ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2011
Volterra, Italy
September 1-2, 2011

Audiovisual perception of counter-expectational questions

Joan Borràs-Comes,[1] *Cecilia Pugliesi, Pilar Prieto* [2,1]

[1] Universitat Pompeu Fabra, Barcelona, Spain
[2] ICREA, Barcelona, Spain
{joan.borras; pilar.prieto}@upf.edu ceciliapuglesi@gmail.com

Abstract

The precise nature of the interaction between acoustic and visual information in the perception of prosodic information is a question that still remains unclear. Though the fuzzy logical model of perception (FLMP) has been shown to explain the recognition of segmental information, this model also needs to be tested in the field of suprasegmentals such as facial gestures. The first goal of this paper is to investigate, by means of a computer-generated 3D video character, the interaction between intonational and gestural information in the detection by listeners of counter-expectational questions compared to narrow focus statements. The second goal is to test which specific facial gesture conveys the counter-expectation meaning most clearly. Our results represent a further step for considering an FLMP approach to the analysis of audiovisual prosody.

Index Terms: audiovisual prosody, intonation, facial gestures, eyebrow, models of perception.

1. Introduction

Several studies have shown that our perceptual system integrates auditory speech information and visual cues from the speakers' face. In a classic study, McGurk & MacDonald [1] showed that perceptual confusions between consonants are different and complementary in the visual and auditory modalities, demonstrating that speech perception is multisensorially integrated. Though crossmodal integration has been studied at the segmental level, our knowledge of audiovisual interactions in the perception of congruent and incongruent prosodic information is more limited.

Dijkstra, Krahmer & Swerts [2] tested which cues participants used to assess the degree of certainty of a person answering factual questions. The auditory-visual (AV) presented materials contained prosodic cues such as fillers ("uh"), rising intonation contours or marked facial expressions, artificially manipulated in such a way that all possible combinations of the cues could be judged by participants. All three factors had a significant influence on the perception results, but facial expressions had by far the largest effect. Given that their results were in line with what has been found for the perception of emotion (in the sense that they found a stronger effect of visual cues; see [3]), they argued that uncertainty is a "social emotion" that plays a role in human communication.

Srinivasan & Massaro [4] looked at how prosody was processed through acoustic and visual information with the aim of testing two quantitative models of perception: the weighted averaging model of perception (WTAV), which predicts that the different sources of information "are averaged according to the weight assigned to each modality" ([4: 10]), and the fuzzy logical model of perception (FLMP), which predicts that the influence of one modality will be greater to the extent that the other is weaker and more ambiguous. They analyzed the perception of questions and statements presenting subjects with auditory and visual (facial) information. They found significant effects for both auditory and visual cues, but the effect of visual information was found to be very small (which is consistent with findings reported in [5]). Consequently, auditory information had a similar effect across the different visual materials, so the data could be explained equally well by either the WTAV or the FLMP (models which are particularly difficult to distinguish if one of the sources of information has a relatively small influence, because their predictions would coincide). The authors argued that the weaker influence of the visible dimension of speech merely meant that the auditory dimension was more informative, but this does not mean that perceivers do not also use visual information to distinguish between questions and statements. Finally, they hypothesized that a shorter test stimulus, consisting of only one word (and not a whole sentence like in their study) might elicit an optimal bimodal integration process and thereby make statement/question identification a more automatic perceptual task and less of a cognitive decision-making process ([4: 20).

Borràs-Comes & Prieto [6] explored the relative importance of pitch accent contrasts and facial gestures in the distinction between two gesturally-marked meanings (i.e., contrastive focus statements and counter-expectational questions) by using a continuum of congruent and incongruent multimodal stimuli. Though listeners paid more attention to the visual component of the AV materials, effects of audiovisual integration were found. Specifically, intermediate visual stimuli caused the acoustic signals to have a greater impact on listener judgments. The authors argued that this was consistent with the FLMP model ([7]).

In the current paper we have the goal of further testing this hypothesis — i.e., that the influence of one modality is greater to the extent that the other is ambiguous — by investigating the audiovisual perception of two linguistic meanings which are produced gesturally in an asymmetric way. Specifically, while narrow focus statements (e.g., ["*What's her name?"*] *"Mary"*) can generally be marked by an optional head-nod, the facial cues for counter-expectational questions (e.g., ["*Her name is Mary"*] *"Mary???"*), which indicate incredulity on the part of the speaker, are emotionally-like, more communicatively-salient marked, and so much stronger (i.e., furrowing the brows, squinting the eyes and moving the head backward; see [6]). Thus in this case according to the FLMP theory we would expect listeners to rely more heavily on acoustic information when processing the weak gestural patterns typical of narrow focus statements than when processing the strong gestural patterns of counter-expectational questions.

