Crossmodal prosodic and gestural contribution to the perception of contrastive focus

Pilar Prieto,[2,1] *Cecilia Pugliesi,*[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

10.21437/Interspeech.2011-397

Abstract

Speech prosody has traditionally been analyzed in terms of acoustic features. Even though visual features and gestures have been shown to help and enhance linguistic processing, the conventional view is that facial and body gesture information in oral (non-sign) 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 conflicting auditory and visual information to investigate two related questions regarding the importance of gestures in conveying prosodic meaning: (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 nod and eyebrow raising) for the perception of this type of focus? Our findings reveal that the visual component is crucial in the semantic interpretation of contrastive focus.

Index terms: facial gestures, statement, contrastive focus, eyebrow, head-nod, talking heads, audiovisual prosody.

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), as well as 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, which can be very helpful in the processing of segmental and prosodic information during online speech communication.

Classic experiments have shown that visual cues contribute to speech intelligibility and the detection of segmental information in noise conditions [1, 2]. On the other hand, other studies have shown that visual features can also be successfully used to identify prosodic information, such as stress or which word in a sentence is emphasized [3]. Prominence is indeed one of the prosodic functions that has been shown to be strongly correlated with facial movements. It has commonly been reported that prominent words are usually marked by means of facial expressions, such as eyebrow movements, or by more exaggerated movements of the articulators [4, 5]. Al Moubayed & Beskow [6] showed that word recognition and intelligibility with animated talking heads was increased when focally-accented (prominent) words were supplemented with head-nods or eyebrow raising gestures. All in all, though the majority of studies suggest that visual facial dynamics convey crucial prosodic information that may improve the perception of focus in conversational situations, it is unclear how important visual cues are compared to auditory cues for the perception of focus [7, 8, 9].

The aim of this work is 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. Previous work on how our perceptual system integrates incongruous auditory speech information and visual cues from a speaker's face has concentrated on effects at the segmental level [10, 11], and our knowledge of audiovisual interactions in the presence of conflicting and non-conflicting information at the prosodic level is more limited (e.g., [4, 5]). In the current study, Experiment 1 will address the potential influence of the presence of facial markers on perception of contrastive focus when they are competing with auditory information.

On the other hand, not much is known about the extent to which specific facial cues are able to affect the linguistic perception of prominence and focus. In general, production studies have claimed that head and eyebrow movements are correlated with prominence marking [4, 5, 6]. Recent work has investigated which visual features are the most effective perceptually in conveying prominence and focus in speech. Though results are partially contradictory, it seems that the visual cues of the upper face (eyebrows) and the head (head movements) are quite important. For instance, Scarborough et al. [9] assessed different facial measures during speech production and the extent to which these correlate with phrasal stress perception in English. While measures such as head and eyebrow movements were correlated with perception performance, chin movement contributed the most to correct perception independently of the other measures, thus suggesting that this is the most effective visual cue to stress. Swerts & Krahmer [5] 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 (eyebrow movements) has stronger cue value for prominence detection than the lower area. The second aim of the present paper is to further investigate which facial movements (namely head nods and raised eyebrows) play the strongest role in the perception of contrastive focus. In Experiment 2, participants were presented with conflicting visual-only information combining four different activations of eyebrow raising and head forward movement of a 3D animated avatar that ranged from the typical configuration of a statement to that of a 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 very difficult to achieve with human actors. The task of the participants was again to indicate for each stimulus whether they perceived the utterance as being a neutral statement or a 'correction' (i.e., contrastive focus) statement.

By enlarging the empirical domain to Catalan, this study contributes to our knowledge of potential crosslinguistic

Copyright © 2011 ISCA

differences in the gestural and intonation marking of contrastive focus. Further, the use of controlled manipulations of intonation and gestural variation allows us to provide more focused analyses of the contribution of visual/intonation features to the perception of focus.

The article is organized as follows. In sections 2 and 3 we will present the methodology and results for both experiments. Finally, we will conclude the article in section 4 with a general discussion about the implications of these results for an audiovisual model of prosody perception.

2. Experiment 1: AV integration

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

2.1. Methodology

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

In Catalan, a pitch range difference in a rising-falling nuclear configuration is the main intonational cue for the distinction between NFS and CFS [13, 14]. To prepare the target auditory stimuli for Experiment 1, we took an audio recording of a contrastive focus statement and manipulated the pitch by means of Praat [15]. By modifying the F0 peak height in 4 steps (distance between each one = 1.5 semitones), we created a synthesized continuum ranging from the typical pitch contour for a NFS to the typical pitch contour of a CFS. A schematic diagram of these manipulations is shown in Fig. 1.

