerer) and M58 (Head Backward).[5]

From the results of the production test it was thus clear that one of the most effective gestural cues for the distinction between contrastive focus statements and echo questions was the pattern of eyebrow movements. A number of crosslinguistic studies have shown that eyebrow movements combine with facial gestures (Cavé et al. 1996; Graf et al. 2002; Beskow et al. 2006; Scarborough et al. 2009) or head movements (Hadar et al. 1983; Graf et al. 2002; Munhall et al. 2004; Beskow et al. 2006; Scarborough et al. 2009) to express prosodic focus. For instance, it has been found that focus production is accompanied by eyebrow raising and/or a head nod (Krahmer and Swerts 2004, for Dutch; Dohen et al. 2006, for French).

It is also interesting to note that in sign languages, eyebrow movements serve various grammatical functions. For example, eyebrows are furrowed in *wh*-questions and raised in yes/no questions in American Sign Language (Baker-Shenk 1983; Grossman 2001; Grossman and Kegl 2006), Swedish Sign Language (Bergman 1984), British Sign Language (Kyle and Woll 1985) and Sign Language of the Netherlands (Coerts 1992) – see Pfau and Quer (2010) for a review.

The prosodic information obtained in this set of audiovisual recordings was used as a basis for the preparation of audiovisual stimuli for use in our two perception experiments (see section 3). While the acoustic information was almost identical in the two experiments (a set of either 11 or 6 pitch range differences created with PSOLA manipulation), the visual information was different, in that we used two unmanipulated video recordings for the contrast for Experiment 1 but used six videos in Experiment 2, with four of these clips being digitally-generated interpolations between part of the two used in Experiment 1.

3. Method

The goal of the two perceptual experiments was to test the relative importance of facial cues with respect to auditory cues for signaling the distinction between contrastive focus statement and echo question meanings in Catalan. Both experiments used an artificially generated pitch range auditory continuum. The second experiment used an artificially generated continuum of visual cues; the first used two unmanipulated video clips.

3.1. *Experiment 1*

The first experiment tested the role of auditory and visual information in pragmatic identification of contrastive focus statements and echo questions by means of an auditory continuum of pitch range which was combined with two video clips depicting the facial gestures characteristic of the two pragmatic meanings in such a way that the audio cue might be congruent or incongruent to a greater or lesser degree with the visual cue.

3.1.1. *Materials* To make sure that participants in our experiments could focus as much as possible on the audiovisual correlates of the two target pragmatic meanings, we selected a very short utterance that would contain the target intonational cues and facial gestures. To generate the audiovisual stimuli for the experiment, a native speaker of Catalan (the first author of this article) was videotaped several times producing natural productions of the noun phrase *petita* [pə.ˈti.tə] (‘small’-fem) with either a contrastive focus contour or an echo question contour. The author tried to imitate the two gestural patterns selected from among our preliminary video recordings as representative of the echo question and contrastive focus meanings (see section 2). The authors then selected the two exemplars that best characterized the contrast, while at the same time making sure that syllabic durations were similar in the two recordings. Figure 3 shows three representative

Figure 3. *Stills from video clips depicting facial gestures during the utterance of a contrastive focus statement (upper panels) and an echo question (lower panels). The three images correspond to three different stages of the gestures: initial expression (left), central expression (centre) and final expression (right).*

stills from the video clips as the subject utters first a contrastive focus statement (upper panels) and then an echo question (lower panels). The three images in each set correspond to three different stages of the facial gesture: initial expression (left), central expression (centre; approximately coinciding with the beginning of the stressed syllable) and final expression (right).

The target utterances were inspected for their prosodic properties. As expected, both target sentences were pronounced with a rising-falling intonational contour (L+H* L%) but differed in pitch range. The observed values for the high tone were 148.1 Hz for the contrastive focus example and 208.7 Hz for the echo question example. As noted above, duration patterns had been controlled for in the original materials. Table 1 shows the duration values of each of the target segments of the utterance *petita* in both readings (focus statement and echo question), revealing very small differences across the two utterances.

To prepare the target auditory stimuli for the experiments, we chose one of the two auditory recordings (the echo question) and manipulated the pitch by means of Praat (Boersma and Weenink 2008). A synthesized continuum was created by modifying the F0 peak height in 11 steps (distance between each one = 0.6 semitones). The pitch values corresponding to the accented syllable of the word *petita* were manipulated so that they would be realized as a 110 ms plateau starting 39 ms after the onset of the accented syllable /ˈti/, and were preceded by a low plateau for the syllable [pə] (102.4 Hz, 97 ms). The posttonic syllable [tə] was produced with a low plateau (94.5 Hz, 163 ms). A schematic diagram of these manipulations is shown in Figure 4.[6]

Each one of the auditory steps was then combined with the two target visual stimuli (see Figure 3), for a total of 22 target audiovisual stimuli. Since the video materials were recorded at 25 frames per second and the observed differences between natural auditory stimuli never surpassed 40 ms., no visual manipulations were needed to prepare the final audiovisual stimuli. An informal inspection of the data did not reveal cases of undesired lip-sync problems and visually the manipulated stimuli appeared natural. To confirm these impressions, we asked a panel of two independent judges to check all the stimuli in terms of whether they felt that

Table 1. *Original values of the duration (in ms.) of the target segments in the auditory sequence* petita *'small' and their difference.*

	original focus	original echo	difference
p	13	17	4
ə	68	80	13
t	41	39	2
ˈi	116	110	6
t	35	39	3
ə	116	124	8
Sum	389	409	

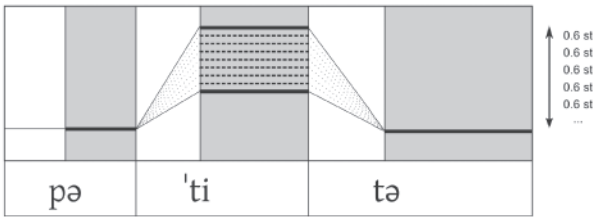

Figure 4. *Schematic diagram with the pitch target manipulation.*

either auditory or visual signals lagged behind, or instead appeared perfect synchronized. This additional check did not reveal any problematic cases of audiovisual mismatches.

