Manipulating Uncertainty

The contribution of different audiovisual prosodic cues to the perception of confidence

Christel Dijkstra, Emiel Krahmer, Marc Swerts

Communication and Cognition
Tilburg University, The Netherlands

Abstract

When answering factual questions, speakers can signal whether they are uncertain about the correctness of their answer using prosodic cues such as fillers ("uh"), a rising intonation contour or a marked facial expression. It has been shown that on the basis of such cues, observers can make adequate estimates about the speaker's level of confidence, but it is unclear which of these cues have the largest impact on perception. To find the relative strength of the three aforementioned cues, a novel perception experiment was performed in which answers were artificially manipulated in such a way that all possible combinations of the cues of interest could be judged by participants. Results showed that while all three factors had a significant influence on the perception results, this effect was by far the largest for facial expressions.

1. Introduction

The idea that non-verbal communication forms a very substantial part of communication is a popular one. We regularly encounter statements, typically without a source or citation, saying that non-verbal communication accounts for more than 90% (or some comparable figure) of a message. Presumably, these statements can be traced back to the work of Mehrabian and colleagues in the second half of the sixties (e.g., Mehrabian and Wiener 1967, Mehrabian and Ferris 1967). They studied how people judged a speaker's general attitude, which could be positive, negative or neutral, based on possibly conflicting (i.e., incongruent) verbal, intonative and facial cues. In the aforementioned studies it was found that the relative weights of these three factors were .7, .38 and .55 respectively (hence .93 non-verbal). Even though the applicability of this result has been stretched beyond recognition, the notion that non-verbal cues such as intonation and facial expressions are important for communication is in itself uncontroversial. However, the relative importance of different auditory and visual cues is far from transparent, and the situation is further complicated by the observation that this seems to depend on which aspects of communication are studied (e.g., Swerts and Krahmer 2005). For instance, even though visual as well as auditory cues can be shown to influence both emotion and speech perception, visual cues are generally believed to have a larger impact for the former and auditory cues for the latter.

In this paper we report on an experiment which studies the importance of different audiovisual cues for the perception of speaker uncertainty. When speakers are asked factual questions (e.g., "Who wrote Faust?") and are not certain about the correctness of their answer, they can signal this uncertainty in a variety of ways. Such uncertain answers may be preceded by fillers such as "uh" or "uhm", uttered with a rising, question-like intonation contour, and pronounced with a marked facial expression (Smith and Clark 1993, Swerts and Krahmer 2005). Moreover, observers can make adequate estimates of speakers' confidence in the correctness of their answers (Brennan and Williams 1997, Krahmer and Swerts 2005). It turns out that observers can make somewhat better estimates of a speaker's level of confidence when they have access to both visual and auditory cues than when they are only offered one of these modalities in isolation (Swerts and Krahmer 2005).

Table 1: *Frequent certain (high FOK) and uncertain (low FOK) settings for the three cues of interest: filler, intonation contour and facial expressions.*

Cue	Certain (+)	Uncertain (−)
Filler	Absent	Present
Intonation	Falling	Rising
Facial expression	Neutral	Marked

While this clearly indicates that auditory and visual non-verbal cues are important for uncertainty perception, it does not tell us which of the audiovisual correlates of uncertainty have the largest influence on perception. To find out, we performed a perception test in which we systematically manipulated features of answers, to obtain all the possible combinations of the auditory and visual cues of interest. The use of incongruent stimuli is a common technique to factor out the relative contribution of various factors, and has been applied successfully in, for instance, McGurk and MacDonald (1976), Massaro et al. (1996) and de Gelder and Vroomen (2000).

The rest of this paper is organized as follows. In section 2 the experimental method is outlined. Since it was important that participants would not note that the stimuli were manipulated, we opted for a between participants design, where each participant saw a very limited set of stimuli. Section 3 describes the findings, and the paper ends with a combined discussion and conclusion in section 4.