The paper reports on the results of two experiments. First, Experiment 1 investigated the interaction between AV cues for the detection of counter-expectational questions (henceforth CEQ) compared to narrow focus statements (henceforth NFS). For this purpose, an acoustic continuum ranging from a typical production of a NFS to a typical production of a CEQ (in both cases a rising pitch accent followed by a low boundary tone; see Fig. 2) was presented to Catalan listeners in co-occurrence with a visual continuum of facial gestures presenting different levels of activation of an incredulity face. Second, Experiment 2 investigated which facial movements were more informative in distinguishing the two meanings (i.e., brow furrowing, eyelid closure and backward head movement). In both experiments, the visual materials involved a computer-generated 3D animated character, which allowed us to finely control the target visual movements. Thus, we also claim that our results are useful to the design of virtual agents.

2. Experiment 1: AV integration

2.1. Methodology

In Experiment 1, a 4-step acoustic continuum was crossed with a 4-step continuum of the visual materials. Eighteen participants were presented with these AV combinations and were asked to indicate which interpretation (NFS vs. CEQ) was more likely for each one. Each stimulus lasted one second of duration.

In terms of facial gestures, while NFS are produced with a neutral facial gesture (with only the specific lip-sync and natural eye blinks), CEQ are generally produced with a very specific pattern of facial gestures. In our materials, the NFS condition only included lip-sync movements, but the CEQ included three additional parameters: eyebrow furrowing, eye squinting (eyelid closure), and a head backward movement. The final alignment between these gestural movements and the acoustic chain was based on previous production data and is shown in Figure 1.

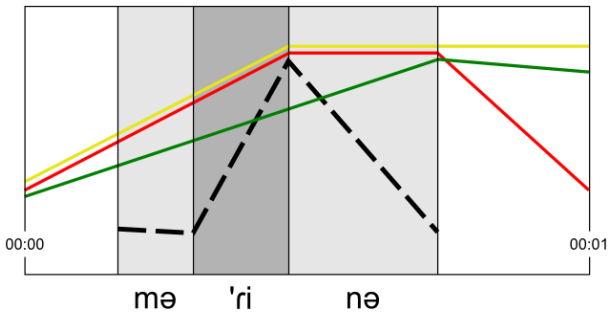

Figure 1. *Idealized activation of each facial movement in the 3D animated character and their alignment with the sound chain in the CEQ condition (yellow line: eyebrow furrowing; red line: eyelid closure; green line: backward head movement; dashed black line: pitch contour).*

Thus the visual information in Experiment 1 appeared in a 4-step continuum, with the animated video sequence showing the character maintaining a neutral expression throughout the utterance at one end and expressing incredulity at the other, and two intermediate video sequences showing gestures between "neutrality" and incredulity (see Fig. 2). This 4-step scale thus differed from the two-way contrast between two specifically marked gestural configurations used in [6].

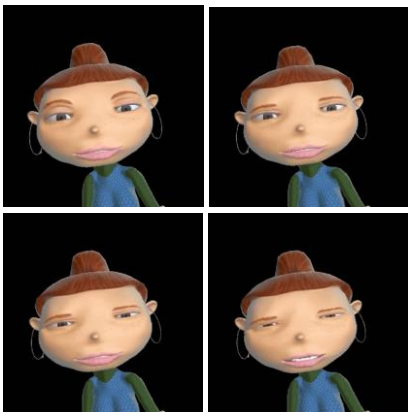

Figure 2. *Stills from the apex of the gesture in the four visual stimuli used in Experiment 1, ranging from the purely neutral facial expression seen in NFS (top-left) to the incredulity expression characteristic of CEQ (bottom-right graph).*

In regard to the intonational information, a pitch range difference in a rising-falling nuclear configuration is the main intonational cue for the distinction between NFS and CEQ in Catalan ([8, 9]; see Fig. 3). The auditory continuum was based on a natural production of a female native speaker of central Catalan of the proper noun *Marina* that was then adapted — using *Praat* ([10]) — to the mean F0 values of a set of NFS and CEQ natural productions.

