Perceiving uncertainty: facial gestures, intonation, and lexical choice

Joan Borràs-Comes,[1] *Paolo Roseano,*[2,1] *Maria del Mar Vanrell,*[1,3] *Aoju Chen,*[4] *and Pilar Prieto*[5,1]

[1] Universitat Pompeu Fabra, Barcelona, Spain
[2] Universitat de Barcelona, Barcelona, Spain
[3] Universitat Autònoma de Barcelona, Barcelona, Spain
[4] Universiteit Utrecht, Utrecht, Netherlands
[5] ICREA, Barcelona, Spain

{joan.borras, mariadelmar.vanrell, pilar.prieto}@upf.edu, paolo.roseano@ub.edu, aoju.chen@uu.nl

Abstract

Languages rely on many verbal and nonverbal sources for the expression of uncertainty, and these linguistic markers are used by hearers to detect degrees of uncertainty in natural communication. An important question is whether the perception of uncertainty is better chracterized by lexical marking, prosody or facial gestures. To test this, a group of Catalan speakers were presented with two perception experiments containing a set of audiovisual materials in which lexical, prosodic and facial gestural cues presented congruent and incongruent combinations for the expression of uncertainty. The results from the two experiments demonstrate that, even though lexical choice is important for conveying pragmatic meanings like uncertainty, it can be easily overridden by prosodic and gestural patterns. Moreover, when gesture and prosody are in conflict, gesture is a more salient and powerful cue.

Index Terms: audiovisual prosody, uncertainty, modal adverbs, intonation, facial gestures.

1. Introduction

As is well-known, languages rely on many sources for the expression of uncertainty, namely lexical marking (by the choice of modal adverbs such as *probably*, *perhaps*), morphological marking (in languages with morphemic marking of epistemicity), prosody, as well as facial gestures (see, e.g., [1]). These linguistic markers are used by hearers to detect degrees of uncertainty in online communication. An important question is whether the perception of uncertainty is better encoded at the lexical or morphological level as opposed to the intonational or gestural level.

In a classic study, Mehrabian and Ferris [2] analyzed how listeners got information about a speaker's general attitude in situations where the facial expression, tone of voice and/or words were sending conflicting signals. 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 suggesting that the combined effect of simultaneous lexical, vocal and facial attitude communications is a weighted sum of their independent effects, with the coefficients of .07, .38 and .55, respectively.

Dijkstra, Krahmer & Swerts [3] tested which cues participants used to assess the degree of certainty of a person answering factual questions. The auditory-visual (AV) materials presented 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. Similarly, Swerts & Krahmer [1] showed that there are clear visual cues for a speaker's feeling of knowing and that listeners are more capable of estimating another person's knowing on the basis of combined auditory and visual information than on the basis of auditory information alone. Rilliard et al. [4] explored the audiovisual perception of attitudinal expressions in Japanese and French and found that the two modalities were widely used by all speakers and decoded adequately by listeners. Moreover, while the expression of obviousness seemed to rely heavily on the speaker's performance, neutral statements and dubitative expressions behaved similarly such that both languages had a systematic opposition between an assertive and a dubitative expression. Even though earlier literature shows that visual cues for uncertainty can override auditory cues and even lexical markers, little is known about the interactions between these three types of information.

Several studies have shown that our perceptual system integrates auditory speech information and visual cues from the speaker's face. In a classic study, McGurk & MacDonald [5] 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.

Borràs-Comes & Prieto [6] explored the relative importance of pitch accent contrasts and facial gestures in the distinction between contrastive focus statements and counter-expectational questions 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 goal of our experiments is to test whether the lexical marking of uncertainty is able to override the prosodic and gestural cues, and vice-versa. The novelty of this paper is that

it introduces a three way opposition between three types of lexical markers (*perhaps*, *probably*, and *obviously*) which express three different degrees of certainty, which interact with congruent and incongruent intonation and gestural patterns. To test this, we ran a set of decision tasks with 30 speakers of Catalan, in which they had to rate the stimuli on a 1-7 scale of certainty.