2. Method

2.1. Participants

Participants were 120 native speakers of Dutch (52 male and 68 female), between 18 and 65 years old.

2.2. Stimuli

Stimuli were selected from a set of 800 utterances, that were collected during an earlier experiment. These utterances consisted of the responses from 20 adult native speakers of Dutch to

Figure 1: Examples of marked facial expressions, naturally produced during low FOK (uncertain) answers.

40 factual questions selected from a Dutch intelligence test (the *Wechsler Adult Intelligence Scale, WAIS*) and from the Dutch version of *Trivial Pursuit*. While answering these questions, the speakers were recorded using a digital camera, filming the speakers' heads from the front. Following Hart (1965), the first question round was followed by a questionnaire in which speakers had to indicate on a 7-point Likert scale for each of the 40 questions how certain they were that they would be able to recognise the correct answer in a multiple choice test (1 = "definitely not recognize", 7 = "definitely recognize"). These scores are referred to as the Feeling of Knowing (FOK) scores. In general, a high FOK score corresponds with speaker certainty, while a low FOK score corresponds to speaker uncertainty.

For all utterances, the presence or absence of a number of auditory and visual features was manually annotated by 4 independent labellers. It was found that high FOK answers (i.e., answers about which the speaker is certain) tend to be associated with little or no filled pauses, a falling intonation contour, and a neutral facial expression, while low FOK (uncertain) answers were frequently preceded by filled pauses, were more often associated with a rising intonation, and were more often uttered with non-neutral facial expressions (see Table 1 for a schematic representation). A special instance of such non-neutral facial expressions were dubbed "funny faces," for want of a better name. These funny faces appear to be similar to the "thinking faces" of Goodwin and Goodwin (1986); in terms of Ekman and Friesen's (1978) *Facial Action Coding Systems*, they typically consist of a combination of action units (AUs) such as lip corner depression (AU 15), lip stretching (AU 20) or lip pressing (AU 24), possibly combined with eye widening (AU 5) and eyebrow movement (AU 1, AU 2), as can be witnessed in Figure 1. For further details on the experimental procedure and annotation of the recordings, the interested reader is referred to Krahmer and Swerts (2005) and Swerts and Krahmer (2005).

For the current experiment we selected one certain (high FOK) and one uncertain (low FOK) answer from 5 speakers. The selected answers had to meet the following constraints:

- They should have substantially different FOK scores and be lexically similar. For example, "Goofy" (low FOK) in response to the question "What is the name of the cartoon character who owns the dog Pluto?" and "Goethe" (high FOK) in response to "Who wrote Faust?".
- Moreover, the high FOK answer should be associated with a facial expression indicating certainty, no filler and a falling intonation, and the low FOK answer with a facial expression indicating uncertainty, a filler and a rising intonation.

Table 2: *Schematic overview of the eight stimuli for each speaker, consisting of all possible combinations of filler, intonation and facial expression; a + indicates the certain variant, a − the uncertain variant.*

Stimulus ID	Filler	Intonation	Facial expr.
1.	+	+	+
2.	−	−	−
3.	+	−	+
4.	−	+	−
5.	+	−	−
6.	−	−	+
7.	+	+	−
8.	−	+	+

Table 3: *Experimental design: distribution of speakers and stimuli over the eight experimental versions A to H. Each experimental version thus contained five different stimuli, one of each speaker.*

	Stimulus ID							
Speaker	1	2	3	4	5	6	7	8
IS	A	B	C	D	E	F	G	H
PM	H	A	B	C	D	E	F	G
KS	G	H	A	B	C	D	E	F
PMO	F	G	H	A	B	C	D	E
ED	E	F	G	H	A	B	C	D