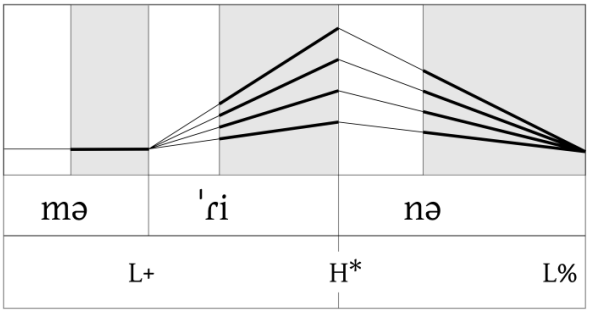


Figure 1. *Schematic diagram showing pitch manipulation.*

Concerning facial gestures, NFS and CFS share the same gestural pattern but show different degrees of activation, ranging from a neutral state (see Fig. 2, left panel) to a more intensely expressive one involving the raising of the eyebrows and a forward inclination of the head (as in Dutch [5]; see Fig. 2, last line, centre). To obtain the visual materials, four subjects (two males and two females) were videotaped in high definition video while producing natural productions of the noun phrase *Marina* with either a NFS or a CFS meaning.[1]

[1] In order to prompt the corresponding answer, subjects were asked to act out the replying speaker in each of the two short dialogues in (1), with (1a) involving a NFS and (1b) exemplifying a CFS.

Based on the recorded videos, the key frames for each gestural movement (the starting and ending points of every transition) were coded for each visual variable. The mean values of the key frames were then used to animate the 3D avatar, using the NINOs Platform, which allows for both assisted and animated production of 3D animated media [16]. Figure 2 shows the transition in four steps from the NFS facial configuration (first row) to the CFS facial configuration (fourth row) in four steps. These stills correspond to the key frames in initial position (first column), eyes completely closed (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), for the four videos.

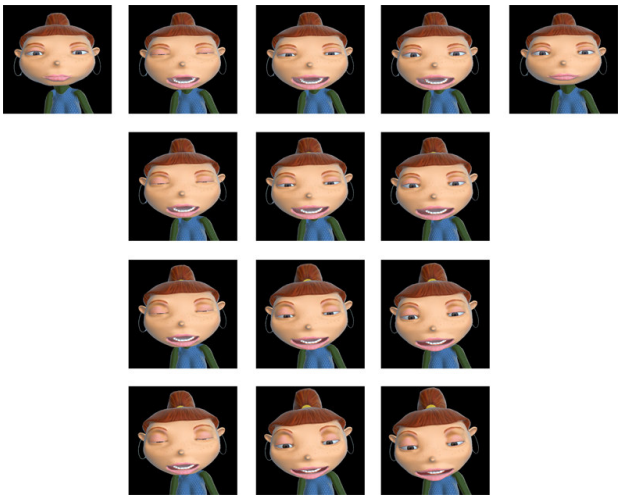


Figure 2. *Facial gesture transition from NFS configuration (first row) to CFS (fourth row) in four steps. The stills correspond to the five key frames in each of four sequences: in initial position (first column), with eyes completely closed (second column), with eyebrows fully raised (third column), with head completely down (fourth column) and in final position (fifth column). The initial and final frames are the same for all four sequences.*

For the audiovisual materials, the 4-step acoustic continuum of the rising-falling intonation contour (see Fig. 1) was crossed with the 4-step continuum of the visual materials created with the animated character (ranging from a facial neutral state which characterizes NFS to a CFS state; Fig. 2).

Eighteen participants were presented with these AV combinations. They were asked to indicate which interpretation (NFS vs. CFS) was more likely for each stimulus by pressing the corresponding computer key, 'A' for NFS ('statement' in Catalan is 'afirmació') and 'C' for CFS (in Catalan, 'correcció'). A total of 1440 responses were obtained for this experiment (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

(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!

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 = 42.451, p < .001) and also a main effect of gesture (F = 64.034, p < .360), with no interaction between the two (F = 1.099, p = .874). Figure 3 shows the identification functions obtained for Experiment 1. The y-axis represents the proportion of CFS identifications within each visual stimulus (different lines) and auditory stimulus (x-axis). Crucially, clear gestural cues combined with appropriate acoustic cues lead to accurate identification responses, while more conflicting gestural and acoustic cues lead to chance-level scores.