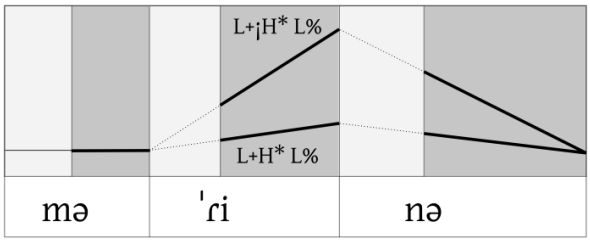

Figure 3. *Idealized intonational contours for the proper name 'Marina' uttered as an NFS (L+H* L%) and CEQ (L+¡H* L%) .*

Therefore, in this experiment, while intonation provided a clear contrast between two well-known nuclear configurations (L+H* L% for NFS, and L+¡H* L% for CEQ), visual information presented different levels of activation of an incredulity gestural pattern.

A set of eighteen native speakers of Central Catalan participated in Experiment 1. All subjects were undergraduates studying journalism or translation at the Campus de la Comunicació of the Universitat Pompeu Fabra, and were paid for their participation. Stimuli were presented to subjects over headphones and a computer screen. They were instructed to pay attention to the audiovisual materials as a whole and indicate which interpretation was more likely for each stimulus.

The task consisted of 5 blocks in which all AV stimuli were presented to the subjects in a randomized order. We thus obtained a total of 1,440 responses for Experiment 1 (4 audio × 4 video × 5 blocks × 18 listeners).

The experiment was set up by means of the psychology software E-prime version 2.0 ([11]), and response frequencies and response times (RTs) were automatically recorded. Subjects were instructed to press the button as quickly as they could. 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. The experiment was set up in such a way that the next stimulus was presented only after a response had been given.

All responses and RT measures were analyzed using a Generalized Linear Mixed Model (GLMM) analysis through IBM SPSS Statistics 19.0 ([12]). GLMM can be also applied to analyze binomial responses and can control for both fixed and random factors. In our analyses, both subject and block of repetition were set as crossed random effects, thus avoiding at the same time inter-subject variation and possible effects of fatigue, boredom, and practice.

2.2. Results

The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 1,440 datapoints, 98 cases (RT ≥ 2792) were treated as outliers, i.e., those 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.

Figure 4 shows the mean 'CEQ' identification responses (y-axis) as a function of visual information (lines) and auditory information (x-axis). The graph shows that the effect of the acoustic information decreases as the visual information shows a clearer pattern of the incredulity gesture.

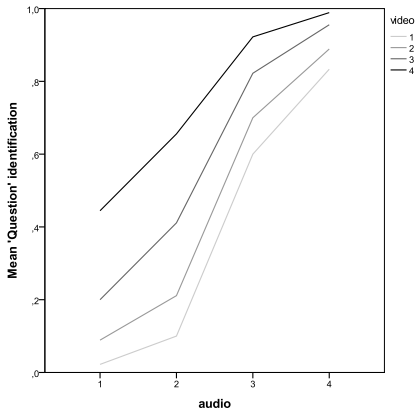

Figure 4. *Mean 'Question' identification (y-axis; 0 = NFS, 1 = CEQ) as a function of visual (lines; 1 = NFS, 4 = CEQ) and auditory information (x-axis; 1 = NFS, 4 = CEQ).*

The subsequent GLMM analysis was set up with the identification rate as the dependent variable, intonation, gesture and their interaction as fixed factors and subject and block as crossed random factors. Main effects of intonation ($F_{3, 1424}$ = 96.724, $p < .001$) and gesture ($F_{3, 1424}$ = 27.131, $p < .001$) were found, with no interaction between the two ($F_{9, 1424}$ = 0.502, $p = .874$). We then applied the Fisher F-test, which can assess whether the expected values of a quantitative variable within several fixed factors differ from each other. Crucially, an analysis of the main effect of auditory information within each visual material revealed that the effect of intonation decreased as the visual CEQ gesture was more available: video 1[NFS] ($F_{3, 1424}$ = 149.292, $p < .001$), video 2 ($F_{3, 1424}$ = 134.994, $p < .001$), video 3 ($F_{3, 1424}$ = 103.114 $p < .001$), and video 4[CEQ] ($F_{3, 1424}$ = 41.885, $p < .001$).