The selected pairs of answers were manipulated with Adobe Premiere™ to obtain all the stimulus variants. These variants consist of all possible combinations of certain and uncertain settings of the three features of interest (filler, intonation and facial expression), which gives rise to the eight variants depicted in Table 2. For some speakers, we could not obtain stimuli pairs meeting all the requirements above. In those cases, intonation contour and facial expression where kept as selection criteria, and an auditory filler (selected from a different answer of the same speaker) was inserted at the appropriate place. This typically was an "uhm" since these are uttered with a closed mouth and hence lend themselves better for combination with film fragments (no lip sync problems). Stimuli 1 and 2 in Table 2 served as the basis for all other manipulations. For instance, by mixing the spoken answer from stimulus 1 (certain) with the visual part of stimulus 2 (uncertain), mixed stimulus 7 was obtained (uncertain facial expression, no filler, falling intonation). Special care was given to the alignment of auditory and visual speech to avoid unwanted McGurk effects (McGurk and MacDonald 1976). If this turned out to be problematic for a given pair of answers, that pair was discarded in favour of a different one. The final 40 stimuli (8 variants × 5 speakers) were pretested on naturalness to see whether any manipulation artefacts could be detected, which turned out to be not the case.

Table 4: *Average FOAK scores (ranked from low to high FOAK, with standard deviations between brackets) for the 8 stimulus types representing all possible combinations of certain (+) and uncertain (−) variants of filler, intonation contour and facial expressions.*

ID	N	Filler	Intonation	Facial expr.	FOAK (s.d.)
5.	75	+	−	−	1.83 (1.12)
2.	75	−	−	−	1.85 (1.12)
7.	75	+	+	−	1.93 (1.07)
4.	75	−	+	−	2.17 (1.42)
3.	75	+	−	+	3.73 (1.40)
6.	75	−	−	+	4.40 (1.40)
1.	75	+	+	+	4.51 (1.44)
8.	75	−	+	+	4.77 (1.52)

2.3. Experimental design

The experiment had a between participants design (a within participants design would not have been feasible, as participants would see the same video images of the same speaker reappear a number of times, which would obviously reveal the manipulations). It was decided to present only one stimulus from each speaker to participants, which led to 8 experimental versions. Strictly speaking this is a mixed within-between design, but we treat it as having a complete between participants-design which is warranted since it does not affect the central comparisons. Within each experimental version, five out of the eight stimulus-versions in Table 2, one for each different speaker, were presented, according to the scheme depicted in Table 3. Thus, participants in experimental version A, watched stimulus 1 of speaker IS (all three cues certain), stimulus 2 of speaker PM (all three cues uncertain), stimulus 3 of speaker KS, etc. The order of presentation within an experimental version was random to avoid potential learning effects.

2.4. Procedure

The 120 participants were randomly assigned to one of the 8 experimental versions. They were told that they would see 5 film clips of answers to questions that were themselves not shown, and that for each answer they had to indicate on a 7 point Likert scale how certain the speaker appeared about the correctness of the given answer (1 = "speaker is very certain"; 7 = "speaker is very uncertain"). Below, these scores will be referred to as the Feeling of Another's Knowing (FOAK) scores, following Jameson et al. (1993) and Brennan and Williams (1997). The experiment lasted approximately 2 minutes, including instructions and debriefing. None of the participants indicated they had noticed that the stimuli had been artificially manipulated, when they were asked about this after completing the experiment.

2.5. Data processing

The analysis of variance (ANOVA) method was used to test for statistical significance, and post hoc tests were performed using the Tukey HSD method.

3. Results

The factor speaker did not have a significant effect on the FOAK scores, hence below we collapse results across speakers.

Table 4 shows the average FOAK scores for the 8 stimulus types for all 5 speakers, ranked here from low to high FOAK.