The audiovisual recordings obtained to assess the audiovisual cues for uncertainty in Catalan are presented in section 2. Experiments 1 and 2 are presented in sections 3 and 4, respectively. Finally, we discuss the results and their main implications in section 5.

2. Audiovisual recordings

The goal of our research is to test whether the lexical marking of uncertainty can override or not prosodic and gestural cues. In order to assess which intonational and gestural cues are used for Catalan speakers in order to mark uncertainty, four Catalan native speakers (henceforth *actors*) were recorded while answering to two *wh*- questions:

1.	*Qui vindrà al sopar?*	*La Marina.*
	'Who will come to the dinner?'	'Marina.'
2.	*On és el sopar?*	*A "La Gavina".*
	'Where is the dinner?'	'At "The Seagull".'

The four actors were asked to pronounce four different types of answers: an isolated condition (e.g, *La Marina*) and three adverbial conditions using the following modal adverbs: *potser* 'perhaps', *segurament* 'probably', and *òbviament* 'obviously' (e.g. *Potser la Marina*, *Segurament la Marina*, and *Òbviament la Marina*). These 8 different utterances (2 item × 4 adverb) were pronounced both in a confident way and in an uncertain way, with no explicit instructions on how to express certainty and uncertainty audiovisual prosody.

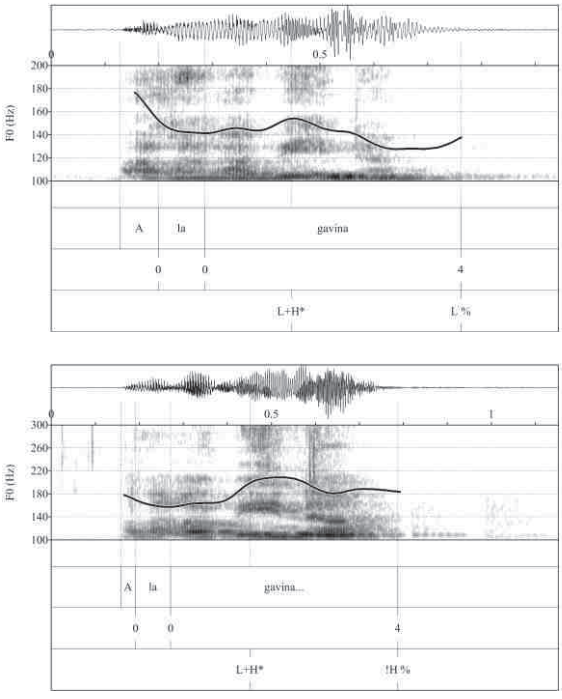

Figure 1. *Representative intonational patterns of* A la Gavina*, with certainty (top panel) and uncertainty (bottom panel) interpretations.*

Concerning the intonational properties, our production results were consistent with [7] and showed that uncertainty statements tend to present mid-level boundary tones, slower speech rates and optionally longer final syllables compared to confident ones. Figure 1 shows two representative intonational patterns of *A la Gavina*, with a certainty and an uncertainty interpretation.

The Elan 4.0.1 software program ([8]) was used to label gestures produced by means of five gestural articulators (head, eyebrows, eyes, mouth, shoulders). Gestures were labeled according to Facial Action Coding System (FACS; [9]), which allows coding visually distinguishable facial expressions and other types of gestures from the upper part of the body. FACS groups muscle 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. The analysis of the gestural cues reveals that the most stable gestural cue in the conveyance of certainty is an "affirmation" head nod (M59), whereas all other articulators (eyes, eyebrows, mouth and shoulders) do not seem to play any clear role (eyes wide open appear to be a secondary gestural correlate of certainty). On the other hand, the gestural expression of uncertainty is reflected in most articulators, with the only exception of eyebrows. When expressing uncertainty, subjects shrug their shoulders (AU82), squint their eyes (AU6), stretch their lips (AU20) and make single head movements (chin-out, side turn, side tilt) which are different from the head movements related to certainty. Figure 2 shows four representative gestural patterns of *La Marina*, with a certainty (left panels) and an uncertainty (right panels) interpretation.