Another GLMM analysis was conducted with RT as the dependent variable and the same fixed and random factors as above. A main effect was found for intonation ($F_{3, 1424}$ = 10.305, $p < .001$) but not for gesture ($F_{3, 1424}$ = 1.039, p = .374). Interestingly, their interaction was found to be significant ($F_{9, 1424}$ = 5.112, $p < .001$).

Figure 5 shows the mean RT (y-axis) as a function of visual information (lines) and auditory information (x-axis). The graph shows, on the one hand, a decrease in RT as the representative movie for CEQ is matched with congruent auditory information; on the other, longer RT for the central auditory stimuli when they co-occurred with the most representative NFS movies.

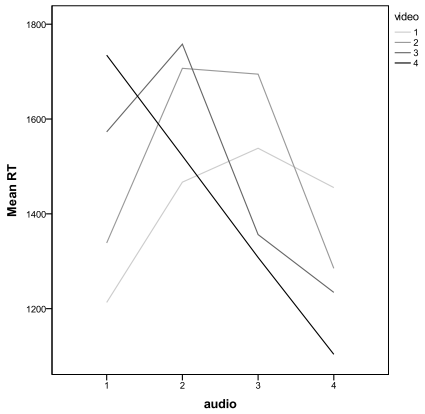

Figure 5. *Mean RT (y-axis; 0 = NFS, 1 = CEQ) as a function of visual (lines; 1 = NFS, 4 = CEQ) and auditory information (x-axis; 1 = NFS, 4 = CEQ).*

3. Experiment 2: visual cues

3.1. Methodology

For Experiment 2, a series of 64 visual stimuli was created using the same 3D animated character. Each of these 64 animated sequences depicted the three gestural elements implicated in the conveyance of incredulity (brow furrowing, eyelid closure and backward head movement) in 4 different levels of activation, in all possible combinations.

The same set of eighteen native speakers participated, though data from one subject was lost due to technical problems. Subjects were presented with the stimuli — this time exclusively visual, with no accompanying auditory stimuli — on a computer screen and were told to indicate which interpretation was more likely for each stimulus by pressing the same two keys used in Experiment 1. The task consisted of 5 blocks in which all stimuli were presented to the subjects in a randomized order. We thus obtained a total of 5,440 responses for Experiment 2 (4 brow × 4 eyelid × 4 head × 5 blocks × 17 listeners). Since Experiment 1 tested auditory-visual integration and Experiment 2 focused only on visual information, participants were always presented first with Experiment 1 and then — after they had viewed a non-related 4-minute documentary — were immediately presented with Experiment 2. The reason for having subjects always do Experiment 1 first was that we did not want subjects to be influenced by gestures, since Experiment 2 represented a visual-only task.

3.2. Results

The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 5,440 datapoints, 362 cases (RT ≥ 2454) were treated as outliers, i.e., those 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.

Figure 6 shows the mean 'CEQ' identification responses (y-axis) as a function of two of the gestural conditions: brow furrowing (lines) and backward head movement (x-axis). The graph shows an effect of brow furrowing, which overrides the effect of backward head movement. However, the effect of backward head movement seems to be greater when eyebrow information is weaker (eyebrow = 1[NFS]).

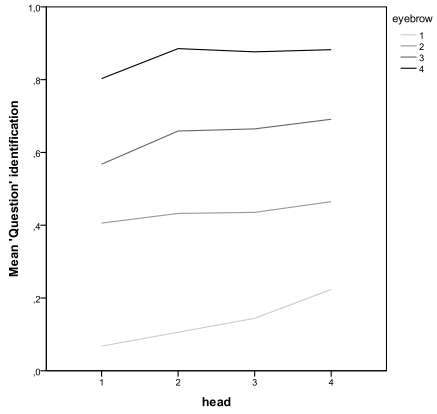

Figure 6. *Mean 'Question' identification (y-axis; 0 = NFS, 1 = CEQ) as a function of brow furrowing (lines; 1 = NFS, 4 = CEQ) and backward head movement (x-axis; 1 = NFS, 4 = CEQ).*

Figure 7 shows the mean 'Question' identification responses (y-axis) as a function this time of brow furrowing (lines) and eyelid closure (x-axis). The graph shows an effect of brow furrowing. Again, the effect of eyelid closure seems to increase when eyebrow information is more ambiguous (eyebrow = 2, eyebrow = 3).