3.1.2. *Procedure* Experiment 1 consisted of 5 blocks in which all 22 stimuli were presented to the subjects in a randomized order. A brief training session was conducted prior to the task in order to get subjects accustomed to the stimuli and the task. In this session, subjects were shown two repetitions of the fully congruent and fully incongruent audio + visual combinations.

Stimuli were presented to subjects using a laptop computer equipped with headphones. Subjects were instructed to pay attention to the auditory stimuli and facial gestures as a whole and decide which interpretation was more likely for each stimulus by pressing the corresponding computer key, "0" for *contrastive focus* and "1" for *echo question*.

The experiment was set up by means of E-Prime version 2.0 (Psychology Software Tools Inc. 2009), which allowed us to record response frequencies automatically. A timer with 1 ms accuracy was activated at the beginning of each stimulus, and the time that elapsed from the beginning of each playback to the striking of a response key was recorded, thus giving reaction time (RT) measurements. Subjects were instructed to press one of the two computer keys as quickly as they could. The experiment was set up in such a way that the next stimulus was presented only after a response had been given.

A total of twenty native speakers of Central Catalan participated in the experiment. The ages of the participants ranged from 18 to 36. All of them were undergraduate or graduate students with no previous experience in audiovisual research. The experiment was set up in a quiet research room at the Universitat Pompeu Fabra. We obtained a total of 2,200 responses (11 auditory steps × 2 visual sequences × 5 blocks × 20 listeners). The experiment lasted approximately 8 minutes.

3.2. *Experiment 2*

Experiment 2 analyzed the identification of contrastive focus statements and echo questions by means of the same auditory continuum used in Experiment 1 but this

(visual stimulus 2) (visual stimulus 3) (visual stimulus 4) (visual stimulus 5)

Figure 5. *Inbetween frames resulting from the digital morphing of the central facial expression between the contrastive focus gesture sequence (left) to the echo question gesture sequence (right).*

time in combination with a continuum of facial gestures produced using a digital image-morphing technique. The goal of this experiment was to test whether the creation of intermediate steps in facial gestures would affect the interpretation of the stimulus materials and how this gradient visual information would interact with the processing of the auditory information.

3.2.1. *Materials* To produce the target visual materials for Experiment 2, four static images were extracted from the target recordings used in Experiment 1, namely the first one for the initial neutral facial gesture, the second at the beginning of the stressed syllable, the third at the beginning of the post-tonic syllable and the last one at the end of the utterance (see Figure 3 above, which illustrates the first, second and fourth moments in time for each gesture pattern). Then, a face morphing technique was applied to the second, third and fourth stills selected (since the first one represented a neutral facial gesture; see Figure 3) in order to create four intermediate videos in between the two original video clips. The morphing was performed by means of Sothink SWF Quicker version 3.0 software (SourceTec Software Co. 2007). With this technique, one can morph one face into another by marking key points on the first face, such as the contour of the nose or location of an eye, and mark where these same points are located on the second face. The program will then create an intermediate frame between the first and second face. The drawings between the key frames are called inbetweens. Once we had the four inbetweens for each moment in time, we concatenated each set of key frames or inbetweens and synchronized them with the auditory materials. Figure 5 illustrates the 4 inbetweens resulting from the face morph manipulation from the contrastive focus gesture pattern (left) to the echo question gesture pattern (right). The total number of target visual stimuli was six.

The duration of this experiment was longer because the auditory materials had to be combined with the set of six video stimuli (instead of the two videos in Experiment 1). Because of this, we selected a subset of the auditory continuum used for Experiment 1, specifically, stimuli numbers 1-3-5-7-9-11 (the distance between each peak height thus becoming 1.2 semitones rather than 0.6). As in Experiment

1, each auditory stimulus was combined with each visual stimulus (6 videotapes), for a total of 36 target stimuli.

3.2.2. *Procedure* Experiment 2 consisted of 5 blocks in which all stimuli (36 in total) were presented to the subjects in a randomized order. Again, a brief training session was conducted prior to the task, in which participants were shown two repetitions of the most congruent and incongruent audio + visual stimuli.

The conditions for Experiment 2 and the instructions for subjects were the same as for Experiment 1, and the same group of twenty native Catalan speakers participated. We obtained a total of 3,600 responses (6 auditory steps × 6 visual sequences × 5 blocks × 20 listeners). The order of the two tasks was counterbalanced. The experiment lasted approximately 10 minutes.

4. Results

4.1. *Experiment 1*

4.1.1. *Identification responses* The graph in Figure 6 shows the mean "echo question" identification rate as a function of video stimulus (solid black line = focus statement video; solid gray line = question video) and auditory stimulus (x-axis), for the 20 subjects. The graph reveals that subjects mostly decided on the interrogativity of the utterance by relying on the visual materials, as the echo question video and the focus statement video responses are clearly separated in the graph (the echo question video elicited from 56% to 96% of "echo question" identification responses and the focus statement video elicited from 3% to 45% "echo question" identifications). Interestingly, there is also a clear effect of the auditory information but it is less robust: the preference for interrogativity is stronger for congruent audio + visual combinations (that is, a question video combined with a question pitch contour obtains 96% of "echo question" responses, and a focus statement video combined with a focus statement pitch contour obtains 3% of "echo question" responses). By contrast, most confusion arises in cases where the auditory cue is incongruent with the visual cue (that is, a question video with a focus statement audio track, or a focus statement video with echo question audio track). In other words, the congruent stimuli reveal more accurate responses than the incongruent ones. The clear congruity effects can be interpreted as evidence for a bimodal integration process.