Figure 2. *Representative gestural patterns of* La Marina*, with a certainty (left panels) and an uncertainty interpretation (right panels).*

The selected stimuli were rated on a 1-7 scale of certainty (1 = totally uncertain; 7 = totally certain) by 4 separate raters, which coincided in the rating of the prosodic and gestural stimuli as being adequate for the expected category.

We thus obtained a corpus of 64 utterances (4 actors × 2 items × 4 adverbs × 2 prosodic/gestural conditions). This set of audiovisual recordings was the basis of the audiovisual stimuli used in our two perceptual experiments.

3. Experiment 1: verbal vs. nonverbal cues

3.1. Methodology

Experiment 1 consisted of three tasks. In these three tasks, the abovementioned 64 stimuli were presented in an auditory-only condition (AO task), in a visual-only condition (VO task) and in an audiovisual condition (AV task; always presenting congruent intonational and gestural combinations to the participants). The order of the AO and VO tasks was counterbalanced among participants.

A total of 30 Catalan native speakers participated, for a total of 5,760 responses (64 stimuli × 3 tasks × 30 subjects). All subjects were undergraduates studying journalism or translation at 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 materials and rate them on a 1-7 scale of certainty (1 = totally uncertain; 7 = totally certain).

The experiment was set up by means of E-prime 2.0 ([10]), and response frequencies were automatically recorded. Subjects were instructed to press the button 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.

All responses were analyzed using a Linear Mixed Model (LMM) analysis through IBM SPSS Statistics 19.0 ([11]). LMM can control for both fixed and random factors. In our analyses, subject, actor and item were set as crossed random effects. First, we present the results concerning the isolated answers and second, the results of the three adverbial conditions.

3.2. Results

3.2.1. Isolated condition

Figure 3 shows the mean 'certainty' rates (y-axis) for the isolated answers, as a function of gestural/prosodic information (x-axis), in each task (different panels). The graph shows clear distinctions between VO and AV, but less clearer decisions for uncertainty-classified AO stimuli, which suggests that auditory information alone could not provide unmistakable cues for uncertainty marking.

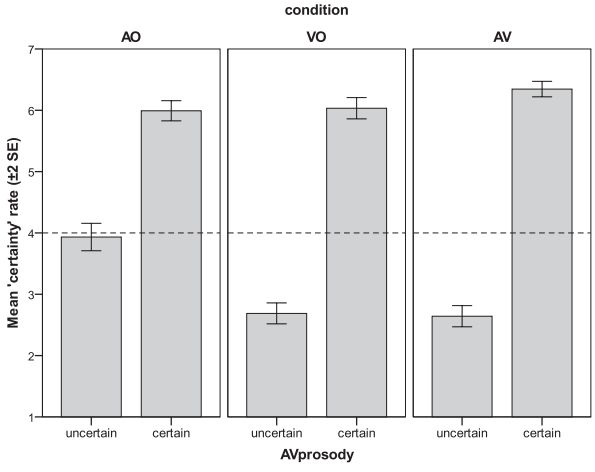

Figure 3. *Mean 'certainty' rates (y-axis) for the isolated answers, as a function of gestural/prosodic information (x-axis), in each task (different panels).*

A LMM analysis was set up with the certainty rate as the dependent variable, AVPROSODY, CONDITION, and their interaction as fixed factors and subject, actor and item as crossed random factors. Main effects of AVPROSODY ($F_{1, 1195}$ = 2146.953, $p < .001$) and CONDITION ($F_{2, 1195}$ = 31.055, $p < .001$) were found, and the interaction between the two was also significant ($F_{2, 1195}$ = 58.160, $p < .001$). We then ran a pairwise comparison between the three different conditions, with a Bonferroni adjustment for multiple comparisons. AO was different from both VO and AV (always $p < .001$), but there were no differences between VO and AV ($p = .291$).