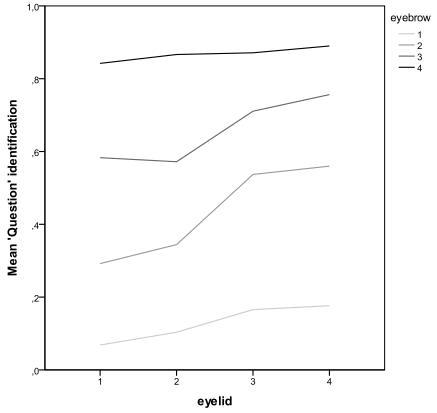

Figure 7. *Mean 'Question' identification (y-axis; 0 = NFS, 1 = CEQ) as a function of brow furrowing (lines; 1 = NFS, 4 = CEQ) and eyelid closure (x-axis; 1 = NFS, 4 = CEQ).*

Once more, a GLMM analysis was conducted, with identification rate as the dependent variable, brow furrowing, eyelid closure, backward head movement and all their possible combinations as the fixed factors, and subject and block as random factors. Main effects of brow ($F_{3, 5376} = 378.384$, $p < .001$), eyelid ($F_{3, 5376} = 29.458$, $p < .001$) and head ($F_{3, 5376} = 16.805$, $p < .001$) were found. In addition, the three paired interactions were also found to be statistically significant: brow*eyelid ($F_{9, 5376} = 2.800$, $p = .003$), brow*head ($F_{9, 5376} = 2.611$, $p = .005$), and eyelid*head ($F_{9, 5376} = 2.436$, $p = .009$). However, their triple interaction was not found to be significant ($F_{27, 5376} = 0.809$, $p = .746$).

Further analyses within each gestural condition revealed that the effect of backward head movement was greater when brow furrowing = 1[NFS] ($F_{3, 5376} = 11.462$, $p < .001$). Moreover, the effect of eyelid closure was greater when backward head movement = 1[NFS] ($F_{3, 5376} = 19.966$, $p < .001$).

Another GLMM analysis was conducted with RT as the dependent variable and the same fixed and random factors as above. A main effect was found for brow furrowing ($F_{3, 1424} = 16.138$, $p < .001$) but not for head ($F_{3, 1424} = 1.947$, $p = .114$) nor for eyelid closure ($F_{3, 1424} = 0.764$, $p = .514$). Two of the three paired interactions were also statistically significant, namely brow*eyelid ($F_{9, 5376} = 4.246$, $p < .001$) and brow*head ($F_{9, 5376} = 4.174$, $p < .001$), but this was not the case for eyelid*head ($F_{9, 5376} = 0.534$, $p = .851$). Their triple interaction was again not found to be significant ($F_{27, 5376} = 1.293$, $p = .142$). Crucially, the fact that eyebrow furrowing is the only statistically significant gestural factor in our experiment suggests that this specific facial gesture is especially relevant in the conveyance of incredulity in human face-to-face communication.

4. Discussion and conclusions

Experiment 1 investigated the interaction between auditory and visual cues in distinguishing between NFS and CEQ. The results indicated main effects of intonation and gesture. Crucially, the effect of intonation decreased as the visual CEQ information was more available, meaning that intonation had a less powerful effect as the incredulity gesture became clearer. The analysis of reaction times revealed a main effect of intonation and interaction between intonation and gesture, which confirms auditory and visual stimuli work together to elicit meaning in the listener.

Experiment 2 tested which of three gestural cues played the greatest role in discriminating between NFS and CEQ. Our results showed that brow furrowing had by far the largest effect, and there was also a significant effect of eyelid closure, backward head movement and all paired interactions. Further analyses within each gestural condition revealed that the effect of backward head movement was greater when brow furrowing showed a NFS configuration; and the effect of eyelid closure was greater when backward head movement showed a NFS configuration. When reaction times were analyzed, only the specific conditions involving brow furrowing were found to be significant. These results mean that listeners clearly rely on brow furrowing information when distinguishing a CEQ from a NFS when dependent on facial gesture information, but that they also integrate the different gestures available in their interlocutor's face in order to detect the incredulity characteristic of CEQ, especially when the main conveyor (i.e., brow furrowing > backward head movement > eyelid closure) is unclear.