A two-factor ANOVA with a 2 × 11 design was carried out with the following within-subjects independent factors: visual stimulus (two levels: focus statement and echo question) and audio stimulus (eleven levels: 11 steps in the pitch range). The dependent variable was the proportion of "echo question" responses. The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 2200 datapoints, 193 cases were treated as outliers, i.e. those

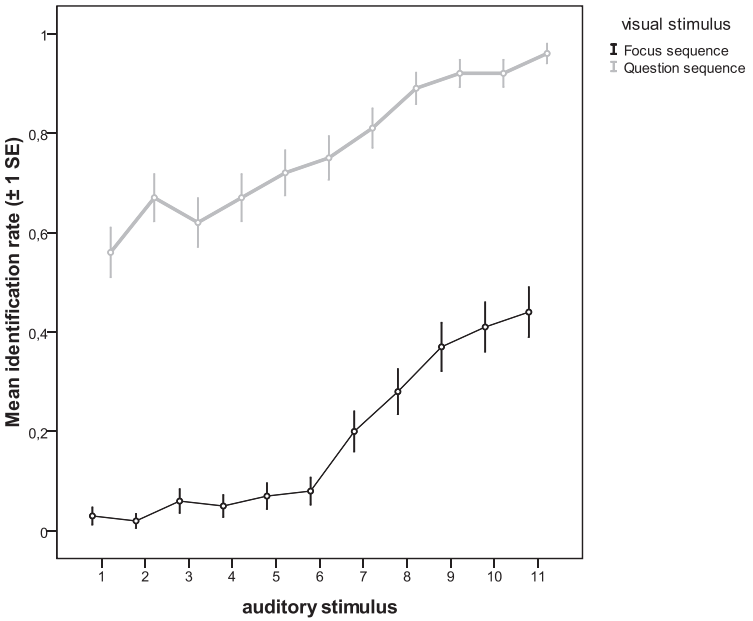

Figure 6. *Mean "echo question" identification rate as a function of video stimulus (solid black line = focus statement video; solid gray line = echo question video) and auditory stimulus (x-axis), for the 20 listeners. Error bars show ±1 Standard Error. In the x-axis, stimulus 1 is a contrastive focus statement and stimulus 11 is an echo question.*

cases where the reaction times were at a distance of at least three standard deviations from the overall mean. These cases were excluded from the analysis.

The analysis revealed a significant main effect of visual stimulus (F (1, 2007) = 1306.798, $p < .001$) and of auditory stimulus (F (10, 2007) = 31.119, $p < .001$) on statement/question identification. The interaction between the two factors was not significant (F (10, 2007) = 1.059, $p = .391$), meaning that the effects of both factors are consistent across factor groups. Thus we can observe a clear preference for visual cues in the listener's main decisions, but also a crucial effect of the auditory stimuli.

4.1.2. *Reaction times* Figure 7 shows mean reaction times (in ms) as a function of video stimulus (solid black line = focus statement video; solid gray line = echo question video) and auditory stimulus (1 = contrastive focus statement contour; 11 = echo question contour), for the 20 listeners. In general, mean RT patterns show that congruent audiovisual stimuli differ significantly from incongruent ones in that the latter trigger consistently slower reaction times. That is, when a question-based visual stimulus occurred with a low-pitched auditory stimulus, this triggered an important time delay in the response (mean RT: 786 ms). This is also the case

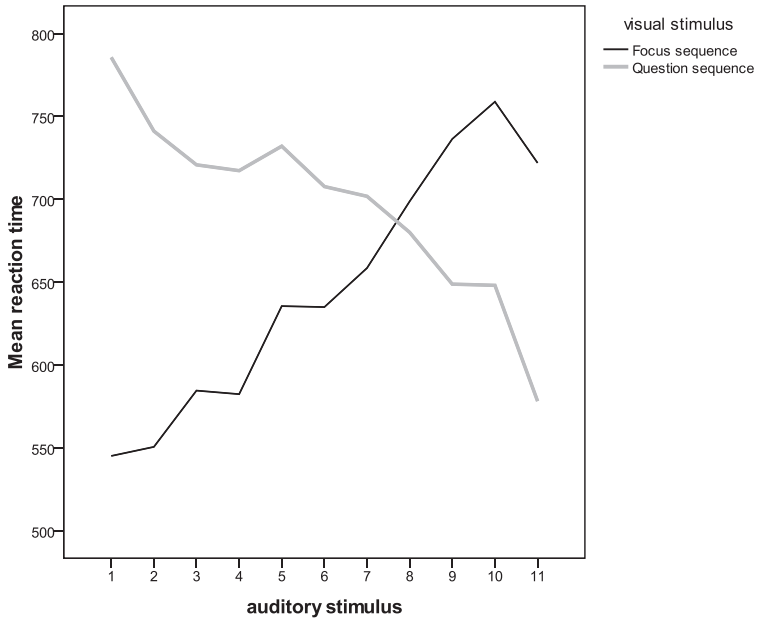

Figure 7. *Mean reaction times in ms as a function of video stimulus (solid black line = focus statement video; solid gray line = echo question video) and auditory stimulus (1 = contrastive focus statement contour; 11 = echo question contour), for the 20 listeners.*

when statement-based visual stimuli occurred with high-pitch auditory stimuli (mean RT: 722 ms). By contrast, congruent audio + visual combinations triggered very fast responses, namely in the combinations of a question video with the highest peak (mean RT: 578 ms) and of a statement video with the lowest peak (mean RT: 545 ms).

To get a first insight into the patterns of the reaction times, we conducted a *t*-test which compared averages for congruent and incongruent stimuli. Thus, for this test, we combined the two conditions for the extreme congruent stimuli (focus statement video with auditory stimulus 1, and echo question video with auditory stimulus 11) and paired those with that for the most incongruent stimuli (focus statement video with auditory stimulus 11, and echo question video with auditory stimulus 1). This *t*-test revealed that congruent stimuli differed significantly from incongruent ones in that the latter yielded consistently slower reaction times (congruent: 670 ms; incongruent: 979 ms) ($t_{(183)} = -3.619$, $p < .001$).

A two-factor ANOVA was carried out on the results. The dependent variable was reaction time measures. The within-subject independent variables were the visual stimulus (two levels: focus statement, echo question) and the auditory stimuli (eleven steps in the pitch range). The analysis revealed a clear effect of the visual factor for reaction times (F (1, 2173) = 6.362, p = .012), and no effect for the audi-

Table 1. *Mean "echo question" identification rates for each visual stimulus when combined with stimuli from each end of the auditory continuum in Experiment 2*

	lowest auditory stimulus	highest auditory stimulus
visual stimulus 1	.010	.475
visual stimulus 2	.030	.515
visual stimulus 3	.050	.592
visual stimulus 4	.340	.888
visual stimulus 5	.536	.970
visual stimulus 6	.582	.960

tory stimuli (F (10, 2173) = .671, p = .752). The interaction between the two factors was statistically significant (F (10, 2173) = 2.815, p = .002). Thus we clearly observe a preference for visual cues in the listener's main decisions, but also a crucial interaction between the visual and auditory information.