3.2.2. Adverbial conditions

Figure 4 shows the mean 'certainty' rates (y-axis) for the adverbial answers, as a function of gestural/prosodic information (x-axis) and the adverb used (different bars), in each of the three conditions (different panels). The graph shows a clear distinction based on prosodic/gestural information, especially for VO and AV. In AO, the uncertainty-classified utterances were at chance level when *probably* and *obviously* adverbs were used.

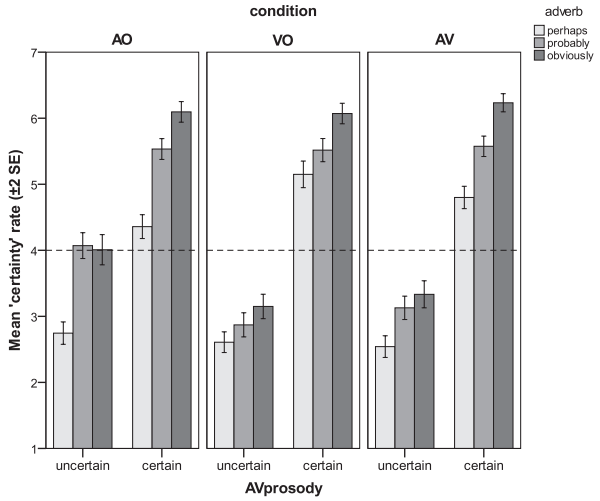

Figure 4. *Mean 'certainty' rates (y-axis) for the adverbial answers, as a function of gestural/prosodic information (x-axis) and adverb used (different bars), in each of the three conditions (different panels).*

A LMM analysis was set up with the certainty rate as the dependent variable, AVPROSODY, CONDITION, ADVERB and all their possible interactions as fixed factors and subject, actor and item as crossed random factors. Main effects of AVPROSODY ($F_{1, 4063}$ = 3770.589, $p < .001$), CONDITION ($F_{2, 4063}$ = 15.536, $p < .001$), and ADVERB ($F_{2, 4063}$ = 301.634, $p < .001$) were found. All their paired interactions were also significant: AVPROSODY × CONDITION ($F_{2, 4063}$ = 64.437, $p < .001$), AVPROSODY × ADVERB ($F_{2, 4063}$ = 17.701, $p < .001$), and CONDITION × ADVERB ($F_{4, 4063}$ = 19.731, $p < .001$). The triple interaction AVPROSODY × CONDITION × ADVERB was not significant ($F_{4, 4063}$ = 1,028, $p = .391$). A pairwise comparison between the three different conditions revealed that AO was different from both VO and AV (always $p < .001$), but there were no differences between VO and AV again ($p = 1$). Another pairwise comparison between the three adverbs revealed significant differences for every comparison (always $p < .001$).

4. Experiment 2: gesture vs. intonation vs. lexical choice

4.1. Methodology

For Experiment 2, the set of 64 stimuli from the previous VO task was presented together with congruent and incongruent auditory information (i.e., gestural information conveying uncertainty was presented with its original audio channel but also with the audio channel of the corresponding "confident" recording, and vice-versa), for a total of 128 AV stimuli.

With respect to AV binding, we placed the new acoustic information where the original had one been, taking the stressed syllable of the narrow focused word in our sentences as an anchor point. By doing this, we kept all the suprasegmental properties of our acoustic stimuli. Despite some cases of temporary asynchrony, two separate raters reported consciousness-related binding and thus an acceptable situation of AV synchrony to allow our participants to integrate auditory and visual information (see, e.g., [12] on audiovisual binding and the ventriloquist effect).

The same 30 Catalan native speakers participated, for a total of 3,840 responses (128 stimuli × 30 subjects). The experimental procedure was the same as in Experiment 1: to pay attention to the audiovisual materials and rate them in a 1-7 scale of certainty. Again, we first present the results concerning the isolated answers, and then the results of the three adverbial conditions.