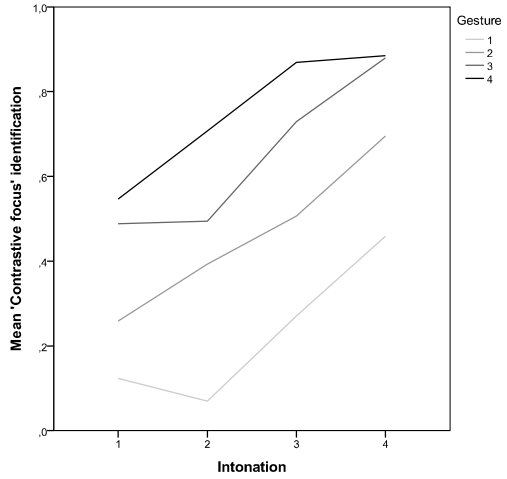


Figure 3. *Mean CFS identification (y-axis) as a function of each particular continuum of visual information (lines) and auditory information (x-axis).*

Another GLMM analysis was conducted with reaction times 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 = 2.316, p = .074), nor gesture (F = 2.197, p = .087) nor the interaction between intonation and gesture (F = 0.879, p = .544), revealing that different combinations of gesture and intonation contours in these cases do not lead to significant differences in reaction times. The fact that the processing time was not reduced in any of the conditions suggests that listeners are not especially puzzled by ostensibly incongruent head nod and eyebrow visual combinations and indeed do not interpret them as incongruent AV combinations.

3. Experiment 2: visual cues

Experiment 2 investigated the extent to which different activations of the gestural cues (namely combinations of competing eyebrow and head movements) are 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 not exposed to audio information.

3.1. Methodology

For the video materials for this experiment, each of the two gestural cues (eyebrow-raising and head inclination) was presented in a continuum of 4 degrees of activation, from less pronounced to more pronounced eyebrow raising, for example, with all the possible combinations between them. Figure 4 shows the peak gesture combination in each of the 16 videos used in Experiment 2 (4 activations of eyebrow raising represented from left to right × 4 activations of head forward represented from top to bottom).

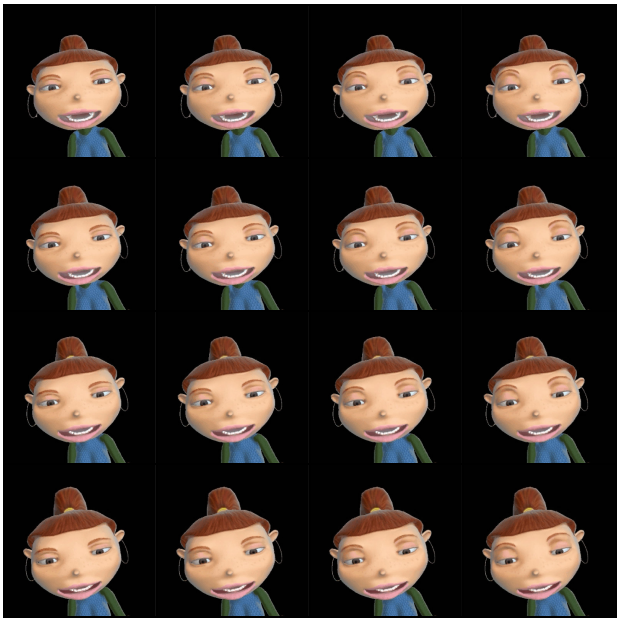


Figure 4. *Stills showing peak gesture combinations in the 16 videos used in Experiment 2. The four activations of eyebrow raising are represented from left to right (from lower to higher degrees of intensity) and the four variations of head inclination from top to bottom.*

The same 18 Catalan participants took part in Experiment 2. The procedure and instructions were the same as in Experiment 1, with the difference that subjects had to judge each video as portraying either a NFS or a CFS without hearing any audio track. A total of 1440 responses were 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 gestures (4 levels) as fixed factors, and subject and block as crossed random factors. Main effects were found for eyebrow (F = 8.480, p < .001) and head (F = 124.038, p < .001) gesture, and their interaction was not statistically significant (F = 1.832, p = .058). Figure 5 shows the mean CFS identification rate (y-axis) as a function of eyebrow activation (x-axis) and head inclination (lines). In general, the effect of head movement is much stronger than that of eyebrow movements, especially when head nods are 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 this we can conclude that the head nod gesture is especially relevant in the conveyance of contrastive focus.