4.2. *Experiment 2*

4.2.1. *Identification responses* Figure 8 shows the mean "echo question" identification rate as a function of video stimulus (different types of lines, ranging from the solid black line = focus statement video to the solid gray line = echo question video) and auditory stimulus (x-axis), for the 20 listeners. The graph reveals a very similar pattern of responses to that obtained in Experiment 1. First, it is clear that the visual materials were crucial in the participants' decision on the interrogativity of the utterance, as again the echo question video responses and the focus statement video responses are clearly separated in the graph (the echo question video elicits from 58.2% to 96% of "echo question" responses while the focus statement video elicits from 1% to 47.5% of "echo question" responses). Table 1 shows the mean "echo question" identification rate for each visual stimulus (visual stimulus 1 = focus statement video; visual stimulus 6 = echo question video) when combined with auditory stimuli from both ends of the continuum, i.e. lowest pitch range and highest pitch range.

Importantly, in all cases we obtain the same effect of the auditory information as in Experiment 1: the preference for interrogativity is stronger for congruent audiovisual combinations (that is, a question video combined with a question pitch contour obtains 96% of "echo question" responses, and a focus statement video combined with a statement focus pitch contour obtains 1% of "echo question" responses). By contrast, most confusion arises in cases where the auditory cue is incongruent with the visual cue.

Interestingly, the tendency to rely on acoustic input is more detectable when the ambiguity of the visual stimulus is more extreme (see Table 1) as can be seen with visual stimulus 4. This elicits 88.8% of "echo question" responses when the audio cue shows an F0 contour with the highest peak (i.e. when the audio track is indeed

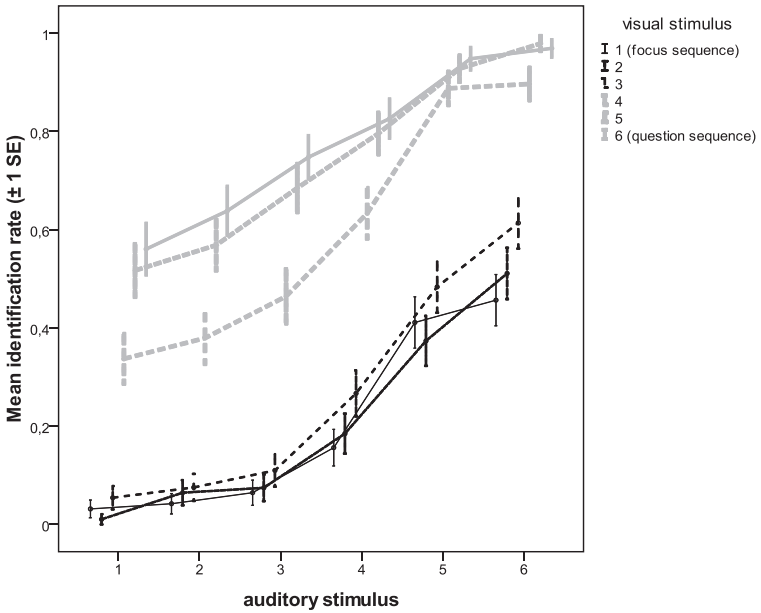

Figure 8. *Mean "echo question" identification rate as a function of video stimulus (different types of lines, ranging from the solid black line = focus statement video to the solid gray line = echo question video) and auditory stimulus (x-axis), for the 20 listeners. In the x-axis, stimulus 1 is a contrastive focus statement and stimulus 6 is an echo question.*

an echo question), and 34% of "echo question" responses when the F0 contour has the lowest peak (i.e. the audio track is a contrastive focus statement).

After completion of the task, several participants reported having seen facial expressions that looked "angry", especially for the most ambiguous visual stimuli. We argue that this collateral identification is an indicator of the ambiguity of the central visual stimuli, which thus increases the effect of the auditory information. In order to compare the curves obtained for the six visual stimuli, we calculated the slope value by means of a logistic regression. This slope value *per se* is not given directly by the function, but the term "b1" is related to the slope, with higher values reflecting shallower curves (Keating 2004). Table 2 shows the b1 value for all tasks. What we can see is that the slope for visual stimulus 4 is the shallowest.

A two-factor ANOVA with a 6 × 6 design was carried out with the following within-subjects independent factors: visual stimulus (six levels: 6 steps from focus statement to echo question) and audio stimulus (six levels: 6 steps in the pitch range). The dependent variable was the proportion of "echo question" responses. Again, the data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 3600 datapoints, 280 cases were treated as outliers.

Table 2. *b1 values of the logistic regression applied to the six visual stimuli across the six auditory stimuli.*

	v1	v2	v3	v4	v5	v6
b1	.482	.418	.489	**.525**	.472	.511

Table 2. *Mean RTs in ms for each visual stimulus across auditory stimuli when combined with auditory stimuli from each end of the continuum.*

	mean	lowest auditory stimulus	highest auditory stimulus
visual stimulus 1	712	**604**	779
visual stimulus 2	687	**575**	743
visual stimulus 3	**792**	730	883
visual stimulus 4	**900**	853	925
visual stimulus 5	691	766	**580**
visual stimulus 6	739	685	**505**

Parallel to the results of Experiment 1, the analysis revealed an effect of visual stimulus (F (5, 3404) = 289.617, $p < .001$) and an effect of auditory stimulus (F (5, 3404) = 149.821, $p < .001$). However, the interaction between the two factors was not significant (F (25, 3404) = 1.391, $p = .093$).

4.2.2. *Reaction times* Figure 9 shows the mean reaction times (in ms) as a function of video stimulus (different types of lines, ranging from the solid black line = focus statement video to the solid gray line = echo question video) and auditory stimulus (1 = contrastive focus statement contour; 6 = echo question contour), for the 20 listeners. Mean RT patterns show that congruent audiovisual stimuli differ significantly from incongruent ones in that the latter trigger consistently slower reaction times. First, the visual sequences closer to the focus gesture pattern (1 and 2) show an increasing function across the auditory stimuli; second, the visual sequences closer to the question gesture pattern (5 and 6) show a decreasing function across the auditory stimuli;[7] third, the most ambiguous visual stimuli (3 and 4) show longer reaction times when combined with almost all auditory stimuli and quite an increase when the auditory stimuli are more ambiguous. Table 2 shows the mean RT values for each visual stimulus, across all auditory stimuli, when combined with the lowest and highest auditory stimuli.