Table 5: *Main effects of the marked and unmarked settings of filler, intonation contour and facial expression on average FOAK scores (with standard deviations between brackets), with F-statistics.*

Factor	Level	FOAK (s.d.)	F-statistics
Filler	Present	3.30 (1.89)	$F(1, 592) = 7.74$,
	Absent	3.00 (1.71)	$p < .01, \eta_p^2 = .013$
Intonation	Rising	2.95 (1.70)	$F(1, 592) = 13.31$,
	Falling	3.35 (1.89)	$p < .001, \eta_p^2 = .022$
Facial expr.	Marked	1.95 (1.19)	$F(1, 592) = 498.33$,
	Neutral	4.35 (1.48)	$p < .001, \eta_p^2 = .457$

The stimulus type had a significant influence on the FOAK scores ($F(7, 592) = 75.25, p < .001$). The average FOAK scores range from $M = 1.83$ for stimulus 5 (no filler, rising intonation, marked facial expression) which was perceived as most uncertain, to $M = 4.77$ for stimulus 8 (filler, falling intonation, neutral facial expression) which appeared most certain. It is worth observing that even the answers perceived as most "certain", receive moderate FOAK scores. The Tukey HSD analysis revealed that three homogeneous groups can be distinguished among the stimulus IDs. One group consists of the stimulus types with uncertain facial expressions (IDs 5, 2, 7, and 4), these are associated with consistently low FOAK scores (hence, they are generally perceived as uncertain). The other groups consist of ID 3 and IDs 6, 1 and 8 respectively, both containing certain facial expressions.

This already suggests that facial expressions are the strongest cue of the three under investigation, which is further confirmed by looking at the main effects associated with each of these cues. Table 5 reveals that statistically significant main effects are found for all three cues. The smallest effect is associated with filled pauses; on average, stimuli with a filled pause are perceived as slightly more certain ($M = 3.30$) than stimuli without a filler ($M = 3.00$), contrary to expectation. Even though the difference is statistically significant, the effect size ($\eta_p^2 = .013$) is very small. Question intonation also has a relatively small effect ($\eta_p^2 = .022$), but this time in the expected direction: on average, stimuli that contain a rising intonation are perceived as less certain ($M = 2.95$) as opposed to stimuli that contain a falling intonation ($M = 3.35$). Finally, a substantial effect (an η_p^2 of .457) is found for facial expressions, in that stimuli with a marked facial expression ("funny face") are perceived as much more uncertain (i.e., having a lower average FOAK score, $M = 1.95$) than stimuli with a neutral facial expression ($M = 4.35$). No significant interactions were found.

4. Summary and Conclusion

When asked a factual question, speakers may be able or unable to provide an answer (even though it may feel as if the answer lies on the "tip of the tongue", in the phrase introduced by James 1890). When speakers do provide an answer, they may be relatively certain or uncertain about the correctness of this answer. Various researchers have shown that speakers employ and observers are sensitive to various non-verbal cues which indicate uncertainty, such as usage of fillers (e.g., "uh", "uhm"), rising intonation and marked facial expressions (e.g., thinking or funny faces). It seems likely that speakers use such cues as a face-saving device, should the answer turn out to be in-

correct after all (Smith and Clark 1993, Brennan and Williams 1997, Swerts and Krahmer 2005). What these and other studies do not reveal is what the relative contributions of these various cues for the perception of uncertainty are, and this question was addressed in the current paper.

To answer this question, a Feeling of Another's Knowing experiment was conducted with manipulated stimuli, derived from on an earlier collected audiovisual corpus of 800 responses to factual questions from 20 adult Dutch speakers. From this corpus, one high FOK and one low FOK answer were selected for 5 speakers, where the former contained cues associated with certainty (no filler, falling intonation, neutral facial expression) and the latter cues associated with uncertainty (filler, rising intonation, uncertain facial expression). By systematic manipulation, all combinations of the cues of interest could be created, resulting in 8 stimuli per speaker. Participants, who were not aware of the manipulations, were asked to rate how certain they felt the speaker was about the answer.