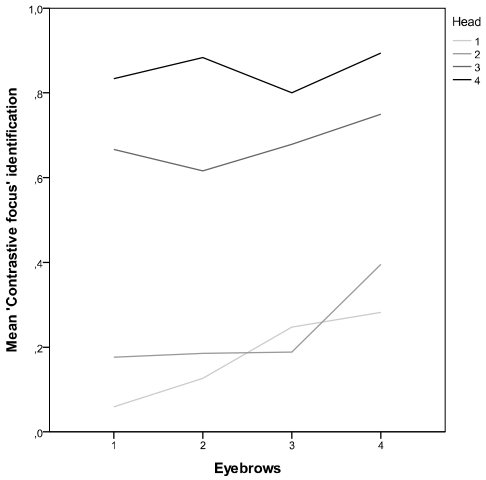


Figure 5. *Mean CFS identification (y-axis) as a function of head inclination (lines) and eyebrow raising (x-axis).*

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

4. Discussion and conclusions

Experiment 1 analyzed the relevance of AV cues for the detection of contrastive focus as compared to narrow focus statements. Listeners were presented with an acoustic continuum with intonation ranging from a typical NFS to a typical CFS. The audio continuum was juxtaposed over a visual continuum of facial gestures that ranged from the gestural sequence characteristic of NFS to that characteristic of a CFS. The results showed a main effect of gesture ($F = 64.034$) and, less powerfully, a main effect of intonation ($F = 42.451$) (in both cases, $p < .001$). Listeners were more accurate in their detection of contrastive foci when clear gestures involving eyebrow raising and head inclination were combined with intonation patterns involving the highest increase in pitch range. In the process of detecting foci, visual cues are thus somewhat more powerful than acoustic ones. Interestingly, the analysis of reaction times revealed no effect of intonation or gesture, and no interaction between the two factors. We argue that these reaction time results indicate that none of the AV combinations presented to the listeners was perceived as incongruent, as listeners can use acoustic and visual information independently of each other. This is consistent with other research whose production results indicate that the visual marking of focus through eyebrow raising and/or a head nod is optional and displays a high degree of inter- and intra-speaker variation (e.g., [4, 12]).

Experiment 2 analyzed the contribution of two visual cues to contrastive focus (namely, eyebrow raising and head inclination) by varying the relative strength of these two cues in a visual-only experiment. This study shows how the use of an animated talking head can allow us to entirely control the different visual cues to prosodic focus. The results reveal that head movements are stronger perceptual cues of contrastive focus than eyebrow raising, possibly due to the stronger visual perceptibility of this gesture.

In sum, our findings lend support to the view that the visual component does not merely accompany acoustic prosodic information but is a crucial component in the semantic interpretation of contrastive focus.

5. Acknowledgements

We are grateful to S. Frota, E. Rodero, and M. Swerts for 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 and CONSOLIDER-INGENIO 2010 'Bilingüismo y Neurociencia Cognitiva CSD2007-00012' and TIN2008-05995 Learn3 CICYT (Spanish Ministry of Science and Innovation), 2009 SGR 701 (Generalitat de Catalunya), and iMP (European Union 7th Framework project).

6. References

[1] Sumby, W. H., & Pollack, I., "Visual contribution to speech intelligibility in noise", Journal of the Acoustical Society of America, 26: 212-215, 1954.

[2] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E., "Visual prosody and speech intelligibility – Head movement improves auditory speech perception", Psychological Science, 15(2): 133-137, 2004.

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

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

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

[6] Al Moubayed, S. & Beskow, J., "Effects of visual prominence cues on speech intelligibility", In AVSP-2009, 43-46, 2009.

[7] Dohen, M., "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, 24-39. Springer: Berlin/Heidelberg, 2009.

[8] Foxton, J.M., Riviere, L.-D. & Barone, P., "Cross-modal facilitation in speech prosody", Cognition 115: 71-78, 2010.

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

[10] McGurk, H., & MacDonald, J., "Hearing lips and seeing voices: A new illusion", Nature, 264: 746-748, 1976.

[11] Summerfield, Q. "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, 1992.

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

[13] Borràs-Comes, J., Vanrell, M. M., & Prieto, P. ,"The role of pitch range in establishing intonational contrasts in Catalan", Proceedings of the Fifth International Conference on Speech Prosody (100103:1-4). Chicago, 2010.

[14] Borràs-Comes, J., Costa-Faidella, J., Prieto, P., & Escera, C., Encoding of intonational contrasts as revealed with the mismatch negativity (MMN). Fourth European Conference on Tone and Intonation. Stockholm, 2009.

[16] Abadia, J., Evans, A., Gonzales, E., Gonzales, S., Soto, D., Fort, S., Romeo, M., & Blat, J.: Assisted animated production creation and programme generation. Proceedings of the International Conference on Advances in Computer Enterntainment Technology (ACE '09). ACM, New York, NY, USA, 207-214, 2009.