As with the results of Experiment 1, we conducted a t-test which compared averages for congruent and incongruent stimuli, the difference being that in this case the auditory stimulus representing the echo question end of the continuum was stimulus 6 (identical to stimulus 11 in Experiment 1). As in Experiment 1, again, this t-test revealed that congruent stimuli differed significantly from incongruent ones in that the latter yielded consistently slower reaction times (congruent: 591 ms; incongruent: 803 ms) ($t_{(180)} = -2.194$, $p = .029$).

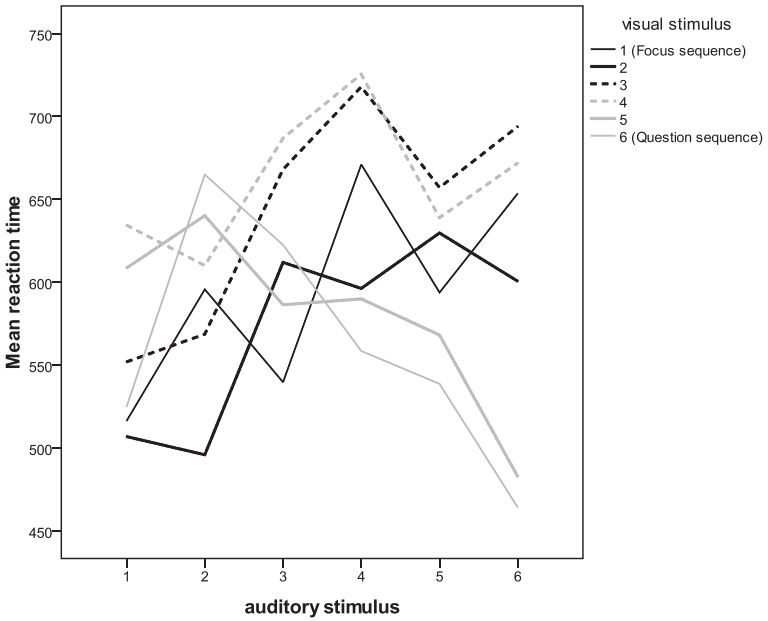

Figure 9. *Mean reaction time measures as a function of video stimulus (black different types of lines, ranging from the solid black line = focus videotape to the solid gray line = echo question videotape) and auditory stimulus (1 = contrastive focus statement contour; 6 = echo question contour), for the 20 listeners.*

A two-factor ANOVA was carried out on the results with the dependent variable again reaction time. The within-subject independent variables were visual stimulus (six steps from focus statement to echo question) and audio stimulus (six levels this time, not eleven). The analysis again revealed a clear effect of the visual factor for reaction times (F (5, 3564) = 11.608, p = .012), and no effect for the auditory stimuli (F (25, 3564) = .730, p = .601). The interaction between the two factors was again statistically significant (F (25, 3564) = 1.579, p = .034). Thus, we again observe a main effect of visual cues but also an important interaction between the visual and auditory input.

5. Discussion and conclusions

To what extent can gestural cues be crucial in encoding a linguistically relevant contrast such as the perception of statements and questions? This is a question that is still subject to debate among linguists and psycholinguists and has important consequences for models of multimodal language processing. In this article, we have explored the relative importance of pitch accent contrasts and facial gestures

in the perception of the contrast between contrastive focus statements and echo questions in Catalan, by using congruent and incongruent multimodal stimuli. Our general goal is to understand interaction in the linguistic processing of audio and visual cues during speech perception.

This paper has presented the results of two perceptual tasks that investigated how Catalan listeners use pitch accent information and facial gestures in making this linguistic distinction. Experiment 1 analyzed whether visual information is a more important cue than auditory information when a continuum of pitch range differences (the main acoustic cue to the distinction between contrastive focus statements and echo questions) co-occur with congruent and non-congruent facial gestures. Experiment 2 analyzed whether the role of auditory information is stronger when visual information is particularly ambiguous. In this case the visual stimuli were created by means of a digital image-morphing technique. Several important conclusions can be drawn from the results of these experiments with regard to the perception of statement and question prosody.

First, in both experiments, the response frequencies given by Catalan listeners revealed a clear preference for giving priority to visual cues when deciding between a contrastive focus statement and echo question interpretation. In both experiments, the listeners' decisions were mainly dependent on whether the video component of the audio + visual material they were watching show facial expressions corresponding to a focus statement or an echo question. Thus the present results show that focus statements and questions can be discriminated predominantly from visual information, with auditory information (on the basis of an F0 pitch range contrast) probably playing a secondary reinforcing role. In these experiments, the facial gesture acts as an integral part of language comprehension and, as such, provides insight into fundamental aspects of prosodic interpretation.

A second result that is obtained in the two experiments (and which can be observed in Figures 7 and 9) is the effect of bimodal audio + visual congruity. In both experiments, stimuli were identified as an "echo question" more quickly and more accurately when question-based visual stimuli occurred with a congruent audio stimulus (i.e. the upstepped pitch accent configuration L+¡H* L%). By contrast, identification became slower and less accurate (more chance-like) when the visual stimuli occurred with exemplars of the incongruent nuclear pitch configuration (i.e. L+H* L%). That is, when Catalan listeners saw a question-based visual stimulus occurring with an incongruent low-pitched auditory stimulus or vice versa, a marked time delay appeared in the response. Importantly, the strong effects of congruity/incongruity both in patterns of results and in reaction time measures represent a clear argument in favor of the view that facial gestures and speech form a single integrated system.

Third, another important result refers to the enhanced importance of acoustic stimuli when visual input is ambiguous. Attenuating the differences in the visual stimuli in Experiment 2 triggered a stronger influence of the auditory signals. Concerning theories of speech perception, integration models predict that both auditory

and visual information are used together in a pattern recognition process. On the one hand, the weighted averaging model of perception (WTAV; see Massaro 1998) predicts that the sources are averaged according to weight assigned to each modality. On the other hand, the fuzzy logical model of perception (FLMP) predicts, moreover, that the influence of one modality will be greater than the other when the latter is more ambiguous. According to the results of our Experiment 2, and in line with the findings of Massaro and Cohen (1993), we argue that this model of speech perception accounts for the processing of prosodic information better than competing models of perception (see also Srinivasan and Massaro 2003).