It was found that all three cues had a significant effect on certainty perception. The smallest effect (in terms of effect size values) could be attributed to fillers. Contrary to our expectation, the presence of a filled pause was associated with a very small though significant increase in certainty perception. We are unsure about the precise nature of this effect. It is worth pointing out that most fillers used in this experiment were of the "uhm" variety since these were easier to manipulate (mouth remains closed). It would be interesting to perform a more detailed study of the effect of fillers on certainty perception, also in view of the not entirely consistent results of Smith and Clark (1993) and Brennan and Williams (1995) in this respect (the former but not the latter finding a functional difference between fillers). Intonation had a marginally stronger effect, this time in the expected direction: when an answer was pronounced with a rising intonation, it was indeed perceived as somewhat less certain. The strongest effect by far could be attributed to the facial expression: when the speaker answered while producing a marked facial expression, this was indeed perceived as much more uncertain. Interestingly, "funny faces" occurred less frequently among the 800 responses than fillers and rising intonation contours, but when they do occur they are a very strong cue for uncertainty perception. It is surprising that no significant interactions were found, which indicates that the three cues are essentially independent for this task. It is also worth pointing out that the average FOAK scores are relatively low overall; even the stimuli that appeared most certain, were rated below 5 on average on the 7-point scale. Presumably this can be explained from the observation that certainty is signalled using normal (more neutral) cues, while uncertainty is signalled using the marked cue settings.

The fact that the visual cues overruled the auditory cues for certainty perception is in line with the findings of emotion perception. In fact, it does not seem to be too far-fetched to argue that uncertainty is a "social emotion" (see e.g., Adolphs 2002), comparable to other social emotions such as for instance embarrassment. When participants have to judge the emotion of incongruent stimuli (e.g., a happy face with a sad voice), the visual cues have also been shown to be have a stronger influence on perception (e.g., in the study of Mehrabian and Ferris 1967 mentioned in the introduction, but also in more recent work such as Hess et al. 1988 or Massaro and Egan 1996, among many others).

Acknowledgements This paper is based on a master thesis written by the first author, under supervision of the other two authors. The research was conducted as part of the VIDI-project "Functions of Audiovisual Prosody (FOAP)", sponsored by the Netherlands Organisation for Scientific Research (NWO), see `foap.uvt.nl`. Many thanks to Lennard van de Laar for technical assistance.

5. References

[1] Adolphs, R. (2002), Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Review* **1**(1):21–61.

[2] Brennan, S.E. and Williams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to observers about the metacognitive states of speakers. *Journal of Memory and Language*, **34**, 383–398.

[3] Ekman, P. and Friesen, W.V. (1978). *The facial acting coding system*. Palo Alto: Consulting Psychologists' Press.

[4] de Gelder, B., and Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, **14**(3), 289–311.

[5] Goodwin, M.H. and Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica*, **62**, 51–75.

[6] Hart, J.T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, **56**, 208–216.

[7] Hess, U., Kappas, A. and Scherer, K. (1988). Multichannel communication of emotion: synthetic signal production. In K. Scherer (ed.), *Facets of emotion: recent research* (pp. 161-182). Hillsdale, NJ: Erlbaum.

[8] James, W. (1890). *The principles of psychology*, Dover.

[9] Jameson, A., Nelson, T.O., Leonesio, R.J. and Narens, L. (1993). The feeling of another person's knowing. *Journal of Memory and Language* **32**, 320 – 335.

[10] Krahmer, E. and M. Swerts (2005). How children and adults signal and detect uncertainty in audiovisual speech. *Language and Speech* **48**(1), 29–54.

[11] Massaro, D. and Egan, P. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review* **3**, 215–221.

[12] Massaro, D., Cohen, M. and Smeele, P. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America* **100**(3), 1777-1786.

[13] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, 746–748.

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

[15] Mehrabian, A. and M. Wiener (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, **6**, 109–114.

[16] Smith, V.L. and Clark, H.H. (1993). On the course of answering questions. *Journal of Memory and Language*, **32**, 25–38.

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