4.2. Results

4.2.1. Isolated condition

Figure 5 shows the mean 'certainty' rates (y-axis) for the isolated answers, as a function of gestural/prosodic information (x-axis). The graph shows that when gesture provides certainty information and intonation provides uncertainty information, participants clearly rely on gestural information.

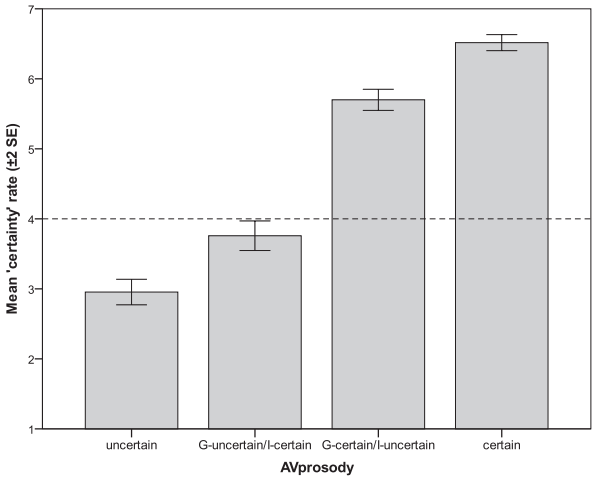

Figure 5. *Mean 'certainty' rates (y-axis) for the isolated answers, as a function of gestural/prosodic information (x-axis): 1 (uncertain AV prosody), 2 (uncertain gesture + certain intonation), 3 (certain gesture + uncertain intonation), and 4 (certain AV prosody).*

A LMM analysis was set up with the certainty rate as the dependent variable, GESTURE, INTONATION, and their interaction as fixed factors and subject, actor and item as crossed random factors. Main effects of GESTURE ($F_{1, 717} = 1321.346$, $p < .001$) and INTONATION ($F_{1, 717} = 114.580$, $p < .001$) were found, but the GESTURE × INTONATION interaction was not significant ($F_{1, 717} = 0.007$, $p = .934$).

4.2.2. Adverbial conditions

Figure 6 shows the mean 'certainty' rates (y-axis) for the adverbial answers, as a function of gestural/prosodic information (x-axis) and the adverb used (different bars). As in Experiment 1, the graph shows an important effect of nonverbal information in the participants' ratings, but also an effect of adverb.

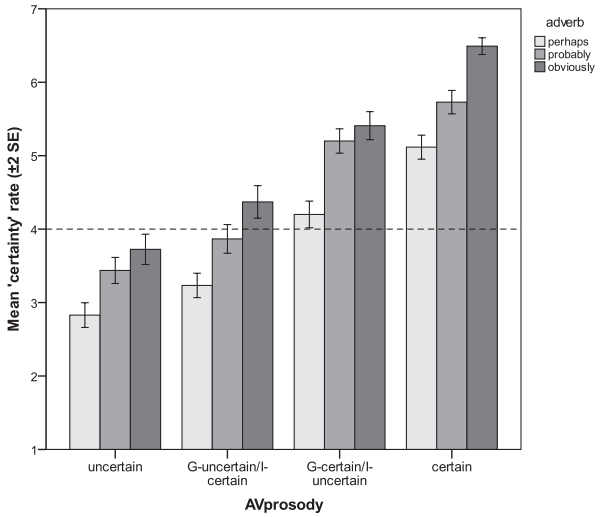

Figure 6. *Mean 'certainty' rates (y-axis) for the adverbial answers, as a function of gestural/prosodic information (x-axis) and adverb used (different bars).*

A LMM analysis was set up with the certainty rate as the dependent variable, GESTURE, INTONATION, ADVERB and all their possible interactions as fixed factors, and subject, actor and item as crossed random factors. Main effects of GESTURE ($F_{1, 2629} = 1676.672$, $p < .001$), INTONATION ($F_{1, 2629} = 236.027$, $p < .001$), and ADVERB ($F_{2, 2629} = 239.206$, $p < .001$) were found. All their paired interactions were also significant: GESTURE × INTONATION ($F_{1, 2629} = 16.196$, $p < .001$), GESTURE × ADVERB ($F_{2, 2629} = 3.468$, $p = .031$), and INTONATION × ADVERB ($F_{2, 2629} = 6.554$, $p = .001$). The triple interaction GESTURE × INTONATION × ADVERB was not significant ($F_{2, 2629} = 2.128$, $p = .119$). A pairwise comparison between the three different adverbs revealed significant differences for every comparison (always $p < .001$).