Our results showing a strong role for visual information in the perception of interrogation seems to partially contradict the results of a large number of studies in audiovisual prosody (e.g. House 2002; Krahmer et al. 2002; Srinivasan and Massaro 2003; Swerts and Krahmer 2004; Dohen and Lœvenbruck 2009; and others). We believe that it is in fact surprising that previous literature on audiovisual speech perception has not found more evidence of the role of visual information in linguistic interpretation. One possible explanation is that the use of real audiovisual recordings is better than the use of embodied conversational agents in avoiding the uncanny valley – the hypothesis in the field of robotics and 3D computer animation which holds that when facsimiles of humans look and act almost, but not perfectly, like actual humans, it causes a response of revulsion among human observers (Mori 1970; Prieto et al. 2011). Moreover, the claim that visual cues simply provide redundant information seems to be at odds with the famous McGurk audiovisual 'illusion' discovered by McGurk and MacDonald (1976). The basic McGurk effect found that an auditory [ba] stimulus combined with a visual [ga] stimulus resulted in a [da] percept. This effect is quite robust and has been replicated for many languages (see Burnham 1998 for an extensive review), thus suggesting that the brain tries to find the most likely stimulus given the conflicting auditory and visual cues, and that visual and auditory information are fused rather than the visual information being superimposed on the auditory one (see also MacDonald and McGurk 1978).

Virtually all studies that have found a complementary effect of visual cues have dealt with the perception of prominence or focus. Yet the studies that have focused on the role of facial expressions as salient indicators of the individual's emotional state (such as incredulity or surprise in echo questions, degree of uncertainty, etc.) have found a very strong effect of these cues. For example, the studies by Mehrabian and Ferris (1967), Swerts and Krahmer (2005), and Dijkstra et al. (2006), found that visual information is far more influential than acoustic information. Dijkstra et al. (2006) dealt with speakers' signs of uncertainty about the correctness of their answer and showed that facial expressions were the key factor in perception. Similarly, Swerts and Krahmer (2005) showed that there are clear visual cues for a speaker's uncertainty and that listeners are better capable of estimating another person's uncertainty on the basis of combined auditory and visual information than on the basis of auditory information alone.

Nevertheless, Srinivasan and Massaro (2003) showed that statements and echo questions were discriminated auditorily and visually, but they also found a much larger influence of auditory cues than visual cues in these judgments. We argue that the discrepancies between our results and theirs might be related to the audiovisual materials used. First, their visual materials were based on a synthetic talking head. The question face was characterized by a significant eyebrow raise and head tilt which extended dynamically across the length of the utterance. Yet it is well known that the eyebrow raise can also mark focalized constituents in statements, thus rendering the visual cues ambiguous between a question interpretation and a focus statement interpretation. Second, their auditory materials were manipulated on the basis of the F0 contour, amplitude and duration. Crucially, their difference in F0 contour implied changing a larger structure of nuclear and prenuclear tonal configurations (e.g. *We owe you a yo-yo | Pat cooked Pete's breakfast | We will weigh you | Chuck caught two cats*), leading to large modifications in the F0 of the stimuli, whereas our F0 changes were limited to changes in the pitch range of a single tonal target that always created a rising-falling intonation sequence. Listeners might have paid more attention to the sentential intonation contour than to the facial cues. As the authors themselves point out: "to assess whether the extended length of the sentence was responsible for nonoptimal integration, a shorter test stimulus (e.g.: "*Sunny.*" | "*Sunny?*") might be used. A short utterance might make statement/question identification a more automatic perceptual task, and less of a cognitive decision-making process. This task might engage an optimal bimodal integration process." (Srinivasan and Massaro 2003: 20)

In addition to the robustness of visual cues in identification tasks, an ongoing experiment involving the present first author and using the gating paradigm has confirmed that visual and audiovisual presentation of the materials triggered faster processing of the same linguistic contrasts, namely focus vs. question interpretation. The experiment tested the perception of a set of gated utterances occurring in the three possible modalities, namely audiovisual (AV), auditory only (AO) and visual only (VO). Preliminary results with 20 Catalan listeners have revealed that echo questions are recognized immediately in the VO condition (from the first gate) and that no differences appear depending on the presence of simultaneous auditory input. The recognition point is first found in the VO condition (between the first gate and the fourth), being closely followed by the AV condition. The responses to the AO condition are late (after the ninth gate).

Summarizing, our results provide clear evidence for the importance of visual cues in the perception of linguistic contrasts (in our case, the perception of statements and questions) and open the way to new investigations in this area. One of the research questions is the relevance of potential facial cues and their contributions to the judgements of statements and questions. We are currently testing this question by using computer-generated 3D talking heads to simulate face gestures during speech production. The visual stimuli used in this study will be implemented in a computer-generated 3D avatar in which each intended facial gesture (in our

case eyebrow position, eyelid closure, and head movement) is manipulated separately and appears on a continuum of four levels of strength.

Acknowledgements

We are grateful to Carme de la Mota, Itziar Laka, Lluís Payrató, Josep Quer, Núria Sebastián-Gallés, Maria-Josep Solé, Marc Swerts and Eric Vatikiotis-Bateson for their comments on an earlier version of this paper. We would like to thank the members of the *Group of Prosodic Studies* (UPF-UAB, Barcelona) for their help in managing E-Prime scripts and statistics. We would also like to thank the subjects who took part in the two experiments and the recordings of the audiovisual materials, especially Eva Estebas-Vilaplana, for their comments on the gestural materials. We also thank the editors of this special issue, Ian Maddieson and Caroline Smith, as well as Beatriz Raposo and two anonymous referees for their insightful comments, which helped clarify and strengthen this paper. This research has been funded by projects FFI2009-07648/FILO and CONSOLIDER-INGENIO 2010 Programme CSD2007-00012 (both awarded by the Ministerio de Ciencia e Innovación) and by project 2009 SGR 701 (awarded by the Generalitat de Catalunya).