Concerning the theories of speech perception (see [7, 13, 4]), 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) predicts that the sources are averaged according to the weight assigned to each modality. On the other hand, the fuzzy logical model of perception (FLMP) predicts that the influence of one modality will be greater than the other when the latter is weaker or more ambiguous. Therefore, the WTAV predicts that acoustic information will have a constant effect across visual stimuli, while the FLMP predicts that acoustic

information will have a diminished effect as the gestural becomes less ambiguous.

The results from our Experiment 2 could be explained by the WTAV, since — as [13] points out — "it is also possible to formulate a weighted average model in which the weight is a function of the ambiguity of a source of information" ([13: 61]), but this conclusion seems premature for the results of Experiment 1. It is true that one can claim that the effect of intonation decreases as the incredulity gesture becomes more activated (and so less ambiguous), but more research is needed to discard the need for an FLMP approach since the effect of audiovisual interaction found in the RT analysis. Moreover, even if NFS are characterized by a non-marked gestural configuration, this still remains as their typical gestural pattern and cannot be simply considered a "non-activated" CEQ gestural pattern.

The study conducted in [15] compared the audiovisual perception of a gradient prosodic contrast, i.e., NFS compared to contrastive focus statements (e.g., ["*Her name is Julia, right?"*] *"Mary!!!"*) by using a very similar experimental design. Their results were clearly in line with the WTAV, with no effects for RT for any of the fixed factors. We argue that our results are affected by the gestural nature of the two gestural configurations tested, with the NFS with a very neutral nature and the CEQ with a very marked one. We claim that this fact could have affected our results in an undesirable way and that comparing the difference between two marked facial configurations (as in [6]) could shed some light about the role of visual prosody and their interaction with the intonational features.

Another important issue is related to the use of a cartoon character to test our hypotheses. This would be seen as a hitch for claiming for the ecological validity of our results and, as such, that they could only serve to test our synthesis procedure and not the way that humans process audiovisual prosody. In regard to our results and the ones from [15], we claim that the use of a cartoonist model could be preferred over more realistic human-like versions since very realistic graphical representations might become a problem, as stated in the Uncanny Valley hypothesis ([16]). As well, in order to claim for the ecological validity of our results, works like [17] have suggested that behavior may be more important than the visual realism of the character in order to obtain a realistic response.

5. Acknowledgements

We are particularly indebted to Javier Agenjo for his help and assistance with the NINOs platform. We also want to thank M. Swerts, J. Blat, E. Arroyo and the reviewers for their insightful comments. This research was supported by grants FFI2009-07648/FILO and CONSOLIDER-INGENIO 2010 Programa CSD2007-00012, awarded by the Spanish Ministry of Science and Innovation, and by project 2009 SGR 701, awarded by the Generalitat of Catalonia.

6. References

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

[2] Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating Uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. *Proceedings of the Third International Conference on Speech Prosody* (025: 1-4). Dresden.

[3] Mehrabian, A., & Ferris, S. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, **31**, pp. 248-252.

[4] Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. *Language and Speech*, **46**(1), pp. 1-22.

[5] House, D. (2002). Intonation and visual cues in the perception of interrogative mode in Swedish. *Proceedings of ICSLP* 2002, 1957-1960.

[6] Borràs-Comes, J., & Prieto, P. (2010). 'Seeing tunes.' The role of visual gestures in tune interpretation. Oral presentation at the *12th Conference on Laboratory Phonology*. Albuquerque (NM).

[7] Massaro, D. W. (2001). Speech Perception. In N. M. Smelser & P. B. Baltes (Eds.) & W. Kintsch (Section Ed.): *International Encyclopedia of Social and Behavioral Sciences* (pp. 14870-14875). Amsterdam: Elsevier.

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

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

[10] Boersma, P., & Weenink, D. (2008). *Praat: doing phonetics by computer* (version 5.0.09). Computer Program.

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

[12] IBM Corporation (2010). *IBM SPSS Statistics* (version 19.0.0). Computer Program.

[13] Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, **86**(3), pp. 236-242.

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

[15] Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2011). Crossmodal prosodic and gestural contribution to the perception of contrastive focus. Presented at the *12th Annual Conference of the International Speech Communication Association*. Florence, August 27-31.

[16] Mori, M. (1970). Bukimi no tani. [The uncanny valley. Translated by K. F. MacDorman & T. Minato]. *Energy*, 7(4), 33-35.

[17] Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., & Slater, M. (2006) Building expression into virtual characters. Presented at *Eurographics 2006 STAR - State of the Art Report* (pp. 21-61). Eurographics Association: Switzerland.