5. Discussion and conclusions

Experiment 1 showed an important effect of audiovisual prosody, i.e., nonverbal information, both in the isolated ($F = 2147$) and in the adverbial condition ($F = 3771$). This effect was found to be especially stronger for the VO and AV conditions, with no differences between these two, which can be taken to mean that gestures convey uncertainty more clearly than intonation.

Moreover, it has been shown that sentences preceded by *probably* and *obviously* produced with uncertain prosody — as well as sentences preceded by *perhaps* produced with certain prosody — were perceived close to chance level by our participants, which suggest an incongruity between these lexical

items and the prosodic information for the expression of uncertainty.

Experiment 2 again showed an important effect of gesture, both in the isolated ($F = 1321$) and in the adverbial condition ($F = 1676$). Moreover, once gesture, intonation and adverb had been analyzed separately, we observe very similar effect sizes for intonation ($F = 236$) and adverb ($F = 239$).

All in all, we have found that while all three cues (lexical choice, intonation, and gestural patterns) have a significant effect on certainty perception, it is mainly the visual (gestural) cues that have overridden the auditory cues.

This phenomenon is in line with the findings of [1], [2] and [3]. The majority of 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, feeling of knowing or of uncertainty, etc.) have found a very strong effect of visual cues. Overall, our results support that both facial gestures and prosody can be considered as two strongly interrelated systems when conveying linguistic meanings such as epistemicity,

6. Acknowledgements

We thank Francesco Cangemi, Marc Swerts, Francisco Torreira and two anonymous referees for their comments on an earlier version of this paper. We also thank the collaboration of Núria Esteve and Gemma Barberà in the recordings. This work has been presented at Phonetics and Phonology in Iberia 2011 (Universitat Rovira i Virgili, Tarragona). This research has been supported by grants FFI2009-07648/FILO and CONSOLIDER-INGENIO 2010 (CSD2007-00012), awarded by the Spanish Ministry of Science and Innovation, and by project 2009 SGR 701, awarded by the Generalitat of Catalonia.

7. References

[1] Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53, pp. 81–94.

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

[3] Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In R. Hoffmann and H. Mixdorff (eds.): *Proceedings of the Third International Conference on Speech Prosody* (025, pp. 1–4). Dresden.

[4] Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., & Aubergé, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. *Language and Speech*, 52(2-3), pp. 223–243.

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

[6] Borràs-Comes, J., & Prieto, P. (in press). 'Seeing tunes.' The role of visual gestures in tune interpretation. *Laboratory Phonology*, 2(2).

[7] Vanrell, M. M., Borràs-Comes, J., Roseano, P., & Prieto, P. (2011). Prosodic cues of confidence and uncertainty in Catalan. Poster presentation at *Phonetics and Phonology in Iberia 2011*, Tarragona, June 21-22.

[8] Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the Neuroges-Elan system. *Behavior Research Methods, Instruments, & Computers*, 41(3), pp. 841–849.

[9] Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

[10] Psychology Software Tools Inc. (2009). *E-Prime* (version 2.0). Computer Program: <http://www.pst-net.com/>

[11] IBM Corporation (2010). *IBM SPSS Statistics* (version 19.0.0). Computer Program: <http://www.spss.com/software/statistics/>

[12] Bischoff, M., Walter, B., Blecker, C. R., Morgen, K., Vaitl, D., & Sammer, G. (2007). Utilizing the ventriloquism-effect to investigate audio-visual binding. *Neuropsychologia*, 45, pp. 578–586.