Correspondence e-mail address: <joan.borras@upf.edu>

Notes

1. Authors refer to uncertainty with the term "feeling of knowing", which is defined as the ability to monitor the accuracy of one's own knowledge or the ability to monitor the feeling of knowing of someone else ("feeling of another's knowing") (see, e.g., Litman and Forbes-Riley 2009).
2. The term 'tone of voice' has to be understood in a non-technical way. In this experiment, subjects were asked to listen to a recording of a female saying the single word 'maybe' in three tones of voice conveying liking, neutrality and disliking.
3. Recent results from a Mismatch Negativity (MMN) analysis by Borràs-Comes et al. (2009) back up this analysis by finding a stronger MMN brain response when contrasting stimuli (statements versus questions; see examples in Figure 1) were presented than when listeners heard pairs of non-contrasting stimuli having the same physical distance between them (either two types of statements or two types of questions).
4. Of particular note is the work by Amades (1957), Mascaró (1978, 1981) and especially Payrató (1989, 1993), which contains a description of a repertoire of 221 emblems and pseudoemblems of Central Catalan. Since the 1990s, two projects lead by Lluís Payrató and financed by the varcom and pragmaestil programmes have analyzed the system of Catalan gestures but have mainly focused on coverbal manual gestures (see e.g. Payrató et al. 2004).
5. As reviewer Beatriz Raposo points out, there is a noticeable lip stretching in the case of the focus gesture. It is interesting to note that the gestural overarticulation of the segments in accented position – in our case, the vowel /i/ – is a common phenomenon among the production of contrastive focus (as described by Dohen and Lœvenbruck 2009; Prieto et al. 2011, and Borràs-Comes et al. submitted).

6. The target acoustic stimuli created for this experiment are identical to those used in Borràs-Comes et al. (2009) and in Borràs-Comes et al. (2010).
7. As for the specific result in the RT values in the incongruent stimulus audio 1 – video 6, we obtain, as Reviewer 1 points out, an unexpected result of a very low RT. This unexpected value is due to the deletion of the outliers for RT values (the ones that were at a distance of at least three standard deviations from the overall mean), which eliminated very high RT values and led, in this case, to an unexpected mean RT value.

References

Aguilar, Lourdes, Carme de-la-Mota & Pilar Prieto (coords.). 2009. *Cat_ToBI training materials*. http://prosodia.upf.edu/cat_tobi/

Amades, Joan. 1957. El gest a Catalunya. *Anales del Instituto de Lingüística* VI. 88–148.

Assmann, Peter & Quentin Summerfield. 2004. The perception of speech under adverse conditions. In Steven Greenberg & William A. Ainsworth (eds.), *Speech processing in the auditory system*, 231–308. New York: Springer Verlag.

Baker-Shenk, Charlotte L. 1983. *A microanalysis of the nonmanual components of questions in American Sign Language*. Berkeley: University of California.

Barkhuysen, Pashiera, Emiel Krahmer & Marc Swerts. 2005. Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication* 45(3). 343–359.

Bergman, Brita. 1984. Non-manual components of signed language: Some sentence types in Swedish Sign Language. In Filip Loncke, Penny Boyes Braem & Yvan Lebrun (eds.), *Recent research on European sign languages*. Lisse: Swets & Zeitlinger, 49–59.

Beskow, Jonas, Björn Granström & David House. 2006. Visual correlates to prominence in several expressive modes. *Proceedings of Interspeech 2006*, Dresden, 2–5 May 2006. 1272–1275.

Boersma, Paul & David Weenink. 2008. *Praat: Doing phonetics by computer* (version 5.0.09). http://www.fon.hum.uva.nl/praat/

Borràs-Comes, Joan, Jordi Costa-Faidella, Pilar Prieto & Carles Escera. Submitted. Specific neural traces for intonational discourse categories as revealed by human evoked potentials. *Journal of Cognitive Neuroscience*.

Borràs-Comes, Joan, Maria del Mar Vanrell & Pilar Prieto. 2010. The role of pitch range in establishing intonational contrasts in Catalan. *Fifth International Conference on Speech Prosody,* Chicago 14–11 May, 2010. Paper 100103: 1–4.

Borràs-Comes, Joan, Cecilia Pugliesi & Pilar Prieto. 2011. Audiovisual competition in the perception of counter-expectational questions. *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing*, Volterra, 31 August–3 September.

Breeuwer, M. & Plomp, Reinier 1984. Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the Acoustical Society of America* 76. 686–691.

Burnham, Douglas. 1998. Language specificity in the development of auditory–visual speech perception. In Ruth Campbell, Barbara Dodd & Douglas Burnham (eds.), *Hearing by eye II: advances in the psychology of speechreading and auditory–visual speech*. New York: Psychology Press. 29–60.

Calvert, Gemma A. & Ruth Campbell. 2003. Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience* 15(1). 57–70.

Cavé, Christian, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay & Robert Espesser. 1996. About the relationship between eyebrow movements and F0 variations. *4th International Conference on Spoken Language Processing* Philadelphia, 3–6 October 1996. 2175–2179.

Coerts, Jane. 1992. *Nonmanual grammatical markers: An analysis of interrogatives, negations and topicalisations in Sign Language of the Netherlands*. Amsterdam: Universiteit van Amsterdam dissertation.

Colin, C., M. Radeau, A. Soquet, D. Demolin, F. Colin & P. Deltenre. 2002. Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology* 113(4). 495–506.

Dijkstra, Christel, Emiel Krahmer & Marc Swerts. 2006. Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. *3rd International Conference on Speech Prosody* (SP) 2006.Paper 025.

Dohen, Marion. 2009. Speech through the ear, the eye, the mouth and the hand. In Anna Esposito, Amir Hussain, Maria Marinaro & Raffaele Martone (eds.), *Multimodal signals: Cognitive and algorithmic issues*, 24–39. Springer: Berlin/Heidelberg.

Dohen, Marion & Hélène Lœvenbruck. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech* 52(2–3). 177–206.

Dohen, Marion, Hélène Lœvenbruck & Harold Hill. 2006. Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variabilities. *Third International Conference on Speech Prosody*. 221–224.

Ekman, Paul & Wallace V. Friesen. 1978. *The facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, Paul, Wallace V. Friesen & Joseph C. Hager. 2002. *The facial action coding system,* 2nd edn *(CD-ROM)*. Salt Lake City: Research Nexus. http://face-and-emotion.com/dataface/facs/new_version.jsp

Graf, Hans Peter, Eric Cosatto, Volker Strom & Fu Jie Huang. 2002. Visual prosody: Facial movements accompanying speech. *5th IEEE International Conference on Automatic Face and Gesture Recognition* Washington, D.C. May 20–21 2002. 396–401.

Grant, Ken W. & Brien E. Walden. 1996. Spectral distribution of prosodic information. *Journal of Speech and Hearing Research* 39. 228–238.

Grant, Ken W., Brien E. Walden & Philip F. Seitz. 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America* 103. 2677–2690.

Grossman, Ruth B. 2001. *Dynamic facial expressions in American Sign Language: Behavioral, neuroimaging, and facial coding analyses for deaf and hearing participants*. Boston, MA: Boston University dissertation.

Grossman, Ruth B. & Judy Kegl. 2006. To capture a face: A novel technique for the analysis and quantification of facial expressions in American Sign Language. *Sign Language Studies* 6(3). 273–305.

Hadar, Uri., T. J. Steiner, E. C. Grant & F. Clifford Rose. 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech* 26. 117–129.

House, David. 2002. Perception of question intonation and facial gestures. *Fonetik* 44(1). 41–44.

Keating, Patricia A. 2004. Statistics. UCLA Phonetics Lab. http://www.linguistics.ucla.edu/faciliti/facilities/statistics/statistics.html

Krahmer, Emiel, Zsófia Ruttkay, Marc Swerts & Wieger Wesselink. 2002. Pitch, eyebrows and the perception of focus. *First International Conference on Speech Prosody*. http://aune.lpl.univ-aix.fr/~sprosig/sp2002/docs/papers.htm

Krahmer, Emiel & Marc Swerts. 2004. More about brows: A cross-linguistic analysis-by-synthesis study. In Zsófia Ruttkay & Catherine Pelachaud (eds.), *From brows to trust: Evaluating embodied conversational agents*, 191–216. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Krahmer, Emiel & Marc Swerts. 2005. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech* 48(1). 29–54.

Krahmer, Emiel & Marc Swerts. 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.

Kyle, Jim G. & Bencie Woll. 1985. *Sign language: The study of deaf people and their language*. Cambridge: Cambridge University Press.

Lapakko, David. 1997. Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education* 46. 63–67.
Litman, Diane & Kate Forbes-Riley. 2009. Spoken tutorial dialogue and the feeling of another's knowing. *The 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL) 2009*. 286–289.
MacDonald, John & Harry McGurk. 1978. Visual influences on speech perception processes. *Perception and Psychophysics* 24(3). 253–257.
Mascaró, Jaume. 1978. *Expresión y comunicación no verbal: Metodología y crítica*. Barcelona: Universitat de Barcelona dissertation.
Mascaró, Jaume. 1981. Notes per a un estudi de la gestualitat catalana. *Serra d'Or* 259. 25–28.
Massaro, Dominic W. 1987. *Speech perception by ear and by eye*. Hillsdale, NJ: Erlbaum.
Massaro, Dominic W. 1998. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.
Massaro, Dominic W. & Michael M. Cohen. 1993. The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General* 122(1). 115–124.
McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices: A new illusion. *Nature* 264. 746–748.
Mehrabian, Albert & Susan R. Ferris. 1967. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology* 31. 248–252.
Mori, Masahiro. 1970. The uncanny valley. *Energy* 7(4). 33–35.
Munhall, Kevin G., Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate & Eric Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science* 15. 133–137.
Nakano, Yukiko I., Gabe Reinstein, Tom Stocky & Justine Cassell. 2003. Towards a model of face-to-face grounding. *Annual Meeting of the Association for Computational Linguistics* (ACL) 2003). 553–561.
Payrató, Lluís. 1989. *Assaig de dialectologia gestual: Aproximació pragmàtica al repertori bàsic d'emblemes del català de Barcelona*. Barcelona: Universitat de Barcelona dissertation.
Payrató, Lluís. 1993. A pragmatic view on autonomous gestures: A first repertoire of Catalan emblems. *Journal of Pragmatics* 20. 193–216.
Payrató, Lluís, Núria Alturo & Marta Payà (eds.). 2004. *Les fronteres del llenguatge: Lingüística i comunicació no verbal* (Col·lecció Lingüística Catalana, 7). Barcelona: Promociones y Publicaciones Universitarias (PPU).
Pfau, Roland & Josep Quer. 2010. Nonmanuals: their grammatical and prosodic roles. In Diane Brentari (ed.), *Sign languages* (*Cambridge Language Surveys*), 381–402. Cambridge: Cambridge University Press.
Prieto, Pilar. In press. The intonational phonology of Catalan. In Sun-Ah Jun (ed.), *Prosodic typology 2*. Oxford: Oxford University Press.
Prieto, Pilar, Cecilia Pugliesi, Joan Borràs-Comes, Ernesto Arroyo & Josep Blat. 2011. Crossmodal prosodic and gestural contribution to the perception of contrastive focus. *12th Annual Conference of the International Speech Communication Association* Florence, 28–31 August 2011.
Psychology Software Tools Inc. 2009. *E-Prime* (version 2.0). Sharpsburg, PA.
Scarborough, Rebecca, Patricia Keating, Sven L. Mattys, Taehong Cho & Abeer Alwan. 2009. Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech* 52(2–3). 135–175.
Sourcetec Software Co. 2007. *Sothink SWF Quicker* (version 3.0). Wuhan.
Srinivasan, Ravindra J. & Dominic W. Massaro. 2003. Perceiving from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech* 46(1). 1–22.
Sumby, William H. & Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26. 212–215.

Summerfield, Quentin. 1992. Lipreading and audio-visual speech perception. In Vicki Bruce, Alan Cowey, W. Ellis & Dawid I. Perrett (eds.), *Processing the facial image*, 71–78. Oxford: Oxford University Press.

Swerts, Marc & Emiel Krahmer. 2004. Congruent and incongruent audiovisual cues to prominence. *Second International Conference on Speech Prosody (SP) 2004*. http://aune.lpl.univ-aix.fr/~sprosig/sp2004/docs/papers_authors.html

Swerts, Marc & Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53(1). 81–94.

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

de Vos, Connie, Els van der Kooij & Onno Crasborn. 2009. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in Sign Language of the Netherlands. *Language and Speech* 52. 